Komputerowa analiza szeregów czasowych 2024/2025

Analiza danych rzeczywistych przy pomocy modelu ARMA

Sprawozdanie

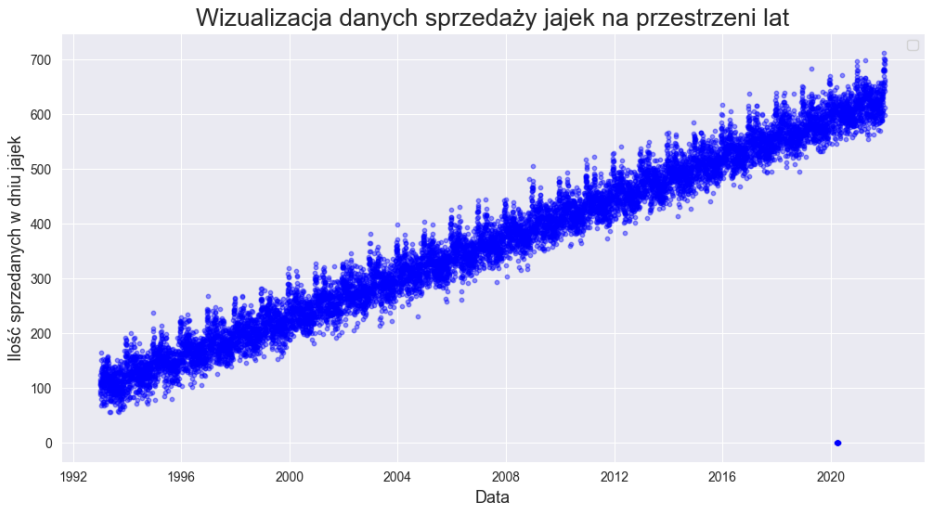
Karolina Bakalarz 276059
Amelia Bieda 275973

Spis treści

1. Wstęp

Celem pracy jest analiza danych o sprzedaży jajek w oparciu o model ARMA.

Wykorzystano dane z konkursu „Egg Sales Forecasting Challenge", obejmujące 30-letnią historię sprzedaży jajek w sklepie na Sri Lance. Dla każdego punktu danych wyróżnione są 2 informacje - ilość sprzedanych jajek w dniu oraz jaki to dzień (data). Zbiór obejmuje zakres dokładnie od 1 stycznia 1993 roku do 31 grudnia 2021 roku, stąd długość próby wynosi 10592. Dane te odzwierciedlają sezonowe wahania wynikające z tradycji i wydarzeń globalnych. Analiza tych danych została przeprowadzona w `Pythonie`, korzystając z biblioteki `statsmodels, pandas` i `matplotlib`.



Rysunek 1. Wykres punktowy badanych danych.

Procesem typu ARMA(p,q), gdzie p, $q \in \mathbb{N}$, nazywamy szereg czasowy $\{X_t\}_{t\in\mathbb{Z}}$, który dla każdego t można zapisać w formie

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

gdzie $\{Z_t\} \sim WN(0,\sigma^2)$ i wielomiany $\phi(z) = 1 - \phi_1 z - \cdots - \phi_p z^p$, $\theta(z) = 1 + \theta_1 z + \cdots + \theta_q z^q$ nie mają wspólnych pierwiastków.

2. Przygotowanie danych do analizy

2.1. Jakość danych

Zbiór analizowanych danych nie ma braków. Wszystkie wartości ilości sprzedanego produktu dzieli 1 dzień. Mamy 18 danych odstających. Są to dane sprzedaży z przełomu marca i kwietnia 2020 roku, w czasie początku pandemii koronawirusa. Sklepy były wtedy oczywiście zamknięte, stąd te wartości wynoszą zero.

Rysunek 2. Wykres punktowy danych odstających.

2.2. Dekompozycja szeregu czasowego

Klasyczną dekompozycją szeregu czasowego nazywamy jego reprezentację jako realizację procesu

$$X_t = m_t + s_t + Y_t,$$

gdzie
- X_t to dane pomiarowe,
- m_t to wielkość opisująca wolno zmienną funkcję - *trend*,
- s_t to periodycznie zmienna funkcja - *sezonowość*,
- Y_t to kompoment opisujący stacjonarny szereg czasowy, często nazywany *szumem*.

Funkcją autokorelacji ACF nazywamy

$$\rho_X(h) \equiv \frac{\gamma_X(h)}{\gamma_X(0)} = Corr(X_{t+h}, X_t)$$

dla stacjonarnego szeregu czasowego $\{X_t\}$, gdzie $\gamma(h) = Cov(X_{t+h}, X_t)$ jest funkcją autokowariancji tego szeregu.

Funkcją autokorelacji cząstkowej PACF dla procesu $\{X_t\}$ (model ARMA) jest funkcja $\alpha(\cdot)$ zdefiniowana równaniami

$$\alpha(0) = 1 \wedge \alpha(h) = \phi_{hh},\ h \geqslant 1,$$

gdzie ϕ_{hh} jest ostatnim wyrazem $\phi_h = \Gamma_h^{-1}\gamma_h$ dla $\Gamma_h = [\gamma(i-j)]_{i,j=1}^h$ oraz $\gamma_h = [\gamma(1),\gamma(2),\ldots,\gamma(h)]'$.

Celem dekompozycji szeregu czasowego jest oszacowanie i ekstrakcja deterministycznych części szeregu (trendu i sezonowości). Chcemy, aby pozostałe dane okazały się stacjonarnym procesem losowym. W takim przypadku możemy przystąpić do przewidywania przyszłego zachowania się szeregu.

