R A P O R T

eksploracyjna analiza danych i inżynieria cech na przykładzie rejestru aptek

Konrad Dąbrowski
nr albumu: 416119

Spis treści

1 Zapoznanie się z biblioteką seaborn

Biblioteka *seaborn* to narzędzie do wizualizacji danych w języku *Python*, oparte na bibliotece *matplotlib* i ściśle zintegrowane z *pandas*. *Seaborn* oferuje szeroki wachlarz gotowych typów wykresów, takich jak wykresy liniowe, słupkowe, pudełkowe czy korelacyjne. Ta biblioteka jest chętnie wykorzystywana w analizie danych, ponieważ umożliwia tworzenie estetycznych i czytelnych wykresów za pomocą prostych komend. Dodatkowo pozwala na łatwe dodawanie informacji statystycznych, co znacząco ułatwia interpretację prezentowanych danych.

2 Wprowadzenie do strony *dane.gov.pl*

2.1 Sposób udostępniania, rodzaje danych i możliwości wykorzystania

Portal dane.gov.pl to oficjalna platforma udostępniająca otwarte dane publiczne w bezpłatny i dostępny sposób, głównie w popularnych i otwartych formatach (CSV, JSON, XML) oraz poprzez interfejsy API, co umożliwia ich łatwe pobieranie i integrację z systemami informatycznymi. Portal zawiera ponad 27 tysięcy zasobów z różnych dziedzin, takich jak gospodarka, zdrowie, środowisko, demografia czy transport. Dane te można wykorzystywać w celach naukowych, analitycznych, komercyjnych i technologicznych. Udostępniane są na różnych typach licencji Creative Commons, a szczegółowe informacje o rodzaju licencji i warunkach korzystania znajdują się na platformie. Taka forma dostępności sprzyja tworzeniu innowacyjnych projektów oraz rozwojowi usług cyfrowych.

2.2 Rodzaje API na portalu

Na portalu dane.gov.pl dostępnych jest wiele różnych interfejsów API. Każde API ma własną specyfikację i zakres udostępnianych danych – przykładowo API Banku Danych Lokalnych GUS to REST API pozwalające na pobieranie danych statystycznych. API dotyczą różnych baz i rejestrów, a szczegółowa dokumentacja każdego z nich jest dostępna na stronie portalu.

2.3 Szybka weryfikacja jakości danych

Portal dokonuje podstawowej weryfikacji jakości danych podczas ich publikacji. Nowe zasoby są analizowane, sprawdzany jest format, struktura oraz dostępność danych. Wyniki są prezentowane w interfejsie portalu, a użytkownik otrzymuje komunikaty o ewentualnych błędach, które należy poprawić przed publikacją. Dzięki temu użytkownicy mają pewność, że pobierane dane spełniają minimalne standardy jakości i są odpowiednio opisane.

3 Wybór zestawu danych

Do przeprowadzenia analizy wybrano Rejestr Aptek, który jest udostępniany przez Centrum e-Zdrowia - instytucję podlegającą Ministerstwu Zdrowia, odpowiedzialną za zarządzanie systemami informatycznymi w sektorze ochrony zdrowia. Dane są aktualizowane codziennie, a do analizy wykorzystano plik w formacie XLS zawierający stan rejestru na dzień 28 kwietnia 2025 roku.

4 Pierwszy etap pipeline'u uczenia maszynowego

4.1 Opis danych

Zbiór danych obejmuje ponad 23 tysiące wierszy oraz 76 kolumn, co wskazuje na jego szczegółowość i zróżnicowany zakres informacji. Z uwagi na dużą liczbę kolumn, dane zostały pogrupowane tematycznie, co ułatwia ich przegląd. Poniżej przedstawiono główne kategorie informacji zawartych w rejestrze:

- **Identyfikacja** – unikalny identyfikator, nazwa i rodzaj apteki (np. ogólnodostępna).
- **Stan prawny i administracyjny** – Informacje o zezwoleniach na prowadzenie działalności apteki, w tym daty wydania, cofnięcia, wygaszenia i inne szczegóły związane z regulacjami prawnymi.
- **Lokalizacja** – Adres, który obejmuje województwo, powiat, gminę, ulicę i inne dane lokalizacyjne (np. kod pocztowy).
- **Dane kontaktowe** – numer telefonu, faxu, email oraz adres strony internetowej.
- **Sprzedaż wysyłkowa** – informacje dotyczące sprzedaży wysyłkowej, w tym data rozpoczęcia oraz dostępność strony internetowej do tego celu.
- **Kadra zarządzająca** – informacje o kierowniku i zastępcy kierownika apteki, takie jak imiona, nazwiska, numery PWZ i daty rozpoczęcia pracy.
- **Właściciel** – informacje o firmie (np. NIP), adresie oraz dane osobowe właściciela.
- **Godziny otwarcia** – Godziny pracy apteki w poszczególnych dniach tygodnia, w tym godziny otwarcia w niedziele handlowe i niehandlowe.

4.2 Możliwość zastosowania danych w kontekście uczenia maszynowego

Zgromadzony zbiór danych może zostać wykorzystany w analizach opartych na metodach uczenia maszynowego, w szczególności w ramach uczenia nadzorowanego. Przykładowym zastosowaniem może być regresja, której celem jest oszacowanie przewidywanego czasu funkcjonowania apteki. Takie podejście pozwala nie tylko prognozować długość działania apteki, ale również znajdować czynniki, które mogą wpływać na jej trwałość. Wyniki analizy mogą stanowić podstawę do podejmowania decyzji strategicznych przy planowaniu otwarcia nowych obiektów.

4.3 Zakres danych użytych w analizie

W związku z ukierunkowaniem analizy na znalezienie czynników wpływających na długość działania aptek, do dalszych etapów pracy wybrane zostały wyłącznie te kolumny, które mogą mieć znaczenie dla tego aspektu. Ze względu na ograniczony czas przeznaczony na analizę, część zmiennych została odrzucona już na etapie wstępnego przeglądu danych. Do dalszej analizy pozostawiono następujące kolumny:

- **stan_apteki** – wskazuje bieżący stan apteki, tj. czy apteka jest aktywna, czy zawieszona lub zamknięta,
- **rodzaj_apteki** – określa typ apteki,
- **data_uruchomienia_apteki** – data, kiedy apteka rozpoczęła swoją działalność,
- **data_wydania_zezwolenia** – data, kiedy apteka otrzymała zezwolenie na prowadzenie działalności farmaceutycznej,
- **data_cofniecia_zezwolenia** – data, kiedy zezwolenie na prowadzenie apteki zostało cofnięte,
- **data_wygaszenia_zezwolenia** – data, kiedy zezwolenie na prowadzenie apteki wygasło,
- **wojewodztwo** – określa województwo, w którym znajduje się apteka,
- **powiat** – określa powiat, w którym apteka funkcjonuje,
- **czy_szprzedaz_wysylkowa** – informacja o tym, czy apteka prowadzi sprzedaż wysyłkową,
- **telefon** – numer telefonu apteki,
- **email** – adres email apteki,
- **kierownik_npwz** – numer prawa wykonywania zawodu farmaceuty (NPWZ), który pełni funkcję kierownika apteki,
- **kierownik_iiw** – numer wpisu do rejestru osób uprawnionych do wykonywania zawodu technika farmaceutycznego (tzw. IIW – identyfikator innej formy uprawnień),
- **wlasciciel_forma_prawna** – określa formę prawną właściciela apteki.

Do analizy uwzględniono wyłącznie apteki ogólnodostępne oraz punkty apteczne, ponieważ stanowią one najbardziej reprezentatywną grupę placówek działających na rynku detalicznym i są poddane podobnym regulacjom. Pozostałe typy aptek, takie jak apteki szpitalne czy zakładowe, charakteryzują się innym profilem działalności, dlatego zostały pominięte.

