Analiza danych rzeczywistych przy pomocy modelu ARMA

Komputerowa analiza szeregów czasowych

Aleksandra Szczur 276047
Agnieszka Staszkiewicz 268791

31.01.2025

Spis treści

1 Wstęp

1.1 Cel pracy

Celem naszego raportu jest analiza danych sprzedaży lodów przy wykorzystaniu metod analizy szeregów czasowych. W szczególności zbadamy długoterminowe trendy oraz sezonowe wzorce sprzedaży lodów na przestrzeni 48 lat. Wyniki analizy umożliwią lepsze zrozumienie mechanizmów zmienności sprzedaży lodów w długim okresie czasu oraz potencjalne zastosowanie modelu w celach prognostycznych.

1.2 Charakterystyka zbioru danych

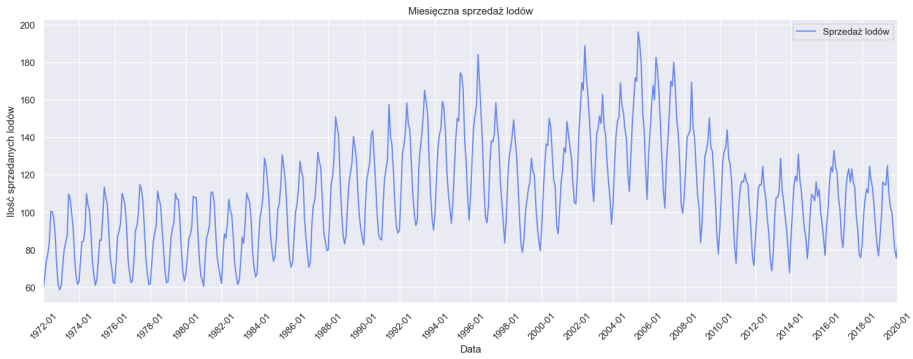
Zestaw danych "Monthly Ice Cream Sales Data (1972-2020)", dostępny na platformie Kaggle, to zbiór reprezentujący miesięczną ilość sprzedanych lodów od stycznia 1972 do stycznia 2020.

Zbiór obejmuje 578 rekordów. Każdy rekord zawiera 2 kolumny:

- DATE - data obserwacji w formacie RRRR-MM-DD
- IPN31152N - miesięczna ilość sprzedanych lodów

1.3 Wizualizacja danych

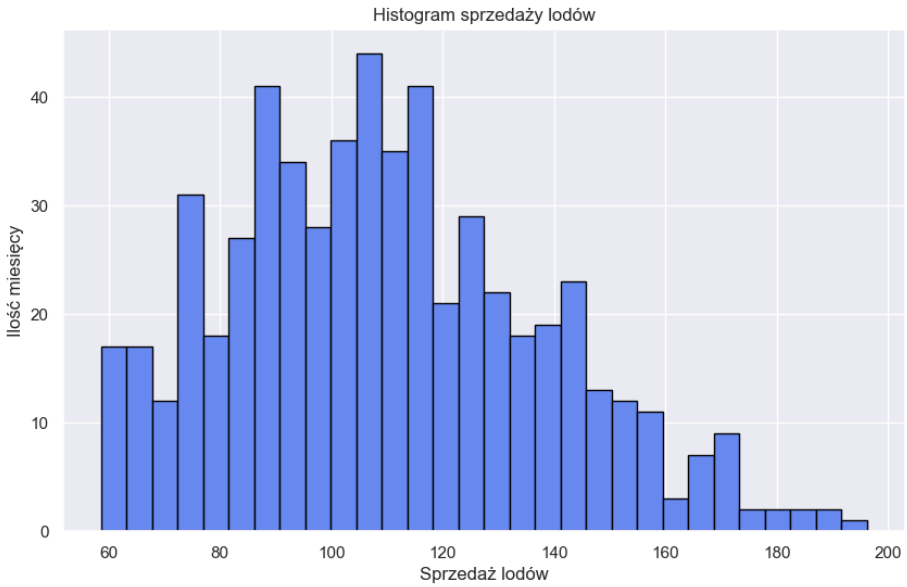
Poniższy wykres przedstawia ilość sprzedanych lodów dla każdego miesiąca.



Rysunek 1: Ilość sprzedanych lodów.

Na powyższym wykresie zauważalna jest sezonowość i powtarzalność trendu w sprzedaży lodów.

Przedstawmy dodatkowo histogram danych.



Rysunek 2: Histogram sprzedaży lodów.

Z wykresu wynika, że większość miesięcy ma sprzedaż w zakresie 80-140 jednostek. Rozkład jest asymetryczny, prawostrannie skośny, co świadczy o znacznie rzadszym pojawianiu się wysokich ilości sprzedanych lodów.

Sprzedaż ma swoje nietypowe wartości. Jest to związane z sezonowoscią danych. Znacznie więcej lodów sprzedawanych jest w miesiące cieplejsze, niż chłodniejsze.

1.4 Podstawowe definicje

1.4.1 Szereg czasowy stacjonarny w słabym sensie

Szereg czasowy X_t jest stacjonarny w słabym sensie, jeśli jego funkcja wartości oczekiwanej jest stała dla każdego t, a funkcja autokowariancji jest skończona i zależy tylko od przesunięcia h:

$$\mu(t) = \mathbb{E}[X_t] = const \quad \forall t \in \mathbb{Z}$$
$$\gamma(t, t+h) = \mathrm{Cov}(X_t, X_{t+h}) = \gamma(h) \quad \forall t \in \mathbb{Z}$$

1.4.2 Model ARMA (autoregressive moving average

Szereg czasowy $\{X_t\}_{t\in\mathbb{Z}}$ jest szeregiem ARMA(p,q) jeśli jest stacjonarny w słabym sensie i dla każdego t spełnia równanie:

$$X_t - \phi_1 X_{t-1} - ... - \phi_p X_{t-p} = Z_t + \theta_1 Z_{t-1} + ... + \theta_q Z_{t-q},$$

gdzie $\{Z_t\}_{t\in\mathbb{Z}} \sim WN(0, \sigma^2)$

oraz wielomiany

$$\phi(z) = 1 - \phi_1 z - ... - \phi_p z^p$$
$$\theta = 1 + \theta_1 z + ... + \theta_q z^q$$

nie mają wspólnych pierwiastków.

1.4.3 Przyczynowość szeregu czasowego

Model ARMA(p,q) jest przyczynowy, jeśli następujący warunek jest spełniony:

$$\phi = 1 + \phi_1 z + ... + \phi_p z^p \neq 0$$

dla wszystkich $|z| \leqslant 1$. To oznacza, że każdy pierwiastek wielomianu $\phi(z)$ spełnia warunek $|z| > 1$.

1.4.4 Odwracalność szeregu czasowego

Model ARMA(p,q) jest odwracalny, jeśli następujący warunek jest spełniony:

$$\theta = 1 + \theta_1 z + ... + \theta_q z^q \neq 0$$

dla wszystkich $|z| \leqslant 1$. To oznacza, że każdy pierwiastek wielomianu $\theta(z)$ spełnia warunek $|z| > 1$.

1.4.5 Autokowariancja

Funkcja autokowariancji (ACVF) rzędu k dla szeregu słabo stacjonarnego to funkcja $\gamma(k)$ określona wzorem:

$$\gamma = \mathrm{Cov}(X_{t+h}, X_t).$$

Próbkowa funkcja autokowariancji (Sample ACVF) jest estymatorem funkcji autokowariancji o wzorze:

$$\hat{\gamma} = \frac{1}{n}\sum_{t=1}^{n-h}(X_{t+h} - \bar{X})(X_t - \bar{X}),$$

gdzie $\bar{X}$ jest średnią próbkową.

1.4.6 Autokorelacja

Funkcja autokorelacji (ACF) rzędu k dla szeregu słabo stacjonarnego to funkcja $\rho(k)$ określona wzorem:

$$\rho(h) = \frac{\gamma(h)}{\gamma(0)},$$

gdzie $\rho(k)$ jest funkcją autokowariancji rzędu k.

Próbkowa funkcja autokorelacji (Sample ACF) jest estymatorem funkcji autokorelacji i jest określona wzorem:

$$\hat{\rho}(h) = \frac{\hat{\gamma(h)}}{\hat{\gamma}(0)},$$

gdzie $\hat{\gamma}(h)$ jest próbkową funkcją autokowariancji rzędu k.