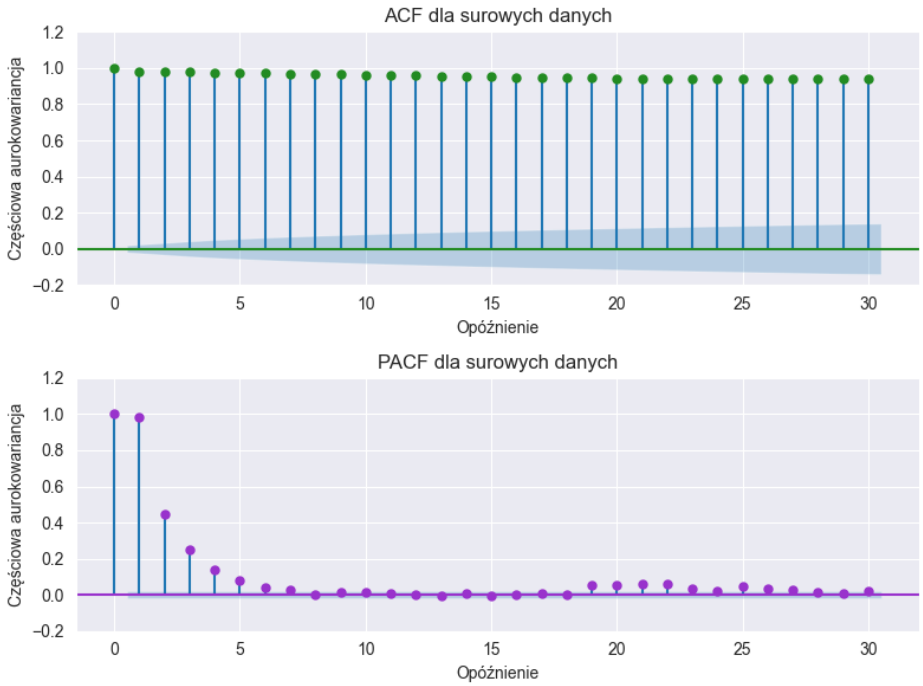
Na początku przeprowadziłyśmy test ADF (Augmented Dickey-Fuller Test) na poziomie istotności 5%. Test weryfikuje hipotezę o niestacjonarności

surowych danych. Statystykę testową porównuje się z wyznaczonymi wartościami krytycznymi. W tym przypadku statystyka testowa jest większa, niż krytyczna wartość dla wybranego poziomu istotności, co mówi o niestacjonarności szeregu czasowego. p-wartość wyniosła aż 0.479, więc nie odrzucamy hipotezy zerowej - szereg jest niestacjonarny.

Wartość krytyczna dla 5%	-2.8618138703454576
Statystyka testowa	-1.609
p-wartość	0.479

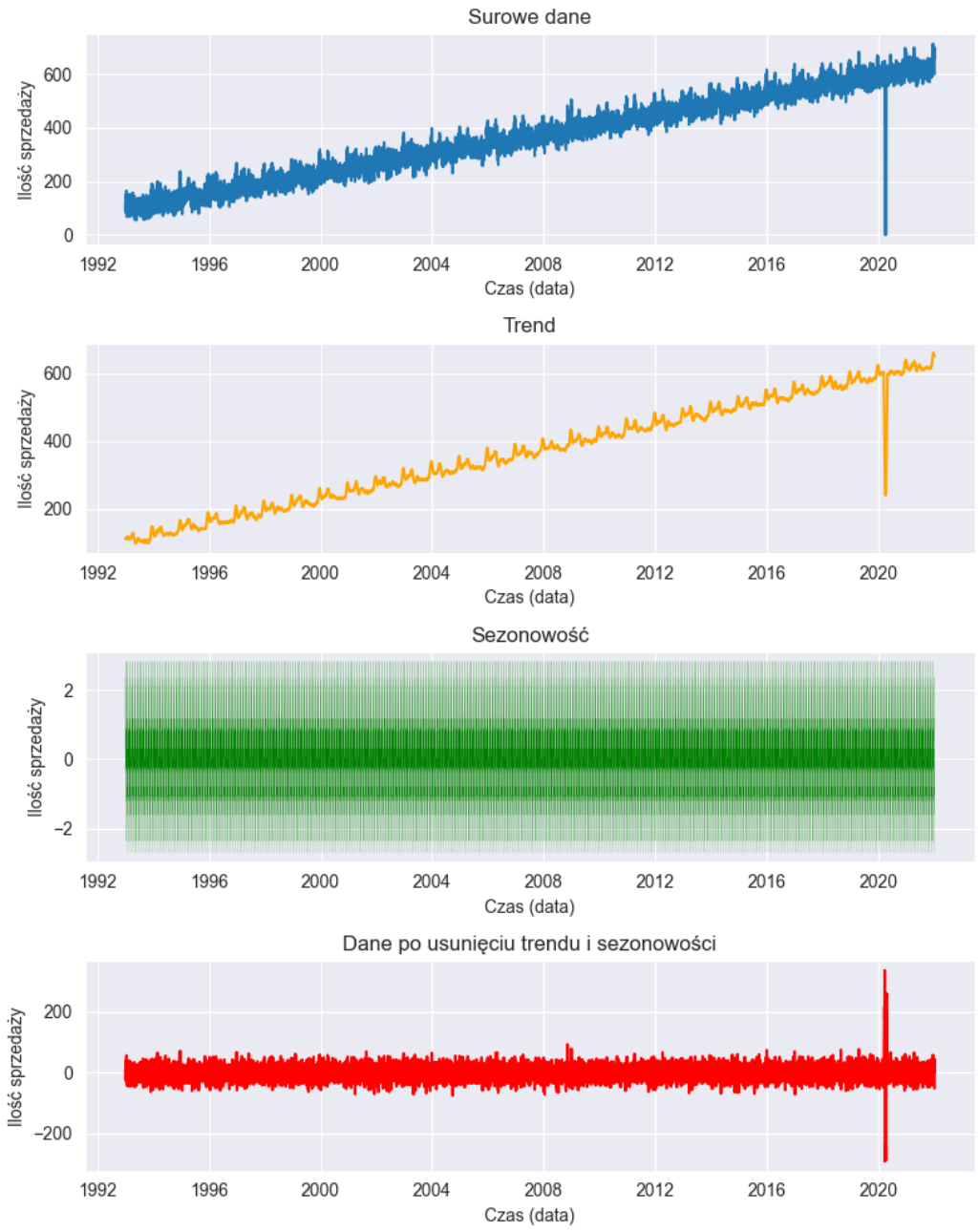
Tabela 1. Wyniki testu ADF dla surowych danych.