4.4 Analiza i uzupełnianie braków danych

Analiza braków danych wykazała, że w dziewięciu cechach występują niepełne informacje, które wymagają szczegółowego przeanalizowania i uzupełnienia. Szczegółowy rozkład braków danych przedstawiono w tabeli 4.1.

Tab. 4.1. – Brakujące dane według cech

Cecha	Ilość brakujących danych
stan_apteki	0
rodzaj_apteki	0
data_uruchomienia_apteki	162
data_wydania_zezwolenia	1
data_cofniecia_zezwolenia	20337
data_wygaszenia_zezwolenia	13788
wojewodztwo	0
powiat	0
czy_szprzedaz_wysylkowa	0
telefon	2278
email	3840
kierownik_npwz	9670
kierownik_iiw	20948
wlasciciel_forma_prawna	796

W celu uporania się z brakami w kolumnie *data_uruchomienia_apteki*, brakujące wartości zostały zastąpione danymi z cechy *data_wydania_zezwolenia*. Analogiczne działanie zastosowano w przypadku kolumny *data_cofnięcia_zezwolenia*, gdzie braki uzupełniono wartościami z cechy *data_wygaszenia_zezwolenia.* Połączone dane z tych dwóch kolumn będą teraz traktowane jako *data_końca_zezwolenia*. Choć takie podejście może wiązać się z pewnym opóźnieniem względem rzeczywistej daty zdarzenia, umożliwia zachowanie ciągłości danych oraz ich pełniejszą analizę. Możliwe jest, dzięki temu, uwzględnienie większej liczby rekordów w dalszych etapach pracy.

W przypadku kolumn *email* oraz *telefon* istotna była nie treść, lecz sam fakt występowania tych wartości. W związku z tym te cechy zostały zbinaryzowane: 0 – brak wartości, 1 – wartość obecna, a nazwy zostały zmienione na *czy_email* oraz *czy_telefon.*

Kolumny *kierownik_iiw* oraz *kierownik_npwz* są względem siebie wykluczające, jeśli jedna z nich zawiera wartość, druga pozostaje pusta i odwrotnie. Z tego powodu obie kolumny zakodowano w zmiennej *kierownik*, przy czym, *1* oznacza obecność danych z kolumny *kierownik_iiw*, a 2 obecność z kolumny *kierownik_npwz*. Nowa kolumna została wstępnie zainicjalizowana wartością *np.nan*, aby uniknąć nadpisania potencjalnych braków danych i umożliwić ich późniejszą identyfikację.

Na obecnym etapie analizy nie zastosowano metod usuwania braków danych ani ich szacowania na podstawie pozostałych cech. W niektórych przypadkach takie podejście okazało się skuteczne, jednak nie we wszystkich. W trzech zmiennych nadal występują braki *(Tab. 4.2)*. Z powodu, że dane obejmują cały kraj, analizowanie ich na tym poziomie może prowadzić do zbyt ogólnych wniosków. Odsetek brakujących danych może się istotnie różnić w zależności od powiatu, dlatego na dalszym etapie niezbędna jest bardziej szczegółowa analiza na poziomie powiatowym. Takie podejście pozwala uniknąć potencjalnego zniekształcenia wyników analizy, do którego mogłoby dojść w przypadku usunięcia zbyt dużej liczby obserwacji z danego regionu.

Tab. 4.2. – Brakujące dane według cech

Cecha	Ilość brakujących danych
stan_apteki	0
rodzaj_apteki	0
wojewodztwo	0
powiat	0
data_uruchomienia_apteki	0
data_konca_zezwolenia	12659
czy_szprzedaz_wysylkowa	0
czy_telefon	0
czy_email	0
kierownik	9142
wlasciciel_forma_prawna	796

4.5 Klasteryzacja powiatów

Przy analizowaniu cechy *powiaty* napotkano na problem związany z liczbą jednostek administracyjnych – znaleziono 375 powiatów. Stwierdzono występowanie 10 par powiatów o identycznych nazwach, lecz zlokalizowanych w różnych województwach:

1. powiat bielski – województwo podlaskie oraz śląskie,
2. powiat brzeski – województwo małopolskie oraz opolskie,
3. powiat grodziski – województwo mazowiecki oraz wielkopolskie,
4. powiat krośnieński – województwo lubuskie oraz podkarpackie,
5. powiat nowodworski – województwo mazowieckie oraz pomorskie,
6. powiat opolski – województwo lubelskie oraz opolskie,
7. powiat ostrowski – województwo mazowieckie oraz wielkopolskie,
8. powiat średzki – województwo dolnośląskie oraz wielkopolskie,
9. powiat świdnicki – województwo dolnośląskie oraz lubelskie,
10. powiat tomaszowski – województwo lubelskie oraz łódzkie.

W celu rozwiązania problemu, do nazw powiatów dodano odpowiednie kody województw, zgodne z dwiema pierwszymi cyframi systemu identyfikatorów TERC (system identyfikacji jednostek podziału terytorialnego kraju). Pozwoliło to na rozróżnienie wszystkich 380 powiatów – 66 miast na prawach powiatu oraz 314 powiatów ziemskich.

Kardynalność rzędu 380 została uznana za zbyt wysoką. W celu redukcji cech pozyskano dane dotyczące liczby ludności oraz gęstości zaludnienia (na 1 km²) w powiatach w latach 2002–2023. Dane te zostały pobrane z Banku Danych Lokalnych GUS, który umożliwia ręczne wybieranie statystyk, regionów oraz przedziału czasowego.

Wybrano maksymalnie dostępny przedział czasowy, ponieważ nowsze dane nie były dostępne, a obecny podział administracyjny powiatów obowiązuje od 2002 roku. Wyjątek stanowi miasto Wałbrzych, które w latach 2003–2013 utraciło status miasta na prawach powiatu. W bazie możliwe było jednak pobranie informacji dotyczących samego miasta, zatem wartości dla powiatu wałbrzyskiego w tamtych latach zredukowano o dane dla miasta. Pozwoliło to na uzyskanie ciągłego zestawu danych dla wszystkich powiatów w Polsce.

Pozyskane dane zostały wykorzystane do klasteryzacji powiatów pod względem gęstości oraz obciążenia aptek, rozumianego jako stosunek liczby ludności do liczby aptek w poszczególnych latach. W celu wyboru odpowiedniej miary tendencji centralnej przeanalizowano rozkłady wartości średniej i mediany, co przedstawiono na wykresie *(Fig. 4.1)*. Rozkład mediany gęstości wykazywał skośność równą 2,39, natomiast skośność rozkładu mediany obciążenia wynosiła 0,68. Oba rozkłady cechowały się asymetrią prawostronną, co oznacza, że większość obserwacji skupiała się wokół niższych wartości. W związku z tym, do dalszych obliczeń wybrano medianę jako bardziej odporną na wartości odstające, które mogłyby zaburzyć średnią.

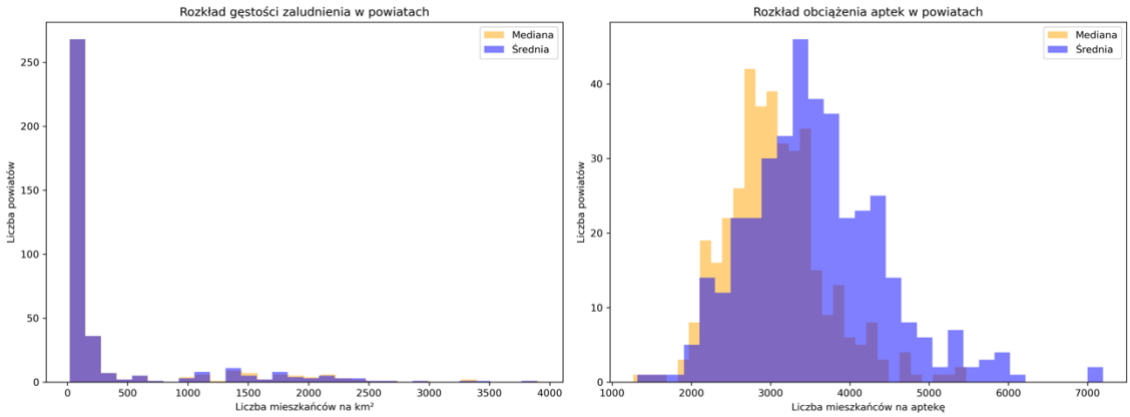

Fig. 4.1.

