Spis treści

1. Wprowadzenie. 3
1.1 Cele Raportu. 3
1.2 Opis Danych. 3
1.3 Dane. 3
1.4 Wzory i definicje. 4
2. Przygotowanie danych do analizy. 6
2.1 Dekompozycja szeregu czasowego. 6
3. Modelowanie danych przy pomocy ARMA(p,q). 10
3.1 Dobranie rzędu modelu. 10
3.2 Estymacja parametrów modelu. 11
4. Ocena dopasowania modelu. 12
5. Weryfikacja założeń dotyczących szumu. 13
5.1 Założenie o stałej zerowej wartości średniej. 13
5.2 Założenie o stałej skończonej wariancji. 14
5.3 Założenie o niezależności. 14
5.4 Założenie o rozkładzie normalnym $N(0, \sigma 2)$. 15
6. Wnioski końcowe. 16

1. Wprowadzenie.

1.1 Cele Raportu.

- Utrwalenie wiedzy z zakresu opisu danych, wykonywaniu wykresów i analizy podstawowych zagadnień.
- Zagadanie jakości danych i przeprowadzenie dekompozycji szeregu czasowego.
- Zaproponowanie i przeanalizowanie modelu ARMA(p,q), dopasowanego do danych rzeczywistych. Ocena dopasowania modelu.
- Weryfikacja założeń dotyczących białego szumu (White Noise).
- Wyciągnięcie wniosków z otrzymanych wyników.

1.2 Opis Danych.

Przeprowadzono analizę danych dotyczących cen akcji po otwarciu giełdy dla dystrybutora filmów Netflix od dnia 04.01.2016 do 12.05.2023. Badano próbkę o długości 1995 (giełda jest zamknięta w święta, soboty i niedziele), z dwoma zmiennymi:

- Cena akcji po otwarciu giełdy (Open),
- Data (Date) – dzień w którym akcja miała podaną cenę.

Dane są częścią większego ich zbioru.

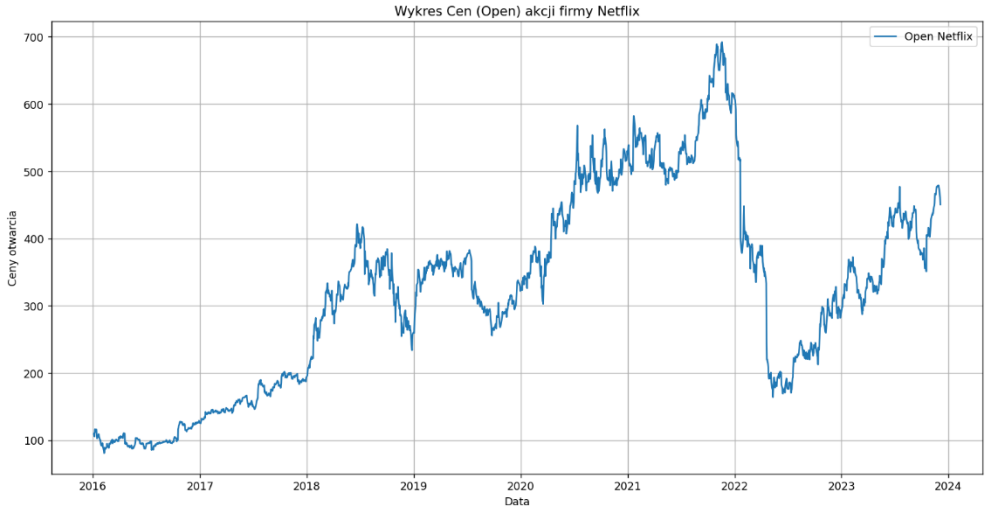
(Dane pochodzą ze strony: https://www.kaggle.com)

1.3 Dane.

W poniższej tabeli (Rys.1), zaprezentowano przykładowe dane z próbki. Kolejno na wykresie przedstawiono dane (Rys.2).