Przeanalizujmy wykresy autokorelacji ACF i częściowej autokorelacji PACF dla badanego zbioru danych. Obserwujemy bardzo silną autokorelację na wszystkich lagach, co sugeruje zarówno trend jak i sezonowość w naszych danych sprzedaży jajek. Dla funkcji częściowej autokorelacji pierwsze opóźnienia powoli spadają.



Rysunek 3. Wykresy ACF i PACF dla surowych danych.

Zidentyfikowaliśmy, że badany zbiór danych jest niestacjonarnym szeregiem czasowym, który ma zarówno trend jak i sezonowość. Przystąpmy do dekompozycji szeregu. Wyodrębnimy tend i sezonowość, a następnie uzyskamy szereg bez części deterministycznych. Używamy do tego metody z biblioteki `statsmodels`.

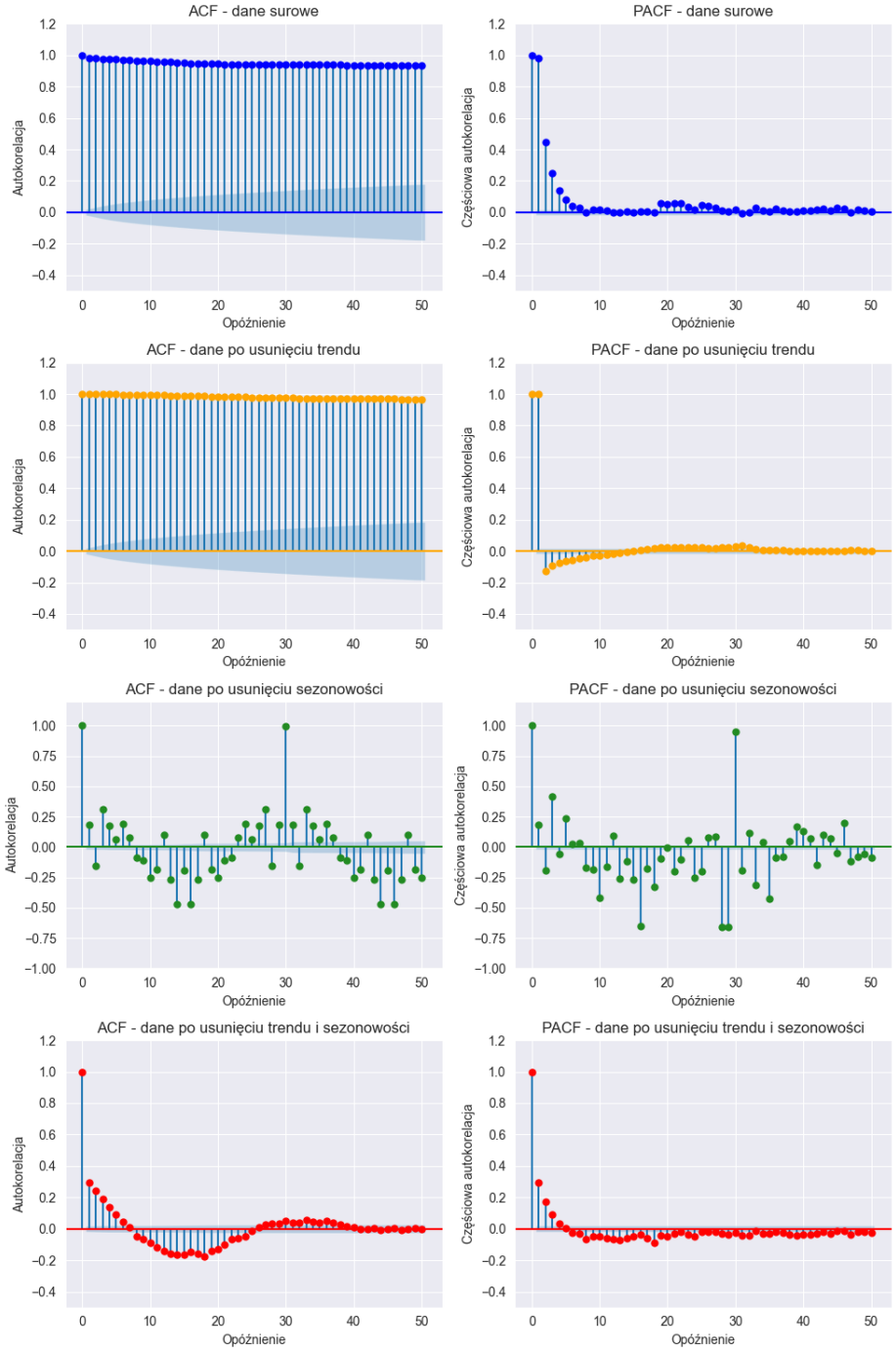


Rysunek 4. Dekompozycja danych.

Dane po usunięciu trendu i sezonowości przypominają biały szum, poza okresem z pierwszego kwartału 2020 roku, gdzie znajdują się wartości odstające.

Przeprowadzając analizę porównawczą wyników funkcji ACF i PACF po klasycznej dekompozycji omawianego zbioru danych, widzimy, że usunięcie trendu nie jest wystarczające, zbiór danych ma nadal silną autokorelację. Gdy pozbędziemy się samej sezonowości, ACF ma zmniejszone wartości, jednak widzimy pewien trend zależności, natomiast PACF wskazuje na mocniejszą częściową korelację niż w poprzednich przypadkach. Po usunięciu zarówno sezonowości jak i trendu wykresy funkcji ACF i PACF wyglądają "najlepiej",

to znaczy wskazują na model stacjonarny, ponieważ wartości funkcji autokowariancji szybko spadają do zera, a funkcja częściowej autokorelacji ma bardzo ograniczoną liczbę istotnych wartości.