W celu określenia optymalnej liczby klastrów, połączono dane dotyczące mediany gęstości zaludnienia oraz obciążenia aptek dla każdego powiatu w jedną tabelę. Następnie dane zostały ustandaryzowane, aby zapewnić porównywalność zmiennych o różnych skalach. Do analizy zastosowano metodę łokcia, która pozwala na wybór najbardziej odpowiedniej liczby klastrów w algorytmie grupowania K-średnich. Metoda ta polega na znalezieniu k środków w taki sposób, aby łączna suma odległości punktów od najbliższego centroidu była jak najmniejsza. W analizie rozważano wartości k od 2 do 20.

Na wykresie *(Fig. 4.2)* oś X przedstawia liczbę klastrów, a oś Y łączną sumę odległości punktów od przypisanych im centroidów. Punkt, w którym dalsze zwiększanie liczby klastrów nie prowadzi już do istotnego zmniejszenia tej sumy, uznawany jest za optymalną liczbę klastrów. Na poniższym wykresie charakterystyczne załamanie, przypominające kształtem zgięcie łokcia, widoczne jest przy 3 klastrach.

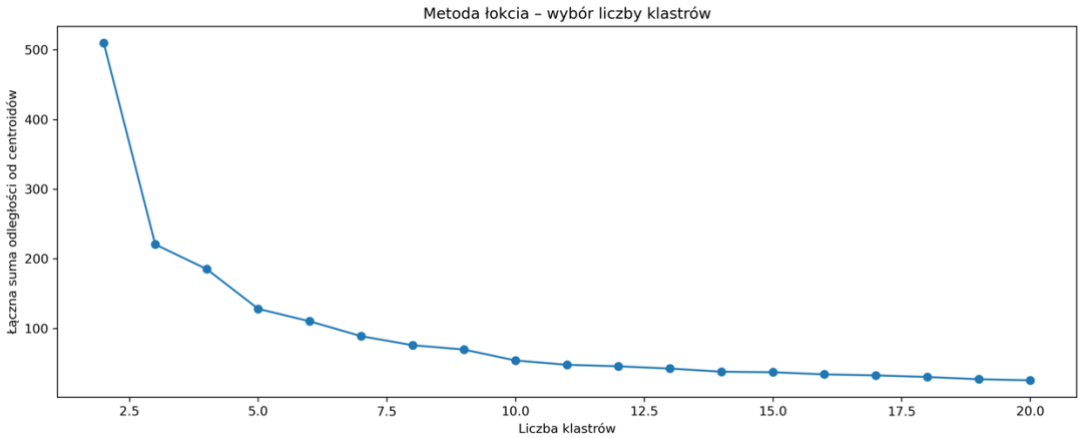

Fig. 4.2.

W celu przeprowadzenia charakterystyki klastrów sporządzono wykresy pudełkowe *(Fig. 4.3)*, przedstawiające rozkład gęstości zaludnienia oraz obciążenia aptek w poszczególnych klastrach. Dodatkowo, w celu analizy przestrzennej, klastry zostały naniesione na mapę Polski *(Fig. 4.4)*. Na tej podstawie można wyróżnić trzy typy powiatów:

- **Klaster 0** – obejmuje głównie powiaty o niskiej gęstości zaludnienia i niskim obciążeniu aptek. Są to w głównie obszary wiejskie i małe miasta.
- **Klaster 1** – skupia powiaty o niskiej gęstości zaludnienia, ale wysokim obciążeniu aptek. Często występują w sąsiedztwie dużych miast (klaster 2), co może wskazywać na zależność od infrastruktury medycznej większych ośrodków miejskich.
- **Klaster 2** – reprezentuje powiaty o wysokiej gęstości zaludnienia i najbardziej zróżnicowanym, choć małym obciążeniu aptek. Klaster ten koncentruje się wokół największych miast i aglomeracji.

Powyższa charakterystyka potwierdza, że dokonany podział powiatów na klastry jest trafny i dobrze odwzorowuje zróżnicowanie terytorialne w kontekście dostępności usług aptecznych. W związku z tym, kolumnę zawierającą nazwę powiatu w rejestrze aptek zastąpiono nową cechą *klaster_powiat*, która zawiera numer przypisanego klastra dla danego powiatu.

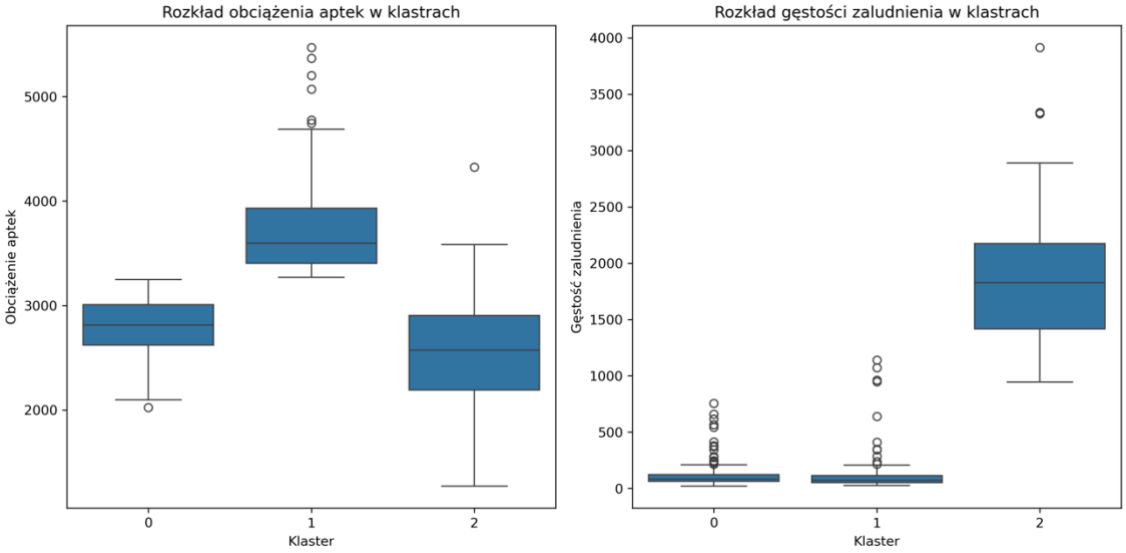

Fig. 4.3.