Data	Open Netflix
04.01.2016	109.00
05.01.2016	110.45
06.01.2016	105.29
07.01.2016	116.36
08.01.2016	116.33

Rysunek 1 – Przykładowe dane.



Rysunek 1 – Wykres danych.

1.4 Wzory i definicje.

1.4.1 Transformacje danych.

Surowe dane, które zostały wybrane, wymagały odpowiednich przekształceń. W tym celu zastosowano:

- transformację Boxa-Coxa, która dla danego parametru λ przedstawia się wzorem:

$$BC_\lambda(X) = \begin{cases} \dfrac{X^\lambda - 1}{\lambda}, & \lambda \neq 0, \\ \ln X, & \lambda = 0; \end{cases}$$

- różnicowanie danych rzędu p – dzięki temu uzyskuje się nowy szereg $\{X_t^*\}_{t\in Z}$ zadany wzorem:

$$X_t^* = X_t - X_{t-p};$$

- centrowanie danych, czyli odjęcie od każdej obserwacji średniej próbkowej z próby.

1.4.2 Szereg stacjonarny w słabym sensie.

Szeregiem stacjonarnym w słabym sensie, nazywamy szereg, który spełnia dwa następujące warunki:

- $\forall t \in Z\ EX_t = const.$ — funkcja średniej jest stała w czasie,
- $Cov(X_t, X_{t+h}) = \gamma X(h)$ — funkcja autokowariancji zależy jedynie od przesunięcia h.

1.4.3 Funkcje autokorelacji.

Użyto wykresów funkcji autokorelacji (ACF) oraz funkcji częściowej autokorelacji (PACF), aby zbadać czy autokorelacja rozpatrywanego szeregu czasowego wskazuje na jego stacjonarność w słabym sensie. Estymator funkcji autokowariancji dany jest wzorem:

$$\hat{\gamma}(h) = \frac{1}{n}\sum_{t=1}^{n-|h|}\left(x_{t+|h|} - \bar{x}\right)(x_t - \bar{x}),$$

gdzie $\bar{x}$ to średnia próbkowa. Empiryczną funkcję autokorelacji definiuje się jako:

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

1.4.4 Model ARMA.

Model ARMA(p,q) nazywamy szereg czasowy stacjonarny w słabym sensie $\{X_t\}$, który spełnia równanie:

$$X_t - \varphi_1 X_{t-1} - \cdots - \varphi_p X_{t-p} = Z_t + \theta_1 Z_{t-1} + \cdots + \theta_q Z_{t-q},$$

gdzie $\{Z_t\}$ to $WN(0, \sigma^2)$ – biały szum, a wielomiany:

$$\varphi(z) = 1 - \varphi_1 z - \cdots - \varphi_p z^p,$$
$$\theta(z) = 1 - \theta_1 z - \cdots - \theta_p z^p,$$

nie mają wspólnych pierwiastków.

1.4.5 Analiza residuów.

Residuum modelu ARMA to różnica wartości rzeczywistej oraz tej zamodelowanej dla konkretnej chwili czasowej t:

$$\varepsilon_t = x_t - \hat{x}_t,$$

$$\hat{x}_t = \sum_{i=1}^{p}\widehat{\varphi_i} x_{t-i} + \sum_{j=1}^{q}\widehat{\theta_j} z_{t-j}.$$

W poprawnie dobranym modelu ARMA residua powinny zachowywać się jak ciąg nieskorelowanych zmiennych losowych z rozkładu normalnego o średniej zero i stałej wariancji σ^2.