Rysunek 5. Wykresy ACF i PACF po dekompozycji.

Sprawdźmy, jakie wyniki daje nam test ADF. Po oczyszczeniu danych statystyka testowa jest dużo mniejsza od wartości krytycznej oraz p-wartość jest znikoma na poziomie istotności 5%. Zatem odrzucamy hipotezę zerową o niestacjonarności szeregu. Na podstawie testu ADF szereg czasowy po dekompozycji jest stacjonarny.

Wartość krytyczna dla 5%	-2.8618138703454576
Statystyka testowa	-27.99
p-wartość	0.0

Tabela 2. Wyniki testu ADF po dekompozycji danych.

3. Modelowanie danych przy pomocy ARMA

3.1. Dobranie rzędu modelu

Dobranie najoptymalniejszego rzędu modelu polega na sprawdzeniu dla jakich p,$q \in N$ model ARMA(p,q) będzie najlepiej opisywał stacjonarne dane przy jednoczesnym unikaniu nadmiernej złożoności.

Kryteria informacyjne są funkcjami logarytmu maksymalnej wartości funkcji wiarygodności. Skorzystamy z trzech kryteriów informacyjnych, które służą do wybrania odpowiedniego modelu, na podstawie odpowiednich wzorów.

Kryterium Akaikego korzysta z równania

$$AIC = -2 \ln L + 2k,$$

gdzie L to maksymalna wartość funkcji wiarygodności oraz k to liczba parametrów modelu.

Kryterium Bayesowskie ma równanie

$$BIC = -2 \ln L + k \ln n,$$

gdzie n jest liczbą obserwacji w danych.

Kryterium Hannan-Quinn korzysta z równania

$$HQIC = -2 \ln L + 2k \ln(\ln n).$$

Kryterium	Najmniejsza wartość	p	q
AIC	98534.394489	1	1
BIC	98563.465906	1	1
HQIC	98544.206904	1	1

Tabela 3. Wyniki kryteriów informacyjnych.

Model ARMA(1,1) ma najniższą wartość zarówno AIC, BIC jak i HQIC spośród wszystkich testowanych modeli, co oznacza, że jednoznacznie najlepiej opisuje dane.

3.2. Estymacja parametrów modelu metodą największej wiarogodności

Metoda największej wiarogodności polega na znalezieniu takich wartości parametrów modelu, które maksymalizują funkcję wiarogodności danych. Dla badanego szeregu czasowego otrzymujemy następujące wyniki

Parametr	Wartość	Błąd	p-wartość
Stała	0.0089	0.453	0.984
ϕ_1	0.7454	0.004	0.000
θ_1	-0.4992	0.005	0.000
σ^2	558.95	2.538	0.000

Tabela 4. Oszacowane parametry modelu.

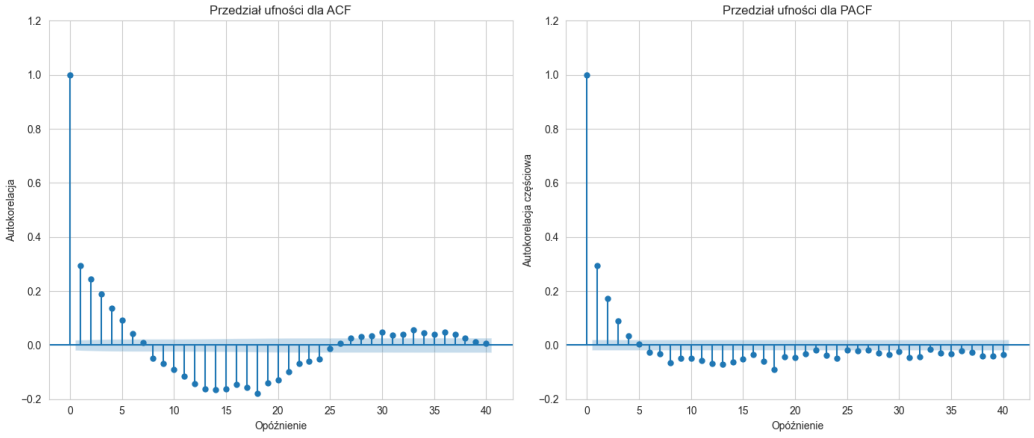
Wartości ϕ_1 (AR) i θ_1 (MA) są istotne statystycznie (p-wartość < 0.05), co oznacza, że rzeczywiście wpływają na model. Stała w modelu jest nieistotna statystycznie ze względu na jej dużą p-wartość. Wariancja składnika błędu jest dość duża, co wynika z bardzo dużego wahania wartości odstających (z okresu Covid-19). Ostatecznie możemy przybliżać oczyszczone dane modelem ARMA(1,1)

$$X_t - 0.7454 \cdot X_{t-1} = Z_t - 0.4992 \cdot Z_{t-1}.$$

4. Ocena dopasowania modelu

4.1. Przedziały ufności

Celem wyznaczenia przedziałów ufności dla ACF i PACF jest ocena, które wartości autokorelacji są istotne statystycznie, a które mogą wynikać z przypadku. Pomagają one w identyfikacji istotnych zależności w danych szeregów czasowych, co jest niezbędne do prawidłowego modelowania. Przedziały ufności wspierają wybór odpowiednich parametrów w modelach ARMA lub ARIMA, umożliwiając skuteczną analizę i prognozowanie.