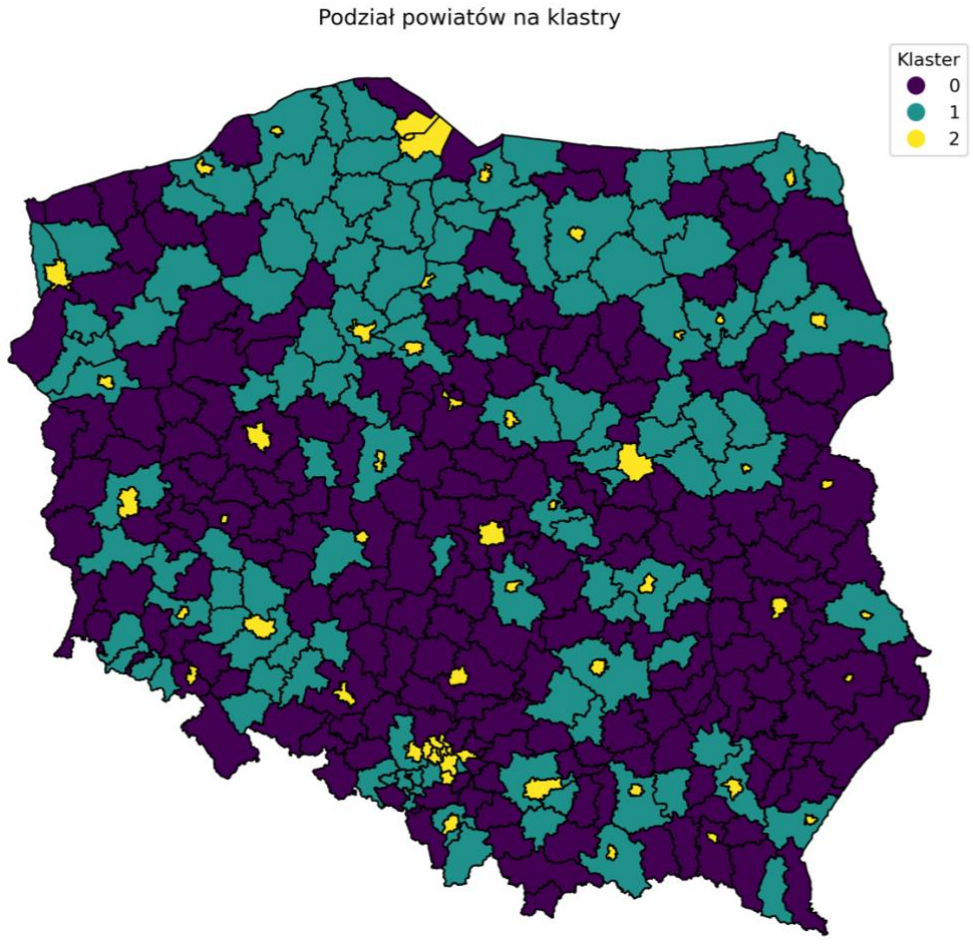

Fig. 4.4.

4.6 Ponowna analiza braków danych w kontekście klastrów powiatów

W kolumnie *data_konca_zezwolenia* zbadano brakujące wartości w podziale na status apteki *(Tab. 4.3)*.

Tab. 4.3. – Liczba brakujących wartości w kolumnie data_konca_zezwolenia w zależności od statusu apteki

Stan apteki	Ilość brakujących dat końca zezwolenia
aktywna	12340
czasowo nieczynna	86
nieaktywna	38
nieaktywna zawieszenie działalności	189
oczekująca	5
unieruchomiona decyzją wif	1

Braki dat końca zezwolenia w przypadku aptek aktywnych jest uzasadniony, dlatego dalszą analizę ograniczono wyłącznie do rekordów, które nie są aktywne. Wśród wszystkich aptek o statusie nieaktywnym, 3,49% nie posiada informacji o dacie końca zezwolenia. Choć wartość ta nie wydaje się znacząca, zdecydowano się sprawdzić jej rozkład w poszczególnych klastrach, w niektórych z nich udział brakujących danych mógłby być na tyle wysoki, że zaburzyłby późniejsze analizy. Jednak, jak pokazano w tabeli 4.4, w żadnym z klastrów udział brakujących danych nie przekracza 5%, co pozwala na bezpieczne usunięcie tych rekordów bez ryzyka istotnego zaburzenia wyników.

Tab. 4.4. – Odsetek brakujących dat końca zezwolenia wśród aptek nieaktywnych według klastrów powiatów

klaster_powiat	Odsetek brakujących dat końca zezwolenia
2	3,80 %
0	3,67 %
1	2,58 %

Cecha *data_konca_zezwolenia* nadal zawiera brakujące wartości dla aktywnych aptek. W związku z tym zdecydowano o jej zastąpieniu nową zmienną, opisującą czas działania apteki. Przed przekształceniem przeanalizowano jednak liczność etykiet dla obu cech czasowych *(Tab. 4.5)*.

Tab. 4.5. – Liczba etykiet dla wybranych cech

Cecha	Liczba etykiet
data uruchomienia apteki	7342
data konca zezwolenia	3191

Kardynalność okazała się zbyt wysoka, dlatego zdecydowano się na redukcję obu cech do poziomu roku. W wyniku tego działania liczba etykiet została zmniejszona o ponad 99%. Po redukcji liczba unikalnych wartości dla roku uruchomienia wynosi 42, a dla roku zakończenia działalności 24. Pozwoliło to na wyznaczenie czasu działania aptek. W przypadku aptek aktywnych okres ten liczony jest do roku 2025, co odzwierciedla ich aktualny staż funkcjonowania.

W przypadku cechy *kierownik* brak danych wynosi 42,47% dla całego zbioru. To na tyle duża wartość, że rzetelne oszacowanie brakujących danych byłoby trudne i potencjalnie obarczone dużym błędem. W celu sprawdzenia czy braki danych mogą być w jakimś stopniu zależne od innych zmiennych, przeanalizowano ich rozkład względem aktywności aptek. Okazało się, że 96% nieaktywnych aptek nie zawiera informacji o kierowniku. Jest to uzasadnione, istnieje duże prawdopodobieństwo, że dane wrażliwe, takie jak numer prawa wykonywania zawodu, są usuwane po zamknięciu apteki. W przypadku aktywnych aptek braki te stanowią jedynie 4%. W związku z powyższym należy zauważyć, że wszystkie kolejne analizy, w których wykorzystywana jest cecha kierownik, są w pewnym stopniu obarczone błędem wynikającym z niepełnych danych. Mimo to, w dalszych krokach przeanalizowano zależność pomiędzy rodzajem apteki a rodzajem kierownika. Wyniki przedstawiono w Tabeli 4.6. Z obserwacji wynika, że wśród aptek ogólnodostępnych aż 99,99% kierowników to zawodowi farmaceuci, natomiast w punktach aptecznych 90,10% kierowników stanowią technicy farmaceutyczni.

Tab. 4.6. – Rozkład rodzaju kierownika w zależności od rodzaju apteki

Rodzaj apteki	Technik farmaceutyczny	Farmaceuta
apteka ogólnodostępna	0.01%	99.99%
punkt apteczny	90.1%	9.9%

Aby statystycznie potwierdzić istnienie zależności między tymi dwiema zmiennymi, zastosowano test chi-kwadrat niezależności. Test ten sprawdza, czy zmienne kategoryczne są od siebie statystycznie niezależne. Wartość obliczonej statystyki chi-kwadrat to 10868,78, a wartość *p-value* to 0,0. Otrzymany wynik *p-value* równy 0.0 (czyli mniejszy niż jakikolwiek przyjęty poziom istotności, np. 0,05) pozwala jednoznacznie odrzucić hipotezę zerową o niezależności zmiennych. Oznacza to, że istnieje statystycznie istotna zależność między rodzajem apteki a typem osoby pełniącej funkcję kierownika. W związku z tym cecha kierownik została wykluczona z dalszej analizy.

Kolumna *właściciel_forma_prawna* zawiera 792 brakujące wartości, co stanowi 3,74% wszystkich rekordów w zbiorze. Analiza braków w podziale na klastry powiatów *(Tab. 4.7)* wykazała, że odsetek brakujących danych nie przekracza 5% w żadnym z klastrów. W związku z tym uznano, że brakujące rekordy można bezpiecznie usunąć bez istotnego wpływu na jakość dalszych analiz.