1.4.7 Częściowa autokorelacja

Funkcja częściowej autokorelacji (PACF) rzędu k dla szeregu stacjonarnego to miara zależności liniowej między X_t i X_{t-k}, która uwzględnia wszystkie zmienne pośrednie. Empiryczna funkcja częściowej autokorelacji (Sample PACF) jest estymatorem funkcji częściowej autokorelacji i jest określona wzorem:

$$\hat{\phi}(h) = \hat{\Gamma}^{-1}(h)\hat{\gamma}(h),$$

gdzie $\hat{\Gamma}(h)$ jest macierzą kowariancji rzędu h, a $\hat{\gamma}(h)$ jest wektorem autokowariancji rzędu h.

2 Przygotowanie danych do analizy

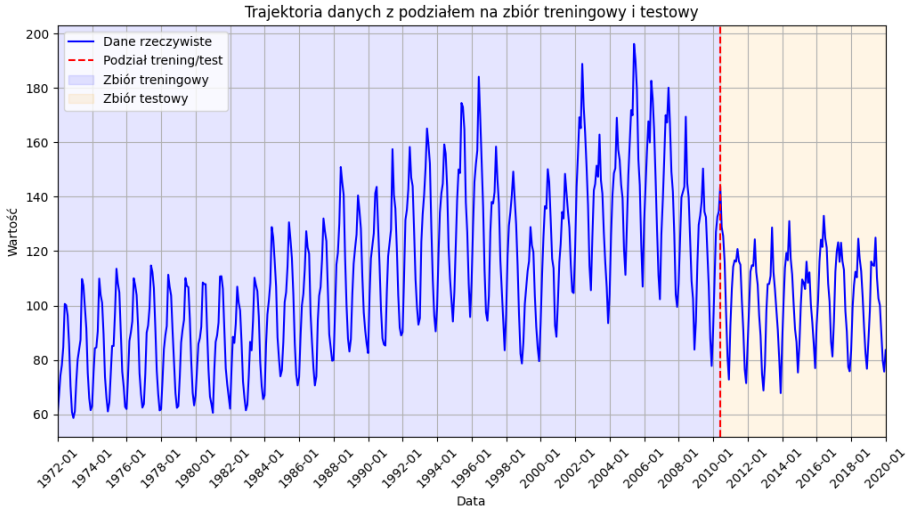
2.1 Zbadanie jakości danych

Aby sprawdzić czy w naszych danych mamy braki danych skorzystałyśmy z funckji `data.isnull().sum()`. Nie wykryto braków danych – wszystkie wartości są kompletne i nie zawierają pustych wpisów. Dodatkowo, po dokładnym sprawdzeniu, nie stwierdzono obecności nietypowych ani błędnych obserwacji, które mogłyby wskazywać na potencjalne anomalie lub błędy w próbkowaniu. Niektóre wartości można uznać za odstające, jednakże widoczny jest trend sezonowy, dlatego nie usuwamy ich. Dane są spójne i możemy wykorzystać je do dalszych analiz.

2.2 Wyodrębnienie obserwacji do zbioru testowego

Dane podzielono chronologicznie, aby zachować rzeczywiste zależności czasowe. 80% obserwacji trafiło do zbioru treningowego (461 wartości), który posłuży do budowy modelu, natomiast pozostałe 20% danych (116 wartości) stanowi zbiór testowy, wykorzystywany do oceny skuteczności prognoz. Wszystkie

transformacje są przeprowadzane wyłącznie na zbiorze treningowym, a następnie stosowane na zbiorze testowym.



Rysunek 3: Trajektoria danych z podziałem na zbiór treningowy i testowy.

2.3 Dekompozycja szeregu czasowego

Dekompozycja szeregu czasowego to proces rozkładu szeregu czasowego na jego podstawowe komponenty w celu lepszego zrozumienia jego struktury oraz przewidywania przyszłych wartości.

Klasyczną dekompozycje szeregu czasowego możemy zapisać jako:

$$Y_t = m(t) + s(t) + X_t.$$

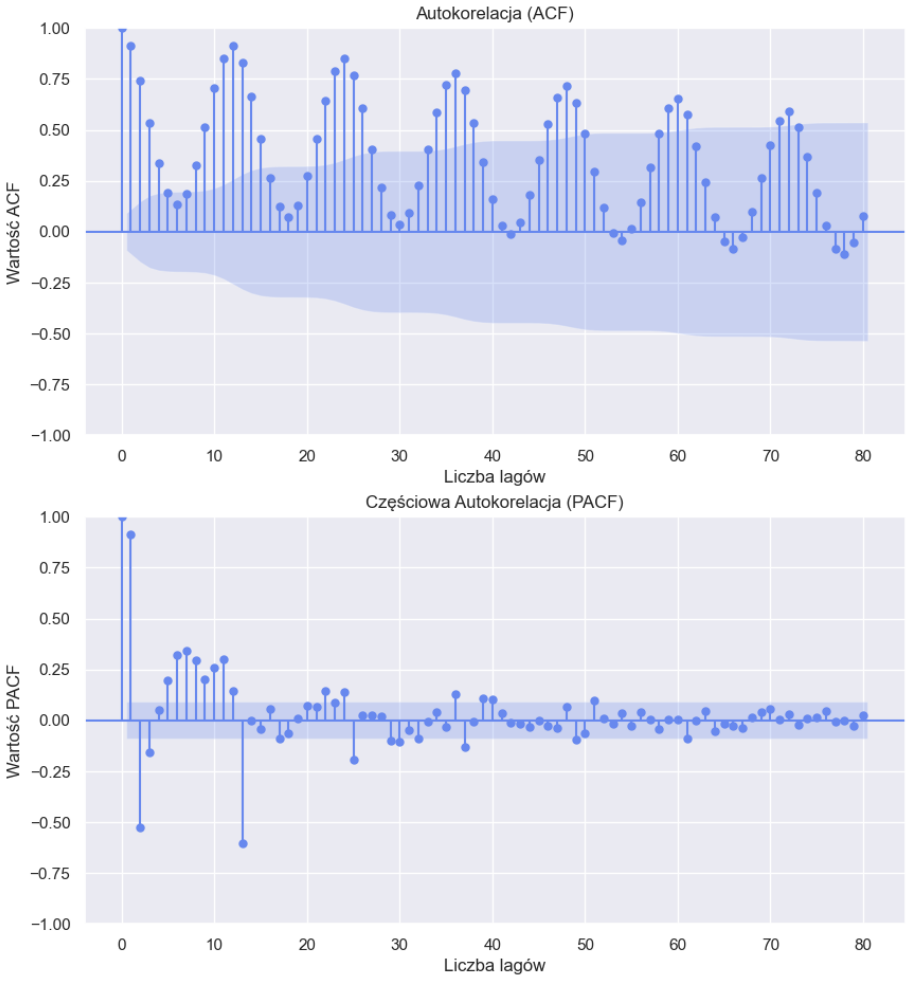
gdzie:

$m(t)$ - trend,
$s(t)$ - sezonowość,
X_t - szereg czasowy stacjonarny w słabym sensie.

2.3.1 Analiza ACF i PACF

Analiza autokorelacji (ACF) i częściowej autokorelacji (PACF) jest kluczowym krokiem w badaniu zależności czasowych w szeregu czasowym. ACF pokazuje, jak wartości szeregu są ze sobą powiązane w różnych opóźnieniach, co pozwala wykryć sezonowość i długoterminowe wzorce. Z kolei PACF uwzględnia tylko bezpośrednie zależności, co ułatwia określenie wpływu wcześniejszych wartości na bieżące obserwacje.

Aby zwizualizować autokorelacje i częściową autokorelacje skorzystałyśmy z funkcji `plot_acf` i `plot_pacf` z biblioteki `statsmodels`.



Rysunek 4: Autokorelacja i częściowa autokorelacja szeregu czasowego.

Na powyższym wykresie ACF wykazuje wyraźne okresowe wzorce, co sugeruje obecność sezonowości, natomiast PACF pokazuje wyraźne zależności dla początkowych lagów, co wskazuje na silne powiązania między kolejnymi obserwacjami. Utrzymująca się wysoka autokorelacja sugeruje, że szereg może wymagać różnicowania w celu osiągnięcia stacjonarności.