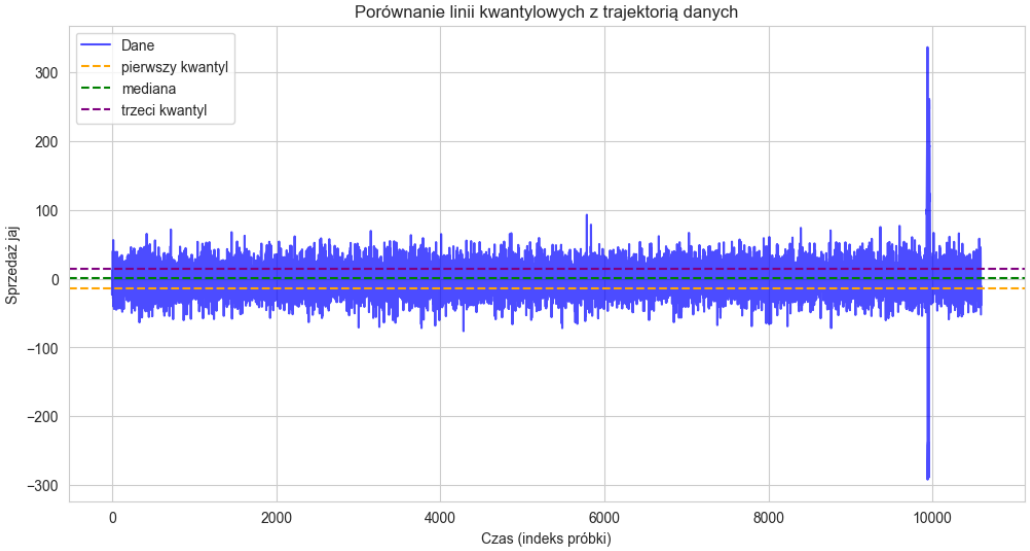
Rysunek 6. Wykresy przedziałow ufności ACF i PACF

Przedziały ufności dla ACF pokazują, że większość wartości mieści się w granicach losowej zmienności, z wyjątkiem kilku początkowych lagów, co sugeruje krótkoterminowe zależności. W PACF pierwsze dwa–trzy opóźnienia przekraczają przedział ufności, co wskazuje na możliwy proces autoregresyjny niskiego rzędu. Brak istotnych przekroczeń przedziałów ufności w dalszych

lagach sugeruje brak długoterminowych zależności w danych. Szerokość przedziałów ufności jest stabilna, co świadczy o dobrze oczyszczonych danych po dekompozycji.

4.2. Porównanie linii kwantylowych z trajektorią

Wykres trajektorii danych z liniami kwantylowymi pozwala analizować zarówno ogólny trend, jak i zmienność badanej wielkości w czasie. Kwantyle pokazują, w jakim zakresie wartości zwykle się znajdują, co pomaga ocenić typowe poziomy sprzedaży i ich rozkład. Dzięki temu można wykryć anomalie, takie jak nagłe spadki lub wzrosty, które mogą wynikać z sezonowości, błędów pomiarowych lub innych czynników.



Rysunek 7. Wykres porównania lini kwantylowych z trajektorią

Większość danych oscyluje wokół mediany, co wskazuje na względnie stabilny rozkład resztowy.Pierwszy i trzeci kwantyl (dolna i górna linia przerywana) obejmują główny zakres zmienności, co sugeruje, że większość wartości mieści się w typowym przedziale międzykwartylowym. Na końcu szeregu występują pojedyncze, ekstremalne wartości (szczyty i spadki), które mogą wskazywać na anomalię lub nietypowe zdarzenia w sprzedaży.

5. Weryfikacja założeń dotyczących szumu

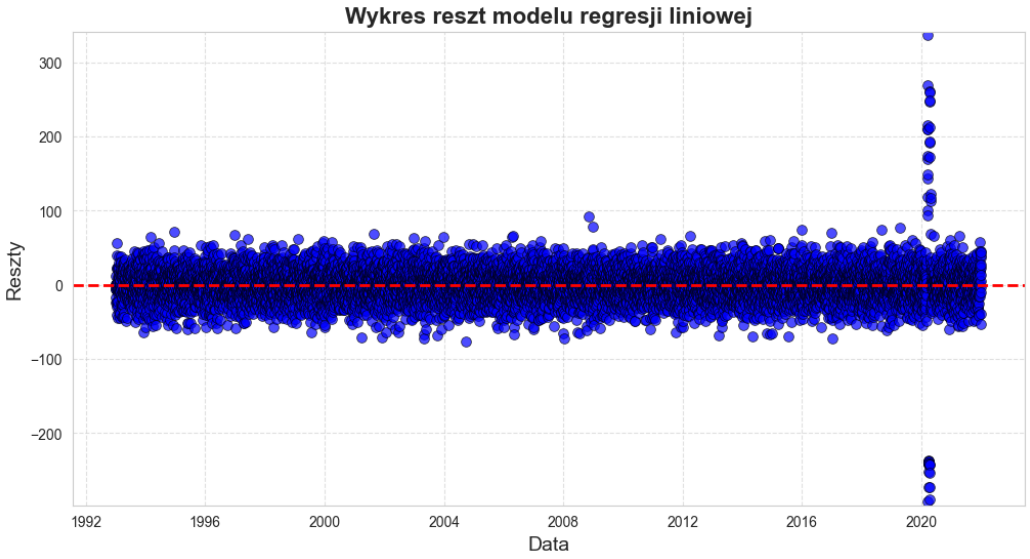
Będziemy weryfikować założenia dotyczące szumu w modelu, analizując jego reszty pod kątem różnych właściwości statystycznych. Sprawdzimy, czy reszty mają zerową średnią za pomocą wykresu wartości resztowych i testu t, a także czy ich wariancja jest stała, wykorzystując wykres reszt, Modified Levene Test i Arch Test. Niezależność reszt ocenimy na podstawie wykresów ACF/PACF oraz testu Ljunga-Boxa, co pozwoli wykryć ewentualne autokorelacje. Na koniec zweryfikujemy, czy reszty mają rozkład normalny,

analizując dystrybuantę, gęstość, wykres kwantylowy oraz stosując testy normalności.

5.1. Założenie o średniej

Wykres wartości resztowych – sprawdza, czy reszty modelu oscylują wokół zera, co sugeruje brak systematycznych błędów.

t-test – testuje hipotezę, czy średnia reszt jest równa zeru. Jeśli p-wartość jest większa od ustalonego poziomu istotności, nie ma podstaw do odrzucenia hipotezy zerowej.