Tab. 4.7. – Odsetek brakujących form prawnych właściciela wśród aptek według klastrów powiatów

Klaster powiatu	Odsetek brakujących form prawnych właściciela
0	4,10 %
2	3,86 %
1	2,81 %

4.7 Analiza kardynalności i redukcja liczby etykiet

Wszystkie cechy zostały oczyszczone z brakujących wartości. Przeprowadzono analizę kardynalności, określając liczbę unikalnych etykiet dla wszystkich kolumn *(Tab. 4.8)*. Powiaty zostały wcześniej sklasteryzowane, co ograniczyło liczbę etykiet z 380 do 3. Kardynalność cech *rok_uruchomienia_apteki* oraz *lata_działania*, mimo wcześniejszej redukcji, pozostaje stosunkowo wysoka w porównaniu do pozostałych cech. Jednak przy liczbie 20365 obserwacji nie stanowi to istotnego problemu.

Tab. 4.8. – Liczba unikalnych etykiet dla wybranych cech w zbiorze danych

Cecha	Liczba etykiet
stan apteki	2
rodzaj_apteki	2
wojewodztwo	16
klaster powiat	3
rok uruchomienia apteki	42
lata dzialania	53
czy szprzedaz wysylkowa	2
czy_telefon	2
czy email	2
wlasciciel_forma_prawna	17

Przeanalizowano również cechę *właściciel_forma_prawna (Tab. 4.9)*. Ze względu na dużą rozbieżność w liczbie aptek przypisanych do poszczególnych form prawnych, dokonano redukcji liczby etykiet do czterech ogólnych kategorii: *Inne*, *Spółka osobowa*, *Spółka kapitałowa* oraz *Osoba fizyczna (Tab. 4.10)*. Zabieg ten pozwala uniknąć sytuacji, w której niektóre formy prawne byłyby reprezentowane przez zbyt małą liczbę obserwacji, co mogłoby negatywnie wpłynąć na jakość dalszych analiz. Nazwę cechy zmieniono na *forma_prawna_kategoria*.

Tab. 4.9. – Udział procentowy form prawnych właścicieli aptek przed redukcją etykiet

Forma prawna właściciela	Procentowy udział
Jednostka budżetowa	0,005 %
Spółdzielnia	0,005 %
Spółka z o.o. komandytowo-akcyjna	0,02 %
Inna instytucja lub osoba	0,029 %
Spółka z ograniczoną odpowiedzialnością z udziałem jednostek samorządu terytorialnego	0,044 %
Spółka partnerska	0,083 %
Kościół lub związek wyznaniowy	0,093 %
Spółka komandytowo-akcyjna	0,098 %
SPZOZ	0,108 %
Fundacja, stowarzyszenie	0,113 %
Spółka z o.o. spółka komandytowa	0,904 %
Spółka komandytowa	1,365 %
Spółka akcyjna	2,642 %
Spółka cywilna	4,591 %
Spółka jawna	21,291 %
Spółka z ograniczoną odpowiedzialnością	29,187 %
Osoba fizyczna prowadząca działalność gospodarczą	39,421 %

Tab. 4.10. – Udział procentowy form prawnych właścicieli aptek po redukcji etykiet

Kategoria formy prawnej właściciela	procent formy prawnej
Inne	0,35 %
Spółka osobowa	28,35 %
Spółka kapitałowa	31,87 %
Osoba fizyczna	39,42 %

4.8 Wizualna analiza cech

Przeanalizowano cechę *stan_apteki*, wykres kołowy przedstawia udział aktywnych i nieaktywnych aptek w Polsce na dzień 28.04.2025 *(Fig. 4.5)*. Z danych wynika, że 60,6% aptek (12346) jest aktywnych, natomiast 39,4% (8019) stanowią apteki nieaktywne. Kolory zielony i czerwony odpowiadają odpowiednio aptekom aktywnym i nieaktywnym. Wykres słupkowy pokazuje procentowy rozkład stanów aptek w trzech klastrach powiatów. W każdym z klastrów udział aptek aktywnych utrzymuje się na poziomie około 60%, a nieaktywnych około 40%. Struktura udziału aptek aktywnych i nieaktywnych jest bardzo podobna we wszystkich analizowanych klastrach powiatów, co sugeruje brak istotnych różnic regionalnych w tym zakresie.

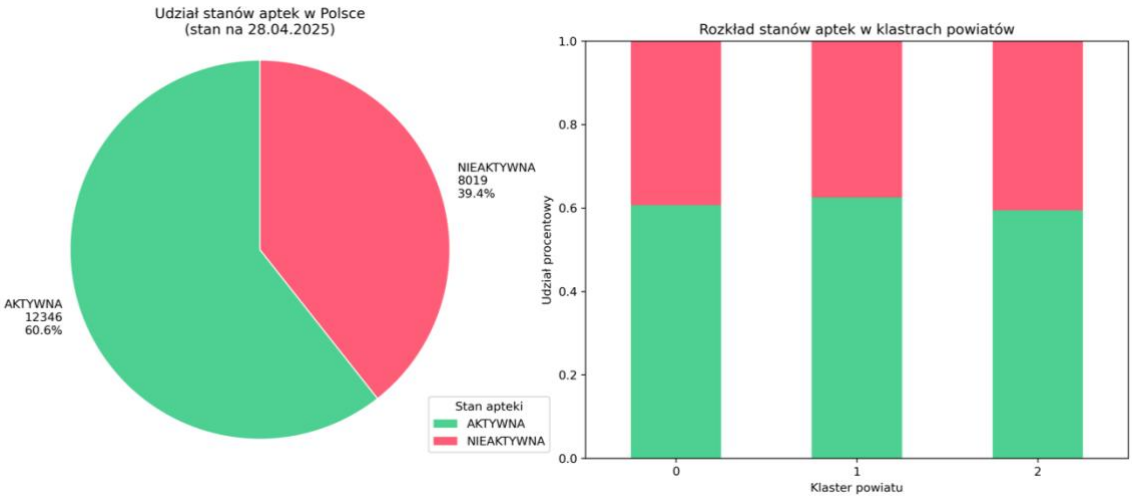

Fig. 4.5.

Kolejnym analizowanym aspektem charakterystyki aptek w Polsce jest ich rodzaj. Wykres kołowy przedstawia udział poszczególnych rodzajów aptek w Polsce *(Fig. 4.6)*. Zdecydowaną większość stanowią apteki ogólnodostępne, których jest 17902, co odpowiada 87,9% wszystkich placówek. Punkty apteczne to 2463 jednostki, czyli 12,1%. Wykres słupkowy ilustruje rozkład rodzajów aptek w trzech klastrach powiatów. W skali kraju apteki ogólnodostępne stanowią zdecydowaną większość wszystkich placówek aptecznych, natomiast punkty apteczne pełnią rolę uzupełniającą. Punkty apteczne są znacznie częściej spotykane w powiatach o niższej gęstości zaludnienia (klastry 0 i 1), co wskazuje na ich istotną rolę w zapewnianiu dostępu do leków na terenach słabiej zurbanizowanych. W największych miastach (klaster 2) dominują wyłącznie apteki ogólnodostępne. Może to świadczyć o wyższym poziomie infrastruktury farmaceutycznej w tych obszarach. Rozkład rodzajów aptek w poszczególnych klastrach odzwierciedla zróżnicowanie potrzeb lokalnych społeczności oraz dostosowanie sieci aptecznej do warunków demograficznych i urbanizacyjnych.