2.3.2 Test Dickeya-Fullera dla surowych danych

Test Dickeya-Fullera (ADF – Augmented Dickey-Fuller Test) pozwala sprawdzić, czy dany szereg czasowy zawiera jednostkowy pierwiastek, co sugeruje jego niestacjonarność. W przeciwnym razie możemy uznać, że proces jest stacjonarny, co umożliwia jego dalsze modelowanie przy użyciu takich metod jak ARMA.

Test ADF pozwala zweryfikować następujące hipotezy:

- H_0 - szereg zawiera jednostkowy pierwiastek (niestacjonarność szeregu)
- H_1 - szereg jest stacjonarny

Za pomocą funkcji `adfuller` z biblioteki `statsmodels.tsa.stattools` w Pythonie sprawdzamy wartości statystyki testowej i p-wartości. Jeśli p-wartość w wynikach testu jest mniejsza niż poziom istotności $\alpha = 0.05$, możemy odrzucić hipotezę zerową i uznać, że szereg jest stacjonarny. Jeśli p-wartość jest wyższa, sugeruje to, że szereg zawiera pierwiastek jednostkowy i jest niestacjonarny.

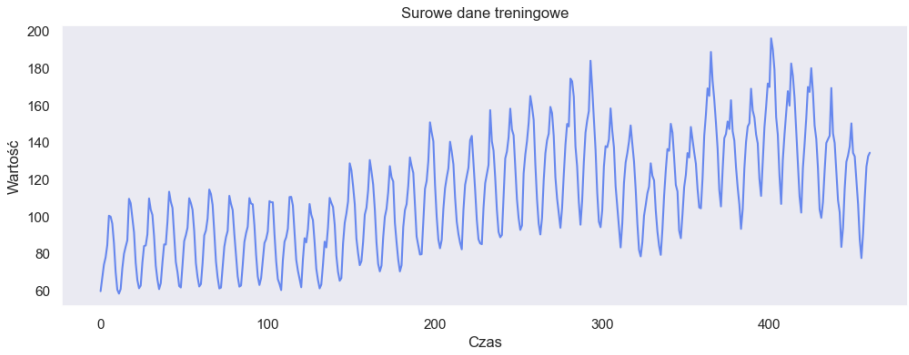
	Statystyka testowa	p-wartość
wartość	-1.58638	0.49043

Tabela 1: Wyniki testu Dickeya-Fullera dla szeregu czasowego.

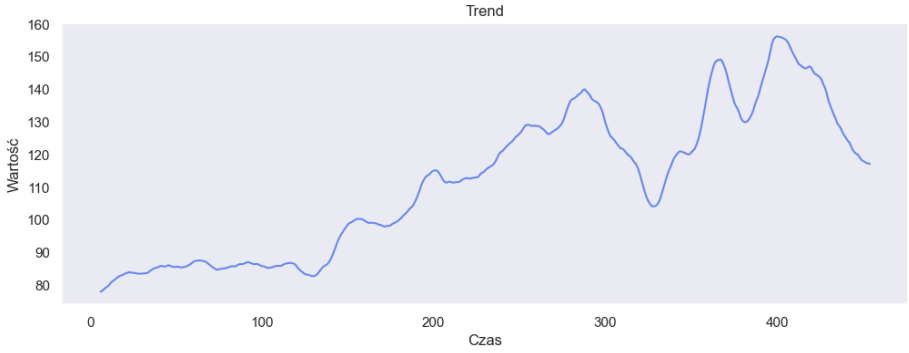
Na podstawie tabeli 1 możemy stwierdzić, że p-wartość jest znacznie większa niż poziom istotności równy 0.05, nie mamy podstaw do odrzucenia hipotezy zerowej. Oznacza to, że szereg czasowy posiada pierwiastek jednostkowy, czyli jest niestacjonarny. W związku z tym nie spełnia on warunków koniecznych do zastosowania klasycznych metod modelowania stacjonarnych procesów, takich jak ARMA.

2.3.3 Dekompozycja

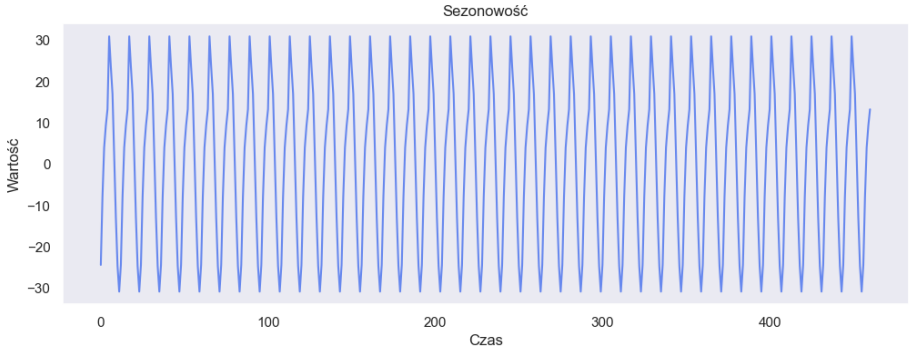
Funkcja `seasonal_decompose` z biblioteki `statsmodels.tsa.seasonal` służy do dekompozycji szeregu czasowego na jego podstawowe składowe: trend, sezonowość i resztę. Dzięki niej możemy badać strukturę szeregu czasowego, identyfikować sezonowość i trend oraz ocenić powtarzające się wzorce. Skorzystamy z niej aby, usunąć trend i sezonowość z naszych danych.



Rysunek 5: Szereg czasowy przed dekompozycją.



Rysunek 6: Trend dla szeregu czasowego.



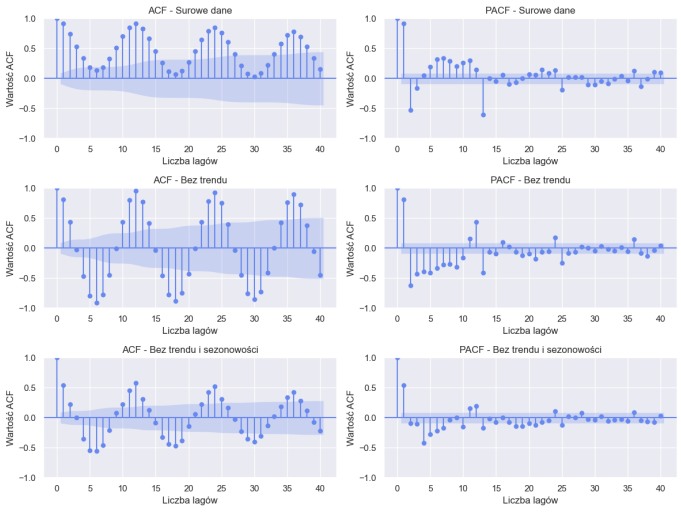
Rysunek 7: Sezonowość dla szeregu czasowego.

Rysunek 8: Szereg czasowy po usunięciu trendu i sezonowości.

Analizowany szereg czasowy przed dekompozycją wykazuje zarówno trend wzrostowy, jak i silną sezonowość, co jest widoczne na Rysunku 5. Po przeprowadzeniu dekompozycji (Rysunki 6 i 7) udało się wyodrębnić składową trendu, która pokazuje długoterminowy wzrost z okresami spadków, oraz wyraźnie powtarzający się wzorzec sezonowy.

Po usunięciu trendu i sezonowości (Rysunek 8) uzyskano szereg czasowy, który jest bardziej stacjonarny, co jest kluczowe dla dalszego modelowania.

2.3.4 Analiza ACF i PACF dla szeregu po dekompozycji



Rysunek 9: Porównanie ACF i PACF dla szeregu czasowego przed dekompozycją, bez trendu i po dekompozycji.

Po dokonaniu dekompozycji szeregu, wciąż zauważamy sezonowość na wykresie "ACF - Bez trendu i sezonowośći", sugeruje to, że szereg wciąż nie jest stacjonarny.

2.3.5 Test Dickeya-Fullera dla szeregu po dekompozycji

Przeprowadźmy jeszcze raz test Dickeya-Fullera dla szeregu po dekompozycji.