Rysunek 8. Wykres reszt modelu regresji liniowej

Reszty modelu regresji liniowej są skoncentrowane wokół wartości 0, co sugeruje brak silnych systematycznych błędów w modelu. W większości okresu reszty są równomiernie rozproszone, jednak pod koniec (ok. 2020) widoczne są duże wartości odstające. Brak wyraźnego wzorca w resztach sugeruje, że model dobrze dopasował się do danych historycznych.

Statystyka	**Wartość**
T-statystyka	0.0000
p-wartość	1.0000

Tabela 5. Wyniki testu t

Wartości t-statystyki oraz p-wartości oznaczają, że nie ma istotnych statystycznie różnic między średnią reszt a zerem. Wynik ten potwierdza, że model regresji liniowej nie wykazuje systematycznego błędu – średnia reszt jest zgodna z założeniem regresji liniowej. Oznacza to, że model nie ma skłonności do systematycznego przeszacowywania lub niedoszacowywania wyników.

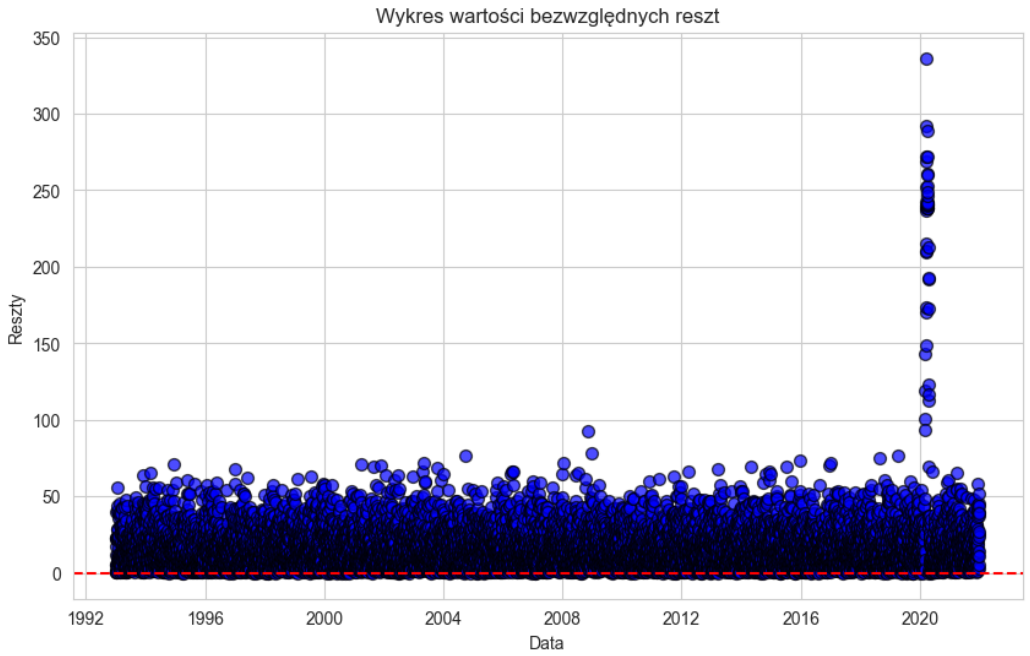
5.2. Założenie o wariancji

Wykres wartości bezwzględnych reszt – pomaga wizualnie ocenić, czy wariancja reszt jest stała w czasie (homoskedastyczność).

Modified Levene Test – statystyczny test sprawdzający, czy wariancja reszt jest stała (test homoskedastyczności).

Arch Test – testuje obecność efektu ARCH, czyli zmiennej wariancji w czasie.

Statystyka F - porównuje wariancje 2 zbiorów danych do testowania stałości wariancji reszt.



Rysunek 9. Wykres wartości bezwzględnych reszt

W latach poprzedzających 2020 rok wartości reszt były relatywnie stabilne i skupione wokół niskich wartości, co sugeruje homoskedastyczność w tym okresie. Rozrzut wartości reszt przed 2020 rokiem był niewielki, a większość reszt oscylowała wokół zera, co potwierdza stabilność modelu w tym przedziale czasowym. Jednak począwszy od roku 2020, zaobserwowano znaczący wzrost rozrzutu wartości reszt, w tym pojawienie się licznych wartości skrajnych. Wartości reszt po 2020 roku są znacznie bardziej rozproszone, a niektóre z nich znacznie odbiegają od średniej. Taka zmiana w rozrzucie reszt wskazuje na możliwość wystąpienia heteroskedastyczności, czyli zmiennej wariancji reszt w czasie.

Wyniki wszystkich trzech testów (ARCH, F-statistic oraz Modified Levene Test) są spójne i jednoznacznie wskazują na obecność heteroskedastyczności w analizowanym szeregu czasowym. Oznacza to, że wariancja reszt nie jest stała w czasie, lecz zmienia się istotnie, co może prowadzić do błędów w modelowaniu i wnioskowaniu statystycznym.

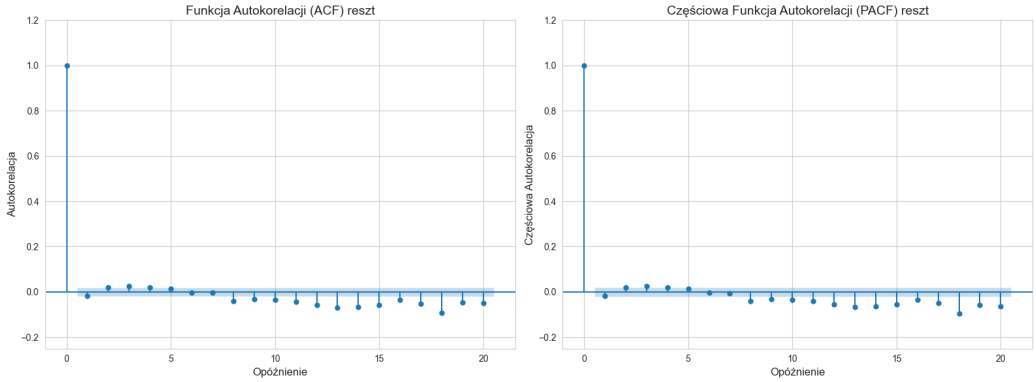
Test	Wartość
Statystyka Modified Levene Test	21.4483
p-wartość Modified Levene Test	3.6783e-06
Statystyka ARCH	9512.06
p-wartość ARCH	0.0000
Statystyka F	9397.96
p-wartość F	0.0000

Tabela 6. Wyniki testów: Modified Levene Test, ARCH oraz F.