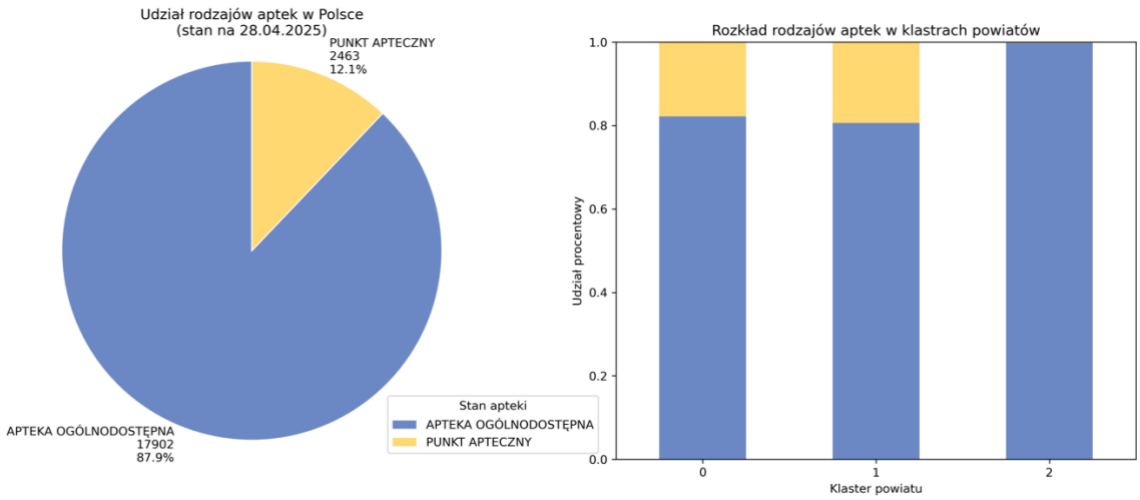

Fig. 4.6.

Następnie przeanalizowano rozmieszczenie aptek w układzie wojewódzkim. Wykres słupkowy *(Fig. 4.7)* przedstawia liczbę rekordów w rejestrze według województw. Najwięcej rekordów odnotowano w województwie mazowieckim, małopolskim, śląskim oraz wielkopolskim. Najmniej rekordów występuje w województwach opolskim, lubuskim, podlaskim oraz świętokrzyskim. Pokrywa się to z najbardziej i najmniej zaludnionymi województwami. Największa liczba rekordów w rejestrze dotyczy województw z największymi miastami i aglomeracjami, co wskazuje na koncentrację infrastruktury medycznej oraz aptek w tych regionach.

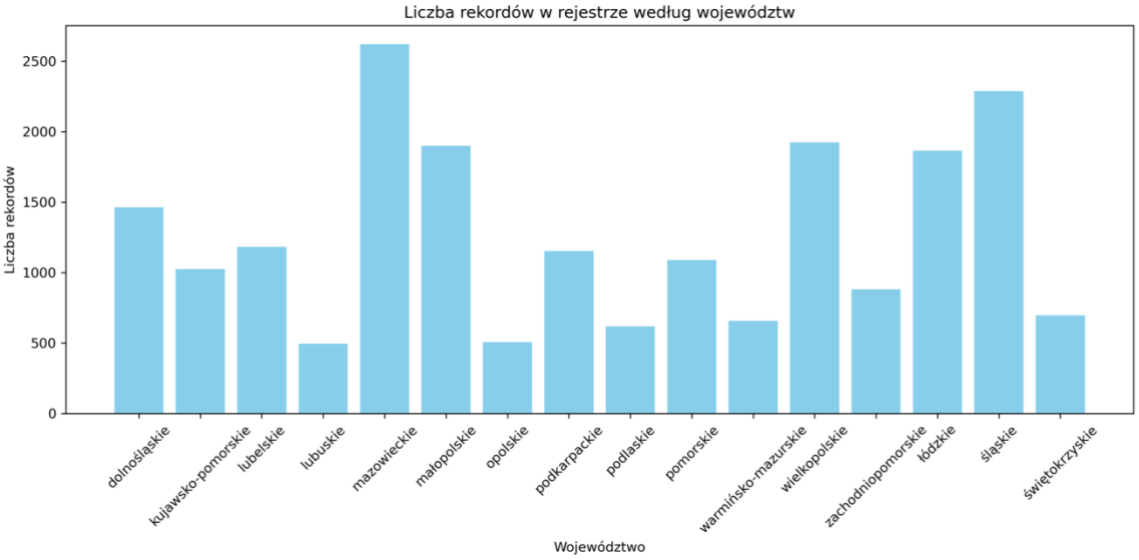

Fig. 4.7.

Rozkład liczby aptek według roku ich uruchomienia stanowi kolejny element analizy *(Fig. 4.8)*. W danych widoczny jest pewien rozrzut w zakresie lat, bezpośrednio po roku 1900 pojawia się rok 1990, a kolejne lata są już reprezentowane w sposób ciągły. Sugeruje to, że apteki otwarte przed 1900 rokiem zostały przypisane właśnie do tej daty, co skutkuje wyraźnym pikiem w tym roku. Z kolei apteki otwarte w latach 1901–1989 przypisano do roku 1990, natomiast pozostałe uwzględniono zgodnie z rzeczywistą datą otwarcia. Zdecydowana większość aptek powstała po 1990 roku, przy czym szczególnie intensywny wzrost liczby nowych placówek obserwowano w latach 2000–2020. Drugi wykres ilustruje rozkład liczby aptek według długości ich działalności. Najwięcej placówek funkcjonuje od kilku do kilkudziesięciu lat, a największa koncentracja przypada na przedział od 5 do 20 lat działalności. Największy odsetek stanowią apteki stosunkowo młode, co wskazuje na dużą dynamikę w zakresie powstawania nowych punktów oraz możliwą rotację na rynku. Obserwuje się wyraźną przewagę aptek o krótkim i średnim stażu nad tymi, które działają od wielu dekad. Rozkład lat działalności potwierdza, że tylko nieliczne apteki mają ponad 100-letnią historię, co jest zbieżne z rozkładem roku ich uruchomienia.

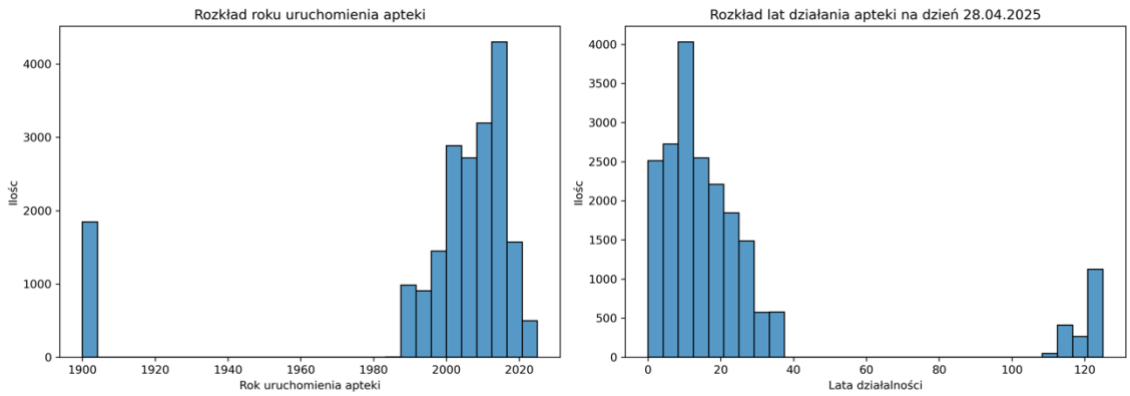

Fig. 4.8.