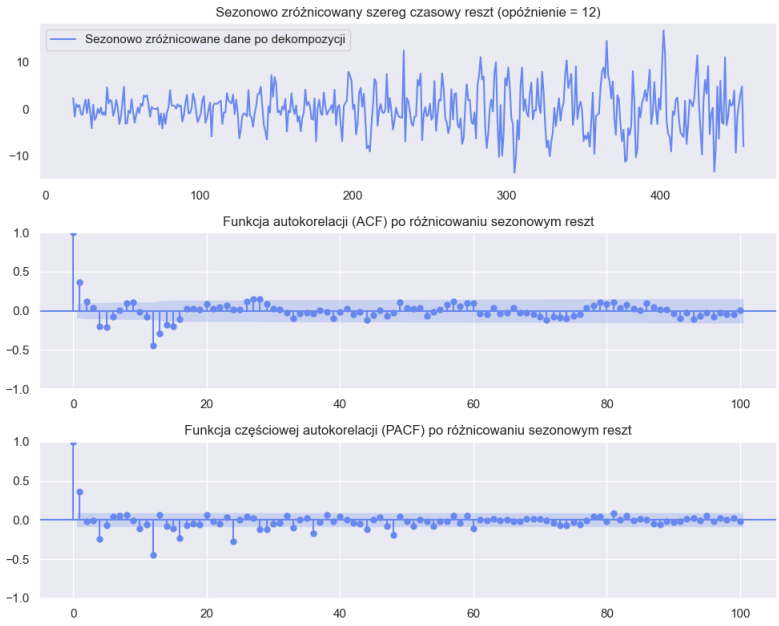
	Statystyka testowa	p-wartość
wartość	-8.30824	3.85128e-13

Tabela 2: Wyniki testu Dickeya-Fullera dla szeregu czasowego po dekompozycji.

Wyniki testu ADF z tabeli 2 wskazują, że p-wartość wynosi 3.85128e-13 co jest bliskie 0. Tak niska wartość p oznacza odrzucenie hipotezy zerowej o obecności pierwiastka jednostkowego, co sugeruje stacjonarność analizowanego szeregu czasowego i skutecznego usunięcią trendu i sezonowości przez dekompozycję, co z wykresu wyżej, wiemy, że jest nieprawdą.
Test Dickeya-Fullera ocenia stacjonarność szeregu czasowego, koncentrując się przede wszystkim na wykrywaniu niestacjonarności związanej z trendem, a nie sezonowością. W naszym przypadku okazuje się niewystarczający do pełnej oceny stacjonarności szeregu.

2.4 Różnicowanie

Aby móc dalej modelować szereg za pomocą metod wymagających stacjonarności, takich jak ARMA, musimy wykonać różnicowanie sezonowe. Różnicowanie szeregu czasowego można wykonać za pomocą funkcji `.diff()` dostępnej w bibliotece `pandas`.



Rysunek 10: ACF i PACF szeregu czasowego po różnicowaniu.

Pierwszy wykres pokazuje przebieg szeregu po zastosowaniu różnicowania sezonowego z okresem 12, co miało na celu usunięcie sezonowości. Drugi i trzeci wykres prezentują odpowiednio funkcję ACF i PACF, które pozwalają ocenić strukturę zależności po przeprowadzonej transformacji. Wartości autokorelacji w ACF znacznie zmalały oraz nie ma widocznej sezonowości, co sugeruje, że usunięto długoterminowe zależności sezonowe. Szereg po transformacji jest stacjonarny, co pozwala na zastosowanie modeli takich jak ARMA do dalszej analizy.

3 Modelowanie danych przy pomocy ARMA

3.1 Dobranie rzędu modelu

Aby odpowiednio dopasować rzędy p i q w modelu ARMA, konieczne jest zastosowanie odpowiednich metod oceny jakości dopasowania modelu. W tym celu wykorzystujemy kryteria informacyjne, które pozwalają na wybór optymalnej liczby opóźnień autoregresyjnych (p) oraz opóźnień składnika średniej ruchomej (q), minimalizując jednocześnie złożoność modelu.

3.1.1 Kryteria informacyjne

Do określenia rzędu modelu ARMA posłużyłyśmy się kryteriami informacyjnymi AIC i BIC.

- Kryterium informacyjne AIC (Akaike Information Criterion)

  Kryterium AIC jest określone wzorem:

  $$AIC = 2k - 2\ln(\hat{L}),$$

  gdzie:

  k - liczba parametrów modelu,
  $\hat{L}$ - maksymalizowana funkcja wiarygodności.

  Wybierając model, należy kierować się zasadą, że niższa wartość AIC wskazuje na lepsze dopasowanie.

- Kryterium informacyjne BIC (Bayesian Information Criterion)

  Kryterium BIC jest określone wzorem:

  $$BIC = k\ln(n) - 2\ln(\hat{L}),$$

  gdzie:

  n - liczba obserwacji w badanym szeregu czasowym,
  k - liczba parametrów modelu,
  $\hat{L}$ - maksymalizowana funkcja wiarygodności.

  Tak jak w przypadku AIC, niższa wartość BIC wskazuje na lepszy model.

- Kryterium informacyjne HQIC (Hannan-Quinn Information Criterion)

  Kryterium HQIC jest określone wzorem:

  $$HQIC = 2k\ln(\ln(n)) - 2\ln(\hat{L}),$$

  gdzie:

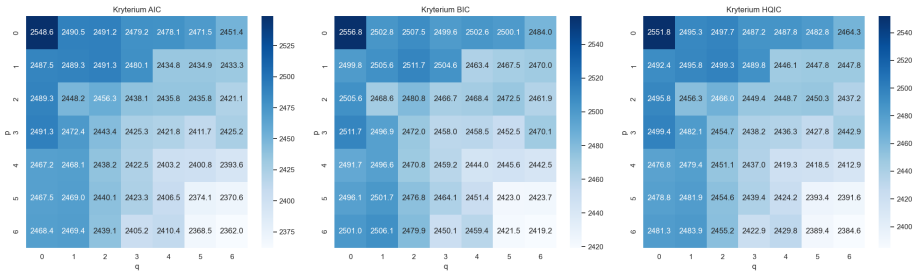
  k - liczba parametrów modelu,
  n - liczba obserwacji,

$\hat{L}$ - maksymalizowana funkcja wiarygodności.

W tym kryterium również najniższa wartość HQIC wskazuje na lepszy model.

Aby wyznaczyczyć wartości kryteriów informacyjnych możemy skorzystać z funckcji `aic`, `bic` i `hqic` z biblioteki `statsmodels`.

Wartości kryteriów informacyjnych dla naszych danych, dla $p, q \in \{1, 2, ..., 6\}$, przedstawia wykres poniżej:



Rysunek 11: Wartości kryteriów innformacyjnych dla modeli ARMA(p,q).

W przypadku wszystkich trzech kryteriów, najniższe wartości przypadają na parametry $p = 6$ i $q = 6$. Oznacza to, że najlepszym modelem dla naszych danych jest model ARMA(6,6).

3.2 Estymacja parametrów modelu

W celu estymacji parametrów naszego modelu korzystamy z wbudowanej funkcji ARIMA z pakietu statsmodels. Wykorzystuje ona metodę największej wiarygodności, która polega na maskymalizacji funkcji wiarygodności:

$$\mathcal{L}(\theta) = \prod_{t=1}^{n} f(X_t; \theta),$$

gdzie θ jest wektorem parametrów, a f jest gęstością rozkładu prawdopodobieństwa.

Wartości wyestymowanych parametrów przedstawiają tabele poniżej:

	ϕ_1	ϕ_2	ϕ_3	ϕ_4	ϕ_5	ϕ_6
wartość	-0.0011	-0.2277	0.8262	-0.2658	-0.5915	0.3442

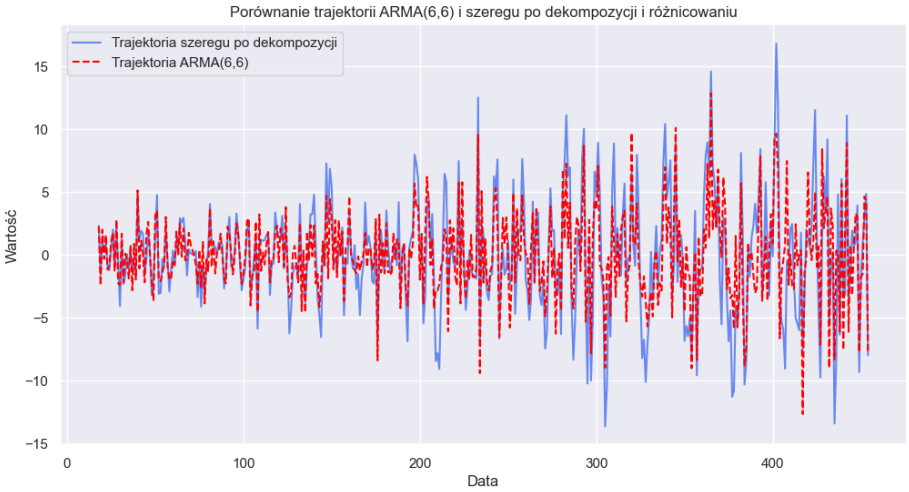
Tabela 3: Wyniki estymacji parametrów modelu AR(6).