5.3. Założenie o niezależności

Wykres ACF/PACF dla reszt – sprawdza, czy reszty są skorelowane w czasie. Brak wyraźnych pików sugeruje, że reszty są niezależne.

Test Ljunga-Boxa – statystyczny test sprawdzający, czy reszty są losowe (brak autokorelacji). Wysoka p-wartość sugeruje brak zależności czasowych.



Rysunek 10. Wykres ACF/PACF dla wartości resztowych

Wartość autokorelacji dla opóźnienia 0 wynosi 1, co jest oczekiwanym wynikiem. Dla kolejnych opóźnień wartości autokorelacji szybko maleją, jednak kilka z nich przekracza granicę istotności. Wskazuje to na obecność pewnej struktury w danych reszt, co sugeruje ich zależność w czasie. Podobnie jak w przypadku ACF, wartości częściowej autokorelacji szybko spadają, ale dla niektórych opóźnień przekraczają granicę istotności. To zjawisko potwierdza, że reszty mogą być częściowo skorelowane w czasie. Hipoteza zerowa testu

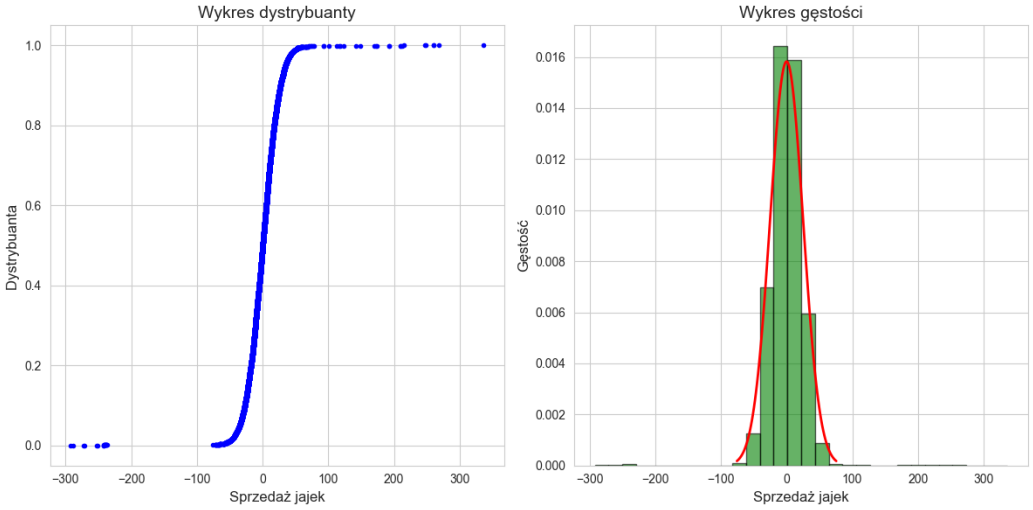
Statystyka testu Ljung-Boxa	p-wartość
lb_stat = 60.69	2.68×10^{-9}

Tabela 7. Wynik testu Ljung-Boxa

Ljunga-Boxa zakłada brak autokorelacji reszt. Odrzucenie tej hipotezy oznacza, że reszty nie są niezależne, czyli istnieje istotna autokorelacja w danych.

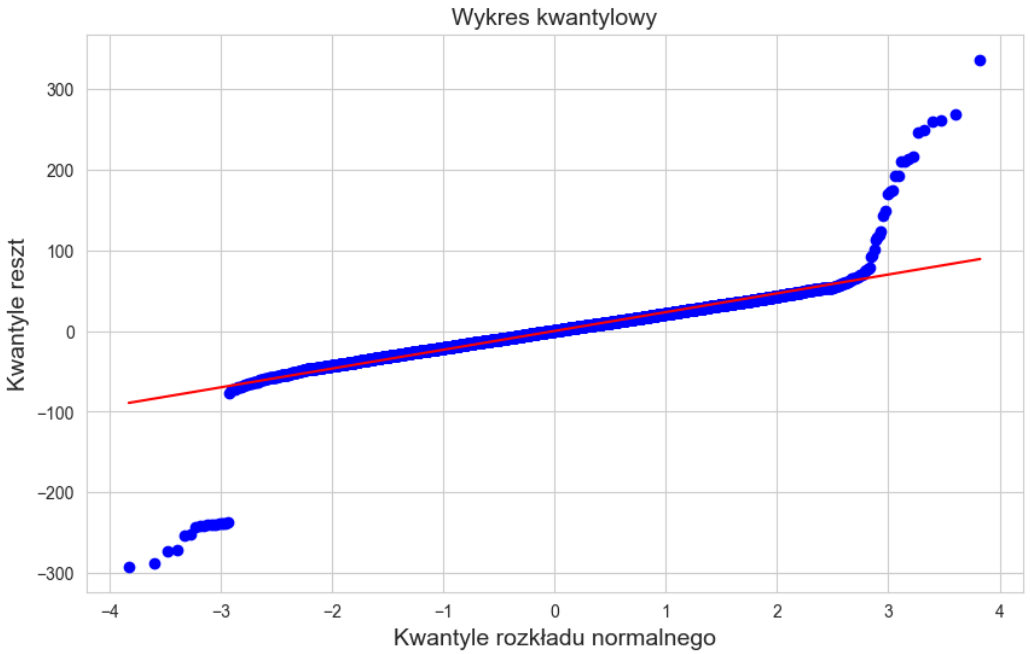
5.4. Założenie o normalności rozkładu

Dystrybuanta – porównuje empiryczny rozkład reszt z teoretycznym rozkładem normalnym.
Wykres gęstości – wizualna ocena kształtu rozkładu reszt.
Wykres kwantylowy (Q-Q plot) – porównuje kwantyle empiryczne z teoretycznymi kwantylami normalnego rozkładu. Jeśli punkty układają się wzdłuż prostej, rozkład jest normalny.
Testy na normalność – statystycznie weryfikują, czy reszty mają rozkład normalny. Niska p-wartość sugeruje odchylenie od normalności.]