Kolejnym elementem analizy były cechy takie jak dostępność numeru telefonu, adresu email oraz informacja o prowadzeniu sprzedaży wysyłkowej *(Fig. 4.9)*. We wszystkich trzech klastrach powiatów zdecydowana większość aptek posiada numer telefonu, odsetek ten przekracza 90%. Udział aptek bez telefonu jest znikomy i bardzo zbliżony we wszystkich klastrach, co pozwala uznać tę formę kontaktu za standard w skali kraju. Adres email deklaruje około 80–85% aptek, nieco mniej niż w przypadku telefonu. Również tutaj różnice między klastrami są niewielkie. Największe zróżnicowanie między klastrami obserwuje się w przypadku cechy sprzedaży wysyłkowej. Usługę tę oferuje jedynie niewielki odsetek aptek (poniżej 5%) we wszystkich klastrach. Jej udział jest jednak nieco wyższy w klastrze 2, obejmującym największe miasta i aglomeracje, co może wynikać z większego zapotrzebowania na tego typu usługi oraz lepszych warunków logistycznych w tych regionach.

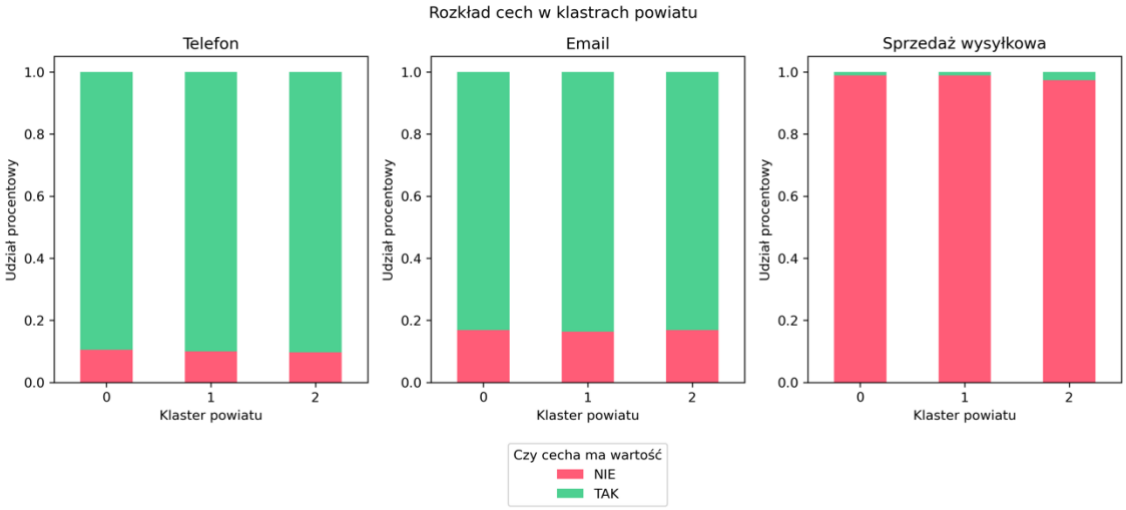

Fig. 4.9.

Rozkład form prawnych prowadzenia aptek znacząco różni się w zależności od charakterystyki powiatów, co można zaobserwować na poniższych wykresach *(Fig. 4.10.)*. Zestawienie w formie wykresu słupkowego oraz mapy ciepła uwzględnia kategorie form prawnych. W klastrach 0 i 1, obejmujących obszary o niskiej gęstości zaludnienia, dominują apteki prowadzone przez osoby fizyczne (ponad 40%), co może wynikać z przewagi lokalnych działalności w tych środowiskach. W klastrze 2, który reprezentuje aglomeracje, zaobserwowano odmienną strukturę. Dominującą formą prawną są spółki kapitałowe, które stanowią aż 43% wszystkich aptek w tej grupie. Udział osób fizycznych spada w tym przypadku do 27%. Może to świadczyć o większej złożoności działalności aptecznej w środowiskach wielkomiejskich, gdzie skala operacyjna, konkurencja oraz wymagania logistyczne są większe. Udział kategorii „inne" we wszystkich klastrach jest znikomy, co potwierdza marginalne znaczenie niestandardowych form prawnych. Stały udział spółek osobowych we wszystkich klastrach sugeruje, że ta forma działalności znajduje zastosowanie w różnych warunkach, zarówno na terenach miejskich, jak i wiejskich.

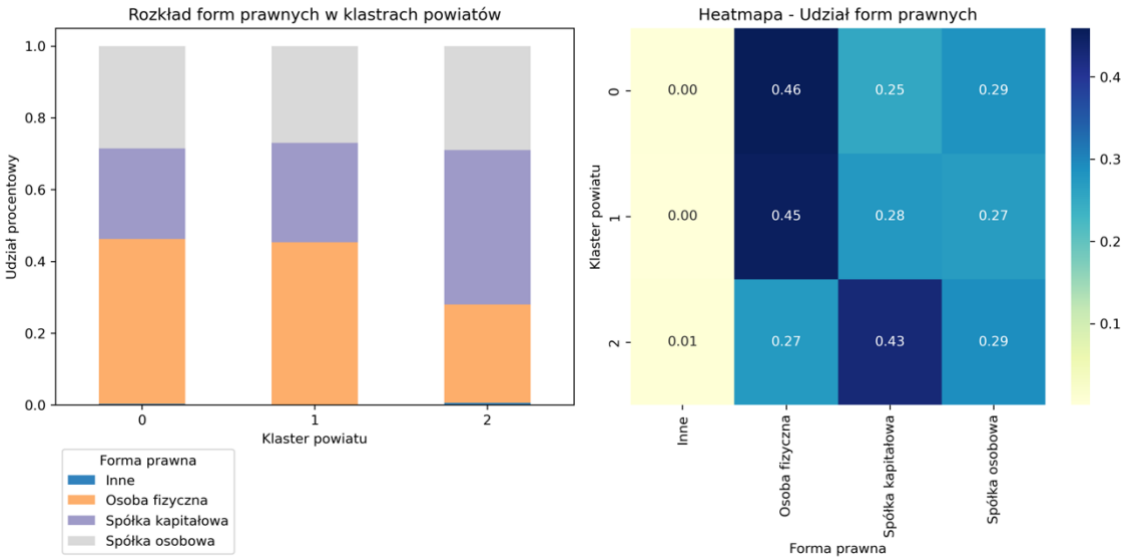

Fig. 4.10.

Na wykresie *(Fig. 4.11)* zaprezentowano zmiany liczby aptek w Polsce w latach 1990–2024. Wcześniejsze dane nie zostały uwzględnione, ponieważ nie były aktualizowane corocznie, przez co wykres zaczyna się od pewnego poziomu liczby aptek. Dane ukazują wyraźny trend wzrostowy od początku lat 90. XX wieku aż do roku 2017, po którym nastąpił systematyczny spadek. W tych latach liczba aptek wzrosła ponad sześciokrotnie – z około 2,5 tysiąca do ponad 14,5 tysiąca. Tak dynamiczny przyrost można wiązać z brakiem ograniczeń regulacyjnych, które umożliwiły szerokiemu gronu przedsiębiorców zakładanie sieci aptecznych. Punkt zwrotny nastąpił po roku 2017, kiedy to wprowadzono nowelizację ustawy prawa farmaceutycznego, tzw. Apteka dla Aptekarza. Przepisy te ograniczyły możliwość zakładania nowych aptek wyłącznie do farmaceutów oraz wprowadziły kryteria geograficzne i demograficzne. Celem regulacji było ograniczenie ekspansji dużych sieci i wzmocnienie pozycji aptek indywidualnych. W efekcie, od 2017 roku obserwowany jest spadek liczby funkcjonujących aptek – do poziomu poniżej 12,5 tysiąca w roku 2024. Spadek liczby aptek może prowadzić do ograniczenia dostępności leków oraz pogorszenia warunków konkurencyjnych na rynku, szczególnie na obszarach wiejskich i w mniejszych miejscowościach. Zmiana ta wskazuje na konieczność dalszego monitorowania efektów wprowadzanych regulacji pod kątem ich wpływu na równomierne pokrycie terytorialne.