	θ_1	θ_2	θ_3	θ_4	θ_5	θ_6
wartość	0.3241	0.5766	-0.8413	0.0832	0.7642	-0.6794

Tabela 4: Wyniki estymacji parametrów modelu MA(6).

4 Ocena dopasowania modelu

Po dokonaniu dekompozycji i przeprowadzeniu różnicowania sezonowego, istotnym etapem analizy jest ocena, na ile model ARMA(6,6) poprawnie odwzorowuje rzeczywisty przebieg szeregu czasowego. Wizualna analiza pozwala zweryfikować, czy model właściwie uchwycił dynamikę danych, odwzorował ich zmienność oraz czy nie występują systematyczne rozbieżności pomiędzy modelem a rzeczywistymi wartościami szeregu. W tym celu porównamy trajektorię rzeczywistych wartości szeregu czasowego po dekompozycji z wartościami wygenerowanymi przez model ARMA(6,6).

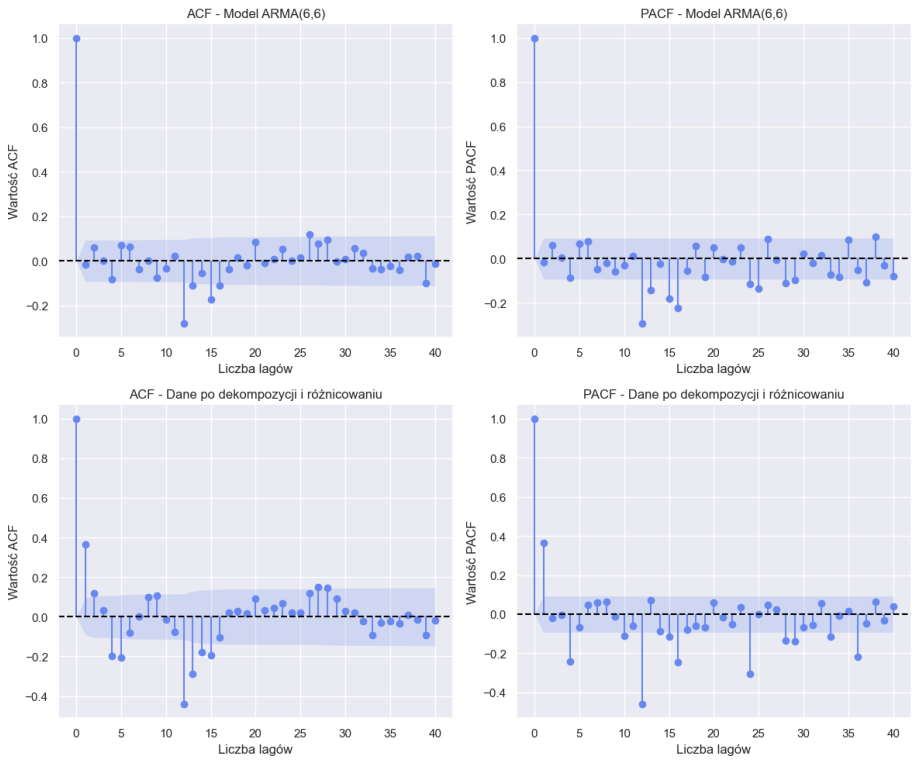


Rysunek 12: Porównanie przebiegu szeregu czasowego po dekompozycji i sezonowym różnicowaniu z trajektorią modelu ARMA(6,6).

Powyższy wykres pokazuje, że model ARMA(6,6) dobrze odwzorowuje ogólny przebieg szeregu. Widoczne są podobne wzorce oraz zgodność w zakresie zmienności.

4.1 Przedziały ufności dla ACF i PACF

Poniższe wykresy przedstawiają, jak dobrze model odzwierciedla wzorce autokorelacyjne. Dzięki nim możemy zobaczyć, jak bardzo model ARMA(6,6) dopasowuje się do wartości szeregu czasowego po dekompozycji i różnicowaniu.

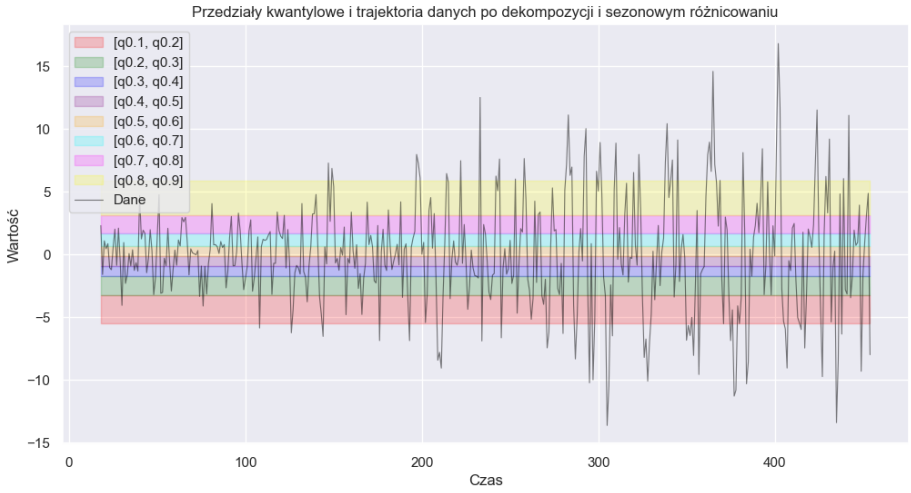


Rysunek 13: Porównanie ACF i PACF dla szeregu po dekompozycji i sezonowym różnicowaniu z modelem ARMA(6,6).

Model ARMA(6,6) dobrze odwzorowuje strukturę szeregu czasowego po dekompozycji i różnicowaniu, co widać po zmniejszeniu autokorelacji w ACF i PACF. Niskie wartości autokorelacji dla większych lagów wskazują, że reszty modelu nie wykazują wyraźnych zależności, co oznacza, że model został dobrze dopasowany do danych. Transformacje usunęły trend i sezonowość, a model skutecznie uchwycił pozostałe zależności w danych.

4.2 Porównanie linii kwantylowych z trajektorią

Analiza kwantylowa szeregu czasowego pomaga zrozumieć, jak wartości zmieniają się w czasie po usunięciu trendu i sezonowości. Przedziały kwantylowe pokazują zakresy, w których z dużym prawdopodobieństwem znajdują się obserwacje, co pozwala lepiej ocenić zmienność danych i ich rozkład.

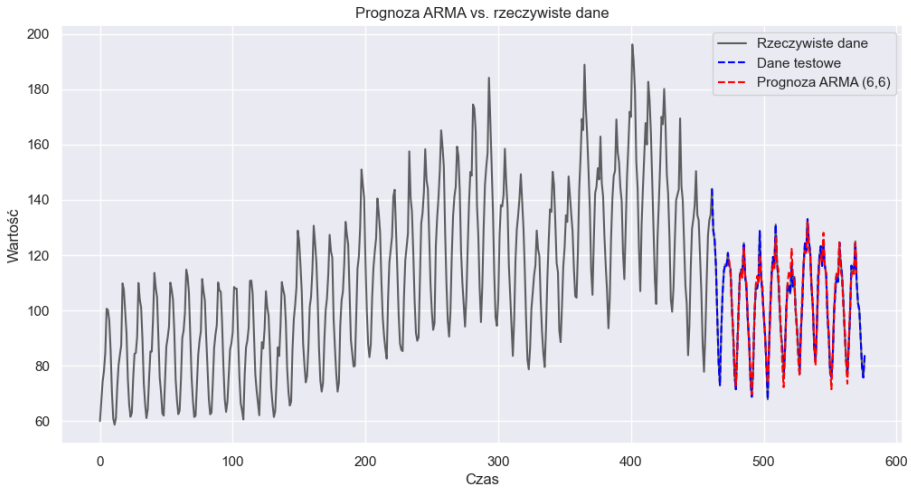


Rysunek 14: Porównanie szeregu czasowego po dekompozycji i sezonowym różnicowaniu z liniami kwantylowymi.