Rysunek 11. Wykresy gęstości i dystrybuanty

Dystrybuanta ma charakterystyczny kształt dla rozkładu normalnego, wykazując szybki wzrost od wartości bliskich 0 dla niskich wartości zmiennej do poziomu zbliżonego do 1 dla wysokich wartości. Jednak na obu krańcach widoczne są odstające wartości, co może wskazywać na niewielkie odstępstwa od założenia normalności. Histogram gęstości przypomina rozkład normalny, zbliżony do symetrii, co potwierdza kształt krzywej gęstości, ale obecność wyraźnych obserwacji odstających sugeruje odchylenia od idealnego rozkładu. Rozkład wydaje się bardziej skoncentrowany wokół średniej, co może oznaczać nieco mniejszy rozrzut danych.



Rysunek 12. Wykres kwantylowy

Wykres kwantylowy wskazuje, że rozkład reszt odbiega od normalności, szczególnie na krańcach, gdzie punkty znacznie odchylają się od linii prostej, sugerując obecność wartości odstających. Środkowa część wykresu dobrze pasuje do linii prostej, jednak skrajne wartości reszt wykazują nieliniowość, co potwierdza problem z dopasowaniem do rozkładu normalnego. Charakterystyczne zakrzywienia na końcach wskazują na rozkład z cięższymi ogonami, co oznacza większą liczbę wartości ekstremalnych niż w przypadku rozkładu normalnego.

Test diagnostyczny	Wyniki
Jarque-Bera	(JB) = 246445.95, Prob(JB) = 0.00
Skośność	-0.19
Kurtoza	26.63
Heteroskedastyczność	(H) = 1.65, Prob(H) = 0.00

Tabela 8. Testy diagnostyczne dla modelu.

Test Jarque-Bera z tak wysokim wynikiem i p-wartością 0 oznacza istotne odchylenie rozkładu od normalności. Skośność dla rozkładu normalnego powinna wynosić 0. Wynik sugeruje lekką asymetrię w lewo, ale nie jest to duże odchylenie. Kurtoza dla rozkładu normalnego powinna wynosić około 3. Bardzo wysoka wartość 26.63 potwierdza wcześniejsze przypuszczenia o rozkładzie z bardzo ciężkimi ogonami. Dla bardzo niskiej p-wartości dla heteroskedastyczności odrzucamy hipotezę zerową o stałej wariancji.

6. Zakończenie

Powyższa praca stanowi kompleksową analizę danych o sprzedaży jajek w sklepie na Sri Lance, opartą na modelu ARMA. Główne etapy pracy obejmują:

— Przygotowanie danych i dekompozycję szeregu czasowego. Wykorzystano 30-letnią historię sprzedaży jajek. Dane wykazały wyraźne sezonowe wahania oraz trend, a także anomalie (m.in. zerowa sprzedaż w czasie pandemii COVID-19). Przeprowadzono klasyczną dekompozycję szeregu, mającą na celu oddzielenie części deterministycznych (trend i sezonowość) od stacjonarnego szumu.

— Modelowanie przy użyciu modelu ARMA. Na podstawie analizy funkcji autokorelacji (ACF) i częściowej autokorelacji (PACF) oraz kryteriów informacyjnych (AIC, BIC, HQIC) dobrano optymalny model ARMA(1,1). Estymacja parametrów metodą największej wiarygodności wykazała, że współczynniki modelu są statystycznie istotne, co potwierdza trafność modelu dla oczyszczonych, stacjonarnych danych.

— Ocena dopasowania modelu i weryfikacja założeń. Przeprowadzono analizę reszt modelu pod kątem średniej, wariancji, niezależności i normalności rozkładu. Wyniki testów wykazały, że choć średnia reszt jest bliska zeru, to występują problemy z heteroskedastycznością (zmienną wariancją w czasie), autokorelacją reszt oraz odchyleniami od rozkładu normalnego (znacząco wysoka kurtoza i istotny wynik testu Jarque-Bera).

Główne wnioski z pracy:

— **Skuteczność modelu ARMA(1,1)**. Model dobrze odwzorowuje zależności czasowe w danych po usunięciu trendu i sezonowości, co umożliwia efektywne prognozowanie przy założeniu stacjonarności szeregu.

— **Wpływ zdarzeń zewnętrznych**. Analiza potwierdziła, że globalne wydarzenia, takie jak pandemia COVID-19, mają znaczący wpływ na sprzedaż, co ujawnia się m.in. poprzez wartości odstające w analizowanym szeregu.

— **Problemy z założeniami modelu**. Mimo trafności modelu w opisie struktury czasowej, wyniki testów diagnostycznych ujawniły problemy z heteroskedastycznością oraz normalnością rozkładu reszt, co sugeruje konieczność zastosowania dalszych metod (np. transformacji danych) w celu poprawy jakości prognoz.

Podsumowując, praca pokazuje, że model ARMA(1,1) może być skutecznym narzędziem do modelowania stacjonarnych szeregów czasowych w kontekście sprzedaży, ale także wskazuje na ograniczenia wynikające z nieregularności i zmiennej wariancji danych, które warto uwzględnić przy dalszych badaniach i prognozowaniu.