Fig. 4.11.

Trend zmian liczby aptek w latach 1990–2024 przedstawiono na wykresie *(Fig. 4.12)*. Potwierdza się na nim, że od 2017 roku nastąpił gwałtowny spadek liczby nowo otwieranych aptek. Już w 2018 roku bilans nowo powstałych placówek stał się ujemny i utrzymuje się na ujemnym poziomie do dziś. Jednocześnie wykres wskazuje na inną istotną tendencję, od 2011 roku obserwowany jest wyraźny wzrost liczby zamykanych aptek, podczas gdy wcześniej ich liczba była niska. W 2011 roku wprowadzono ustawę refundacyjną, która ustanowiła sztywne ceny i marże na leki refundowane oraz ograniczyła możliwość stosowania promocji i rabatów przez apteki. W efekcie sytuacja finansowa wielu małych, indywidualnych aptek uległa znacznemu pogorszeniu, co przyczyniło się do ich upadku oraz przyspieszyło ekspansję dużych sieci aptecznych.

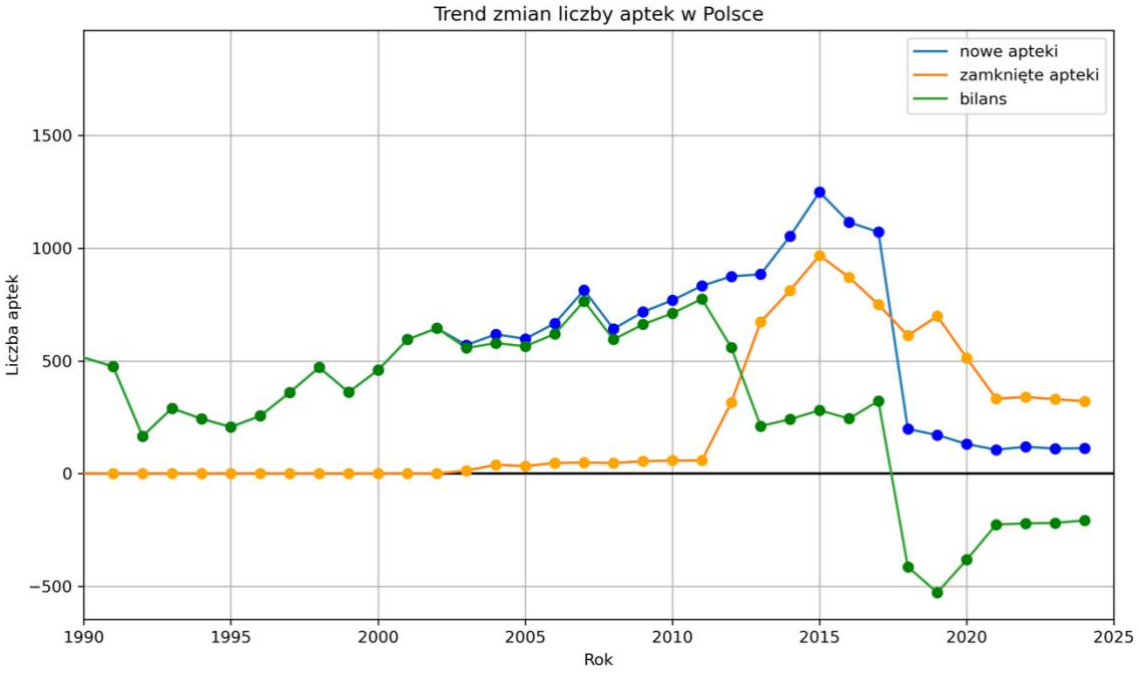

Fig. 4.12.

4.9 Kodowanie zmiennych kategorycznych

Kolejnym krokiem było przygotowanie zmiennych kategorycznych do wykorzystania w modelach uczenia maszynowego, które wymagają danych w formacie numerycznym. W tym celu zastosowano kodowanie numeryczne wybranych cech, tak aby w zbiorze danych nie pozostały już żadne nienumeryczne kolumny. Zmienne tekstowe zostały zmapowane na wartości całkowite przy użyciu predefiniowanych słowników:

- kolumna **stan_apteki** została przekształcona na wartości binarne *(0 dla NIEAKTYWNA, 1 dla AKTYWNA),*
- kolumna **rodzaj_apteki** *(0 dla APTEKA OGÓLNODOSTĘPNA, 1 dla PUNKT APTECZNY),*
- kolumna **czy_szprzedaz_wysylkowa** *(0 dla NIE, 1 dla TAK),*
- kolumna **wojewodztwo** przekształcona na część wojewódzką kodów TERC,
- kolumna **forma_prawna_kategoria** została zmapowana na wartości całkowite od 0 do 3, reprezentujące różne kategorie form prawnych.

5 Wybór zmiennej *target* do uczenia nadzorowanego

Do uczenia nadzorowanego jako zmienną *target* wybrano kolumnę lata_dzialania. Zmienna ta jest dobrym kandydatem, szczególnie w kontekście analizy regresyjnej. Celem takiego modelu mogłoby być przewidywanie, jak długo dana apteka będzie funkcjonować na rynku. Informacja ta ma wartość biznesową, ponieważ pozwala na identyfikację czynników wpływających na stabilność i długowieczność aptek. Co więcej, dane dotyczące lat działalności zostały wcześniej przygotowane i przetworzone, m.in. poprzez redukcję kardynalności do poziomu roku, co ułatwia ich modelowanie. W przypadku aptek wciąż aktywnych, okres ten jest liczony do roku 2025, co odzwierciedla ich aktualny staż funkcjonowania.

6 Wybór zmiennych *features*

Do wyznaczenia zmiennej *target (lata_dzialania)* można wybrać następujący podzbiór zmiennych *features*:

- **rodzaj_apteki**: Rodzaj apteki może wpływać na jej stabilność i długość działania. W raporcie zauważono, że punkty apteczne pełnią rolę uzupełniającą, szczególnie na terenach słabiej zurbanizowanych, co może wiązać się z innymi uwarunkowaniami rynkowymi niż w przypadku aptek ogólnodostępnych.
- **klaster_powiat**: Zmienna ta powstała w wyniku klasteryzacji powiatów pod względem gęstości zaludnienia i obciążenia aptek. Charakterystyka klastra, w którym działa apteka, może mieć istotny wpływ na jej perspektywy i czas funkcjonowania.
- **wojewodztwo**: Lokalizacja na poziomie województwa może odzwierciedlać regionalne uwarunkowania ekonomiczne, demograficzne oraz specyfikę rynku farmaceutycznego, co może wpływać na długość działania apteki.
- **czy_szprzedaz_wysylkowa**: Prowadzenie sprzedaży wysyłkowej może być czynnikiem różnicującym, wpływającym na zasięg i potencjalne przychody apteki, a tym samym na jej żywotność.
- **czy_telefon** oraz **czy_email**: Posiadanie danych kontaktowych, takich jak telefon czy email, może świadczyć o pewnym poziomie organizacji i profesjonalizmu apteki, co pośrednio może przekładać się na jej stabilność.
- **forma_prawna_kategoria**: Kategoria formy prawnej właściciela apteki może mieć związek z jej stabilnością finansową, dostępem do kapitału i strategią działania. Analiza wykazała różnice w dominujących kategoriach form prawnych w zależności od klastra powiatu, co sugeruje ich potencjalny wpływ na funkcjonowanie aptek.
- **rok_uruchomienia_apteki**: Chociaż lata_dzialania są bezpośrednio powiązane z rokiem uruchomienia, sama data startu może nieść dodatkowe informacje o warunkach rynkowych panujących w momencie otwarcia apteki (np. zmiany regulacyjne, nasycenie rynku), które mogły wpłynąć na jej dalsze losy.