Na początku szeregu większość wartości utrzymuje się w środkowych zakresach kwantylowych, co sugeruje stabilność danych. W dalszej części widać większe wahania i częstsze przekraczanie przedziałów kwantylowych, co może oznaczać rosnącą zmienność lub niespodziewane zmiany w danych. Gdy wartości wychodzą poza najwyższe i najniższe kwantyle, może to wskazywać na okresy większej nieprzewidywalności lub występowanie nietypowych zdarzeń.

4.3 Prognoza dla przyszłych obserwacji

Przewidywanie przyszłych obserwacji szeregu czasowego polega na prognozowaniu przyszłej trajektorii na podstawie wzorców historycznych. W tej części przewidzimy wartości na podstawie danych historycznych i ocenimy skuteczność modelu poprzez porównanie wyników prognozy z rzeczywistymi danymi testowymi.
W tym celu skorzystamy z funkcji `model fit.forecast` z pakietu `statsmodels`.



Rysunek 15: Wykres dopasownia i prognozy modelu ARMA(6,6).

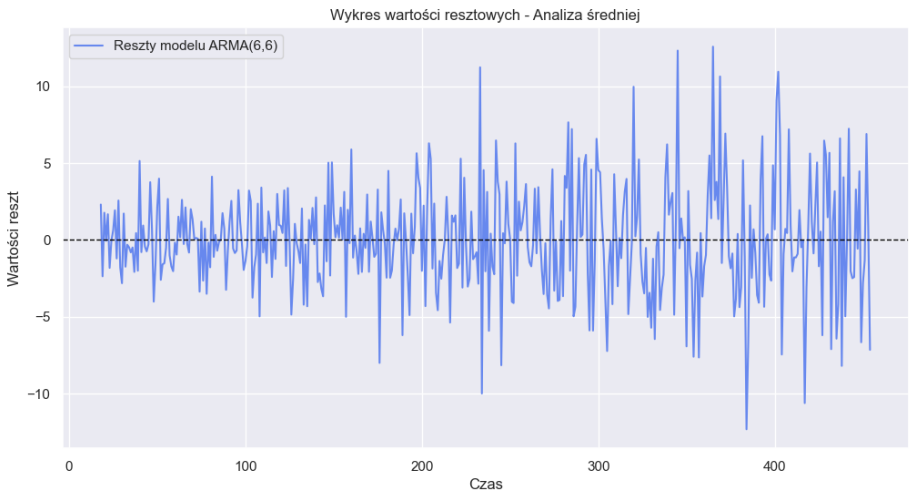
Z analizy wykresu wynika, że model ARMA(6,6) dobrze odwzorowuje dane testowe, szczególnie w zakresie sezonowości. Prognoza pokrywa się z trajektoriądanych testowych, co sugeruje, że model poprawnie identyfikuje ich wzorce.

5 Weryfikacja założeń dotyczących szumu

5.1 Założenie dotyczące średniej

5.1.1 Analiza wykresu wartości resztowych

Na wykresie poniżej przedstawiamy wykres wartości resztowych modelu ARMA(6,6).



Rysunek 16: Wykres wartości resztowych modelu ARMA(6,6) - analiza średniej.

Reszty oscylują wokół zera i nie wykazują wyraźnego trendu, co sugeruje, że model dobrze odwzorowuje średnią procesu.

5.1.2 Test t-Studenta

W celu sprawdzenia wniosku z poprzego wykresu, że wartości resztowe oscylują wokół 0, przeprowadzimy test t-Studenta dla średniej, który polega na sprawdzeniu hipotez:

- Hipoteza zerowa H_0: $\mu = 0$, czyli średnia wartości reszt jest równa zero,
- Hipoteza apternatywna $H_1 : \mu \neq 0$, czyli średnia wartości reszt różni się od zera.

Test został przeprowadzony na poziomie ufności $\alpha = 0.05$
W teście używamy statystyki testowej:

$$t = \frac{\bar{X} - \mu}{\frac{S}{\sqrt{n}}},$$

gdzie $\bar{X}$ to śrenia z próby, S to odchylenie standardowe z próby, a n to liczebność próby.

Do wykonania testu skorzystamy z funkcji `ttest_1samp` z pakietu `scipy.stats`. Wyniki testu przedstawione zostały w tabeli poniżej:

	Statystyka testowa	p-wartość
wartość	0.1117	0.9110

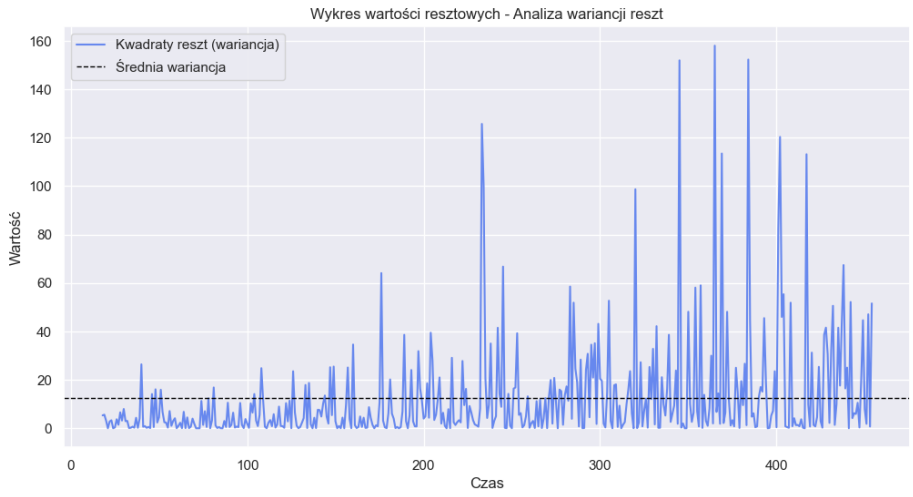
Tabela 5: Wyniki testu t-Studenta dla średniej szumu modelu ARMA (6,6).

Widzimy, że p-wartość jest większa od poziomu istotności, więc nie ma podstaw do odrzucenia hipotezy zerowej. Potwierdza to wniosek z wykresu. Średnia wartości resztowych jest równa 0.

5.2 Założenie dotyczące wariancji

5.2.1 Analiza wykresu wartości resztowych

Na wykresie poniżej przedstawiono wykres kwadratów reszt. Wykres taki pozwala ocenić, czy wariancja reszt jest stała, co stanowi podstawowe założenie homoskedastyczności.



Rysunek 17: Wykres wartości resztowych modelu ARMA(6,6) - analiza wariancji.

Na wykresie widzimy, że wartości kwadratów reszt są niskie w pierwszej fazie, a następnie zwiększają się, osiągając znacznie większe wartości. Oznacza to, że wariancja nie jest stała.

5.2.2 Modified Levene Test

W celu zweryfikowania hipotezy, że wariancja nie jest stała, przeprowadzimy test Levene. Polega on na sprawdzeniu hipotez:

- Hipoteza zerowa $H_0 : \sigma_1^2 = \sigma_2^2$, czyli wariancje w dwóch różnych grupach są równe.
- Hipoteza alternatywna $H_1 : \sigma_1^2 \neq \sigma_2^2$, czyli wariancje w dwóch grupach nie są równe.

Test został przeprowadzony na poziomie ufności $\alpha = 0.05$

W teście używamy statystyki testowej:

$$W = \frac{(N-k)}{(k-1)} \frac{\sum_{i=1}^{k} N_i (\bar{Z}_i - \bar{Z})^2}{\sum_{i=1}^{k} \sum_{j=1}^{N_i} (Z_{ij} - \bar{Z}_i)},$$

gdzie N jest liczebnością próby, k jest liczbą grup, N_i jest liczebością i-tej grupy, $\bar{Z}_i$ to średnia z i-tej grupy, $\bar{Z}$ średnią ze wszystkich grup, a Z_{ij} j-tą obserwacją i-tej grupy.

Do wykonania tego testu korzystamy z funkcji `levene` z pakietu `scipy.stats`. Przeprowadza ona zmodyfikowany test Levene'a, który zamiast brać pod uwagę różne próbki porównuje wariancje w różnych przedziałach czasowych jednej próbki. W tym cely dzielimy próbkę na połowę i sprawdzamy, czy wariancja w tych przedziałach jest stała.

Wyniki testu przedstawione są w tabeli poniżej:

	Statystyka testowa	p-wartość
wartość	39.5911	7.65 e^{-10}

Tabela 6: Wyniki testu Modified Levene Test dla wariancji szumu modelu ARMA (6,6).

P-wartość jest mniejsza od poziomu istotności, co daje podstawy do odrzucenia hipotezy zerowej na rzecz hipotezy alternatywnej. Test potwierdził, że wariancja nie jest stała.

5.2.3 ARCH Test

Ostatnią metodą na sprawdzenie wariancji wartości resztowych jest ARCH test. Test polega na sprawdzeniu hipotez:

- Hipoteza zerowa $H_0 : \alpha_1, \alpha_2, ..., \alpha_p = 0$, czyli współczynniki regresji liniowej są równe zeru, mamy brak efektu ARCH (stała wariancja).

- Hipoteza alternatywna $H_1 : \exists_{i \in \{1,2,\ldots,p\}} \alpha_i \neq 0$, czyli przynajmniej jeden współczynnik regresji liniowej nie jest równy zeru, mamy efekt ARCH (wariancja nie jest stała).

Aby stwierdzić, czy mamy efekt ARCH skorzystamy z funkcji `arch model` z pakietu `arch` i obliczymy wartości współczynników regresji liniowej.
Wartości współczynników α przedstawione zostały w tabeli poniżej:

	α_1	α_2	α_3	α_4	α_5	α_6
wartość	0.04	0	$4.75\ e^{-13}$	$5.50\ e^{-2}$	$1.77\ e^{-11}$	$2.63\ e^{-12}$

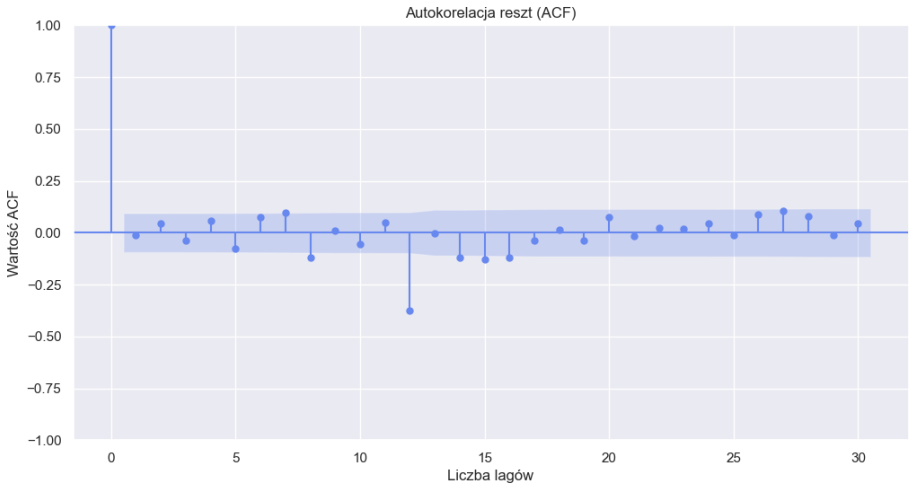
Tabela 7: Wyniki testu ARCH Test dla wariancji szumu modelu ARMA (6,6).

Widzimy, że istnieją współczynniki α, które nie są równe zeru. Oznacza to, że mamy efekt ARCH. Wariancja nie jest stała.

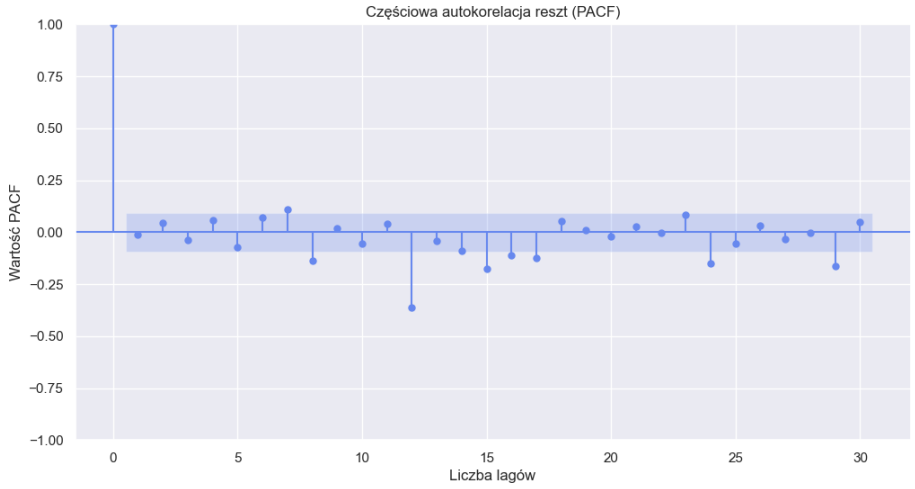
5.3 Założenie dotyczące niezależności

5.3.1 Wykresy ACF i PACF dla wartości resztowych

W celu sprawdzenia niezależności poniżej przedstawione zostały wykresy ACF i PACF dla wartości resztowych modelu ARMA(6,7):



Rysunek 18: Wykres ACF wartości resztowych modelu ARMA(6,6) - analiza niezależności.



Rysunek 19: Wykres PACF wartości resztowych modelu ARMA(6,6) - analiza niezależności.

Jeśli punkty na wykresie ACF i PACF są ograniczone granicami błędu (oznaczonym niebieskim obszarem), to model jest niezależny.
W naszym przypadku, większość reszt mieści się w granicach błędu. Model jest bliski niezależności, jednak nie mamy pewności co do tego.

5.3.2 Test Ljunga-Boxa

W celu bardziej dokładnego sprawdzenia niezależności, wykonamy test Ljunga-Boxa. Jest to test statystyczny wykorzystywany do sprawdzania, czy w szeregu czasowym występuje autokorelacja reszt. Test polega na sprawdzeniu hipotez:

- Hipoteza zerowa $H_0 : \rho_1 = \rho_2 = ... = \rho_m = 0$, czyli wszystkie autokorelacje do rzędu m są równe zeru, czyli brak autokorelacji
- Hipoteza alternatywna $H_1 : \exists_{k \in \{1,2,...,n\}} \rho_k \neq 0$, czyli przynajmniej jedna z autokorelacji jest różna od zera, czyli w danych występuje autokorelacja

Test został przeprowadzony na poziomie ufności $\alpha = 0.05$ W teście używamy statystyki testowej:

$$Q = n(n+2) \sum_{k=1}^{m} \frac{\hat{\rho_k}^2}{n-k},$$

gdzie n to liczebność próby, $\hat{p_k}$ to k-ty współczynnik autokorelacji, a m liczbą opóźnień.

Do wykonania testu skorzystałyśmy z funkcji `acorr_ljungbox` z pakietu `statsmodels`. Wyniki testu przedstawione są w tabeli poniżej:

Opóźnienia	P-wartość
lag 1	$4.36\ e^{-2}$
lag 3	$6.80\ e^{-3}$
lag 5	$3.28\ e^{-4}$
lag 10	$2.51\ e^{-6}$
lag 20	$2.21\ e^{-14}$
lag 30	$2.41\ e^{-13}$

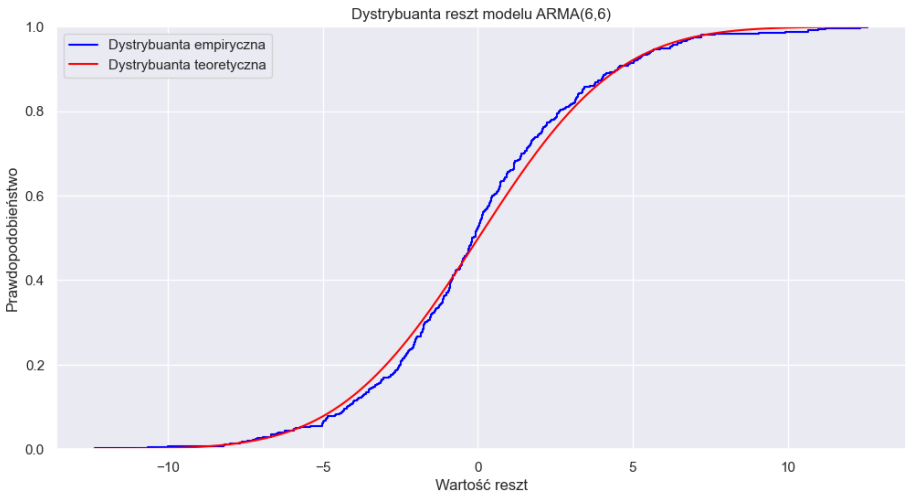
Tabela 8: Wyniki testu Ljunga-Boxa dla modelu ARMA (6,6).

Dla wszystkich opóźnień z tabeli, p-wartości są mniejsze od poziomu istotności, co oznacza, że odrzucamy hipotezę zerową. Szum nie jest niezależny.

5.4 Założenie dotyczące normalności rozkładu

5.4.1 Analiza dystrybuanty wartości resztowych

W celu sprawdzenia normalności rozkładu wartości resztowych porównany wykres dystrybuanty empirycznej z dystrybuantą teoretyczną, będącą dystrybuwantą rozkładu normalnego o średniej i wariancji równym średniej i wariancji wartości resztowych.
Skorzystamy w tym celu z funkcji `ecdfplot` z pakietu `seaborn` i funkcji `norm.cdf` z pakietu `scipy.stats`.

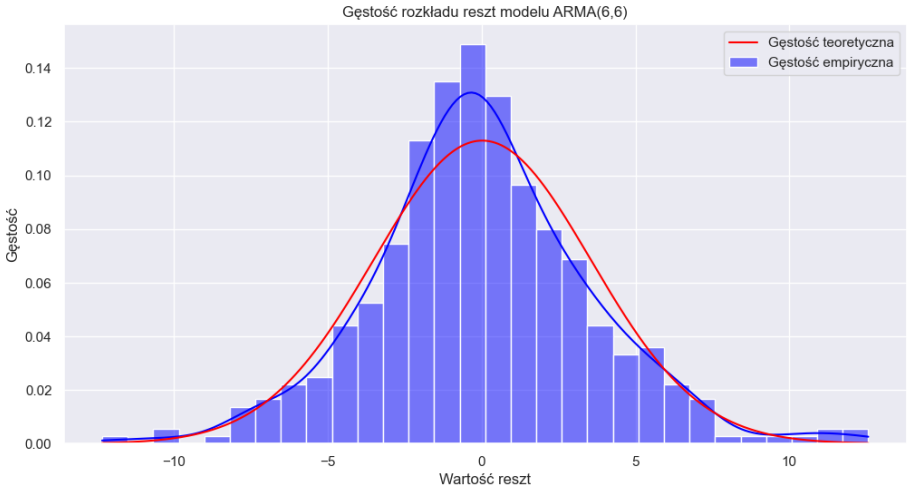


Rysunek 20: Wykres dystrybuanty wartości resztowych modelu ARMA(6,6).

Widzimy delikatne różnice między dystrybuantą empiryczną, a dystrybuantą rozkładu normalnego. Może to znaczyć, że wartości resztowe nie mają rozkładu normalnego.

5.4.2 Analiza gęstości wartości resztowych

Dla lepszego zbadania normalności rozkładu wartości resztowych porównamy ich gęstość empiryczną z gęstościa rozkładu normalnego o średniej i wariancji równym średniej i wariancji wartości resztowych.
Skorzystamy w tym celu z funkcji `histplot` z pakietu `seaborn` i funkcji `norm.pdf` z pakietu `scipy.stats`.

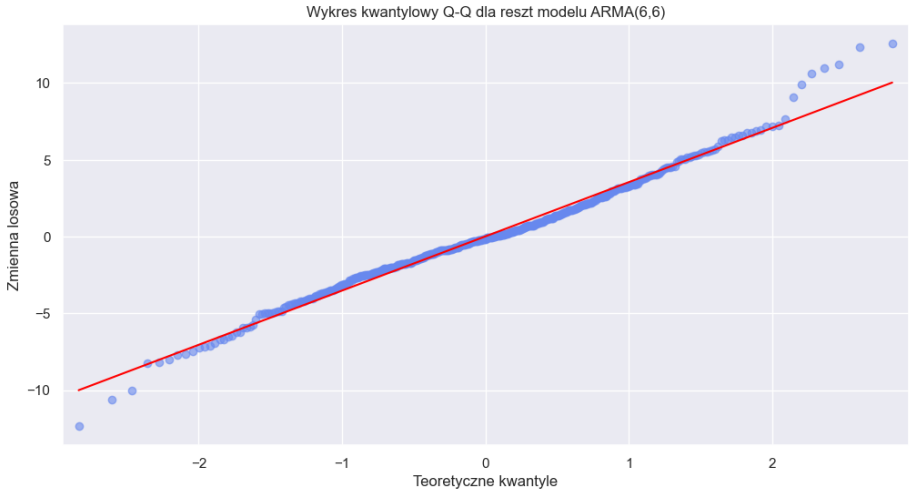


Rysunek 21: Wykres gęstości wartości resztowych modelu ARMA(6,6).

Tak jak w przypadku dystrybuanty, wykres gęstości empirycznej różni się od rozkładu normalnego, co ponownie potwierdza brak rozkładu normalnego w przypadku wartości resztowych.

5.4.3 Analiza wykresu kwantyl-kwantyl

W celu sprawdzenia normalności rozkładu przedstawimy również wykres kwantyl-kwantyl. Na wykresie tym porównywane są kwantyle wartości resztowych z kwantylami rozkładu normalnego o średniej i wariancji wartości resztowych.



Rysunek 22: Wykres Q-Q wartości resztowych modelu ARMA(6,6).

Nie wszystkie reszty przechodzą przez linię teoretyczną, co oznacza brak rozkładu normalnego.

5.4.4 Test Shapiro-Wilka

Ostatnią metodą sprawdzenia normalności rozkładu wartości resztowych będzie przeprowadzenie testu Shapiro-Wilka, który polega na sprawdzeniu hipotez:

- Hipoteza zerowa H_0 : wartości resztowe mają rozkład normany
- Hipoteza alternatywna H_1 : wartości resztowe nie mają rozkładu normalnego/

Test został przeprowadzony na poziomie ufności $\alpha = 0.05$ W teście używamy statystyki testowej:

$$W = \frac{(\sum_{i=1}^{n} a_i x_{(i)})^2}{\sum_{i=1}^{n}(x_i - \bar{x})},$$

gdzie $x_{(i)}$ to i-ta najmniejsza wartość resztowa, x_i to i-ta wartość resztowa, $\bar{x}$ jest średnią wartością resztową, a a_i są współczynnikami zależnymi od liczebności próby n.

W celu przeprowadzenia testu skorzystałyśmy z funkcji `shapiro` z pakietu `scipy.stats`. Otrzymane wyniki testu przedstawia tabela poniżej:

	Statystyka testowa	p-wartość
wartość	0.98677	0.00052

Tabela 9: Wyniki testu Shapiro Wilka dla normalności szumu modelu ARMA (6,6).

P-wartość jest mniejsza od poziomu istotności. Wnioski z poprzednich podpunktów są prawidłowe. Wartości resztowe nie mają rozkładu normalnego.

6 Podsumowanie

Celem naszej pracy była analiza danych sprzedaży lodów za pomocą modelu ARMA. Praca obejmowała przygotowanie danych, dekompozycję szeregu czasowego, testy stacjonarności oraz dobór optymalnego modelu.

Wyniki wykazały sezonowość oraz istnienie trendu danych, co wymagało zastosowania dekompozycji oraz różnicowania w celu osiągnięcia stacjonarności. Na podstawie kryteriów informacyjnych do naszych danych dopasowałyśmy model ARMA(6,6).

Ocena dopasowania modelu oraz prognozowanie przyszłych wartości potwierdziły, że ARMA(6,6) dobrze odwzorowuje wzorce sezonowe i zależności w danych, ale analiza wartości resztowych wykazała, że założenia dotyczące wariancji i normalności rozkładu nie zostały spełnione. Spełnione zostało jedynie założenie o średniej.

Podsumowując, odstępstwa od założeń mogą wpłynąć na jakość prognoz i sugerują konieczność dalszej optymalizacji modelu lub znalezienie innej metody, w celu lepszego uwzględnienia zmienności danych.