1.4.6 Kryterium informacyjne Akaikego.

W celu dobrania optymalnych parametrów modelu ARMA wykorzystano kryterium informacyjne Akaikego (AIC). Jego wartość jest zadana wzorem:

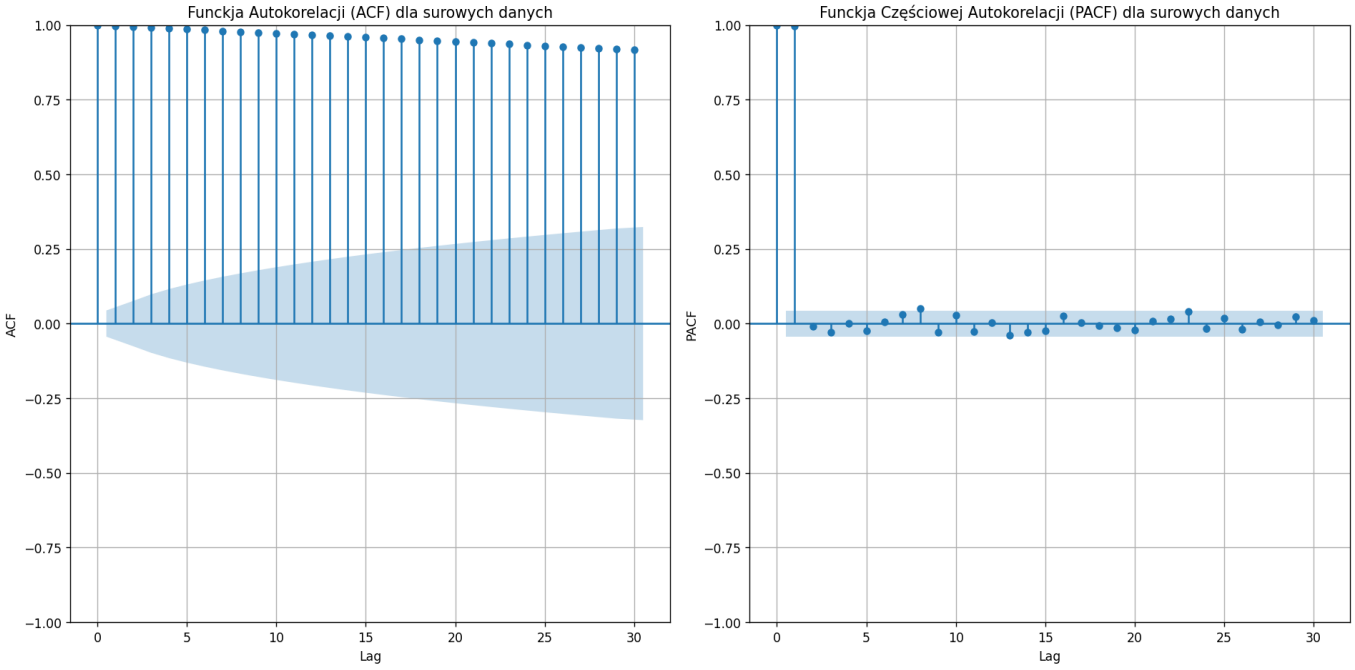
$$AIC = -2\ln(L) + 2k,$$

gdzie L to maksymalna wartość funkcji wiarygodności modelu, a k to liczba parametrów modelu. Spośród rozpatrywanych modeli wybrano ten, dla którego AIC jest najmniejsze.

2. Przygotowanie danych do analizy.

2.1 Dekompozycja szeregu czasowego.

Dla surowych danych z próbki wykonano wykresy (Rys.3) autokorelacji (ACF) i częściowej autokorelacji (PACF).



Rysunek 3 – Wykresy ACF i PACF dla surowych danych.

Z wykresu funkcji autokorelacji (ACF), widać, że występują wysokie wartości autokorelacji dla wielu opóźnień, co sugeruje niestacjonarność szeregu czasowego. Dodatkowo z wykresu funkcji częściowej autokorelacji (PACF), widzimy, że nie wszystkie wartości „chowają" się w przedział ufności, co świadczy, o zależności danych. Przeprowadzono test ADF (Augmented Dickey-Fuller Test), weryfikujący hipotezę o niestacjonarności surowych danych. Wyniki przeprowadzonego testu zamieszczono w poniższej tabeli (Rys.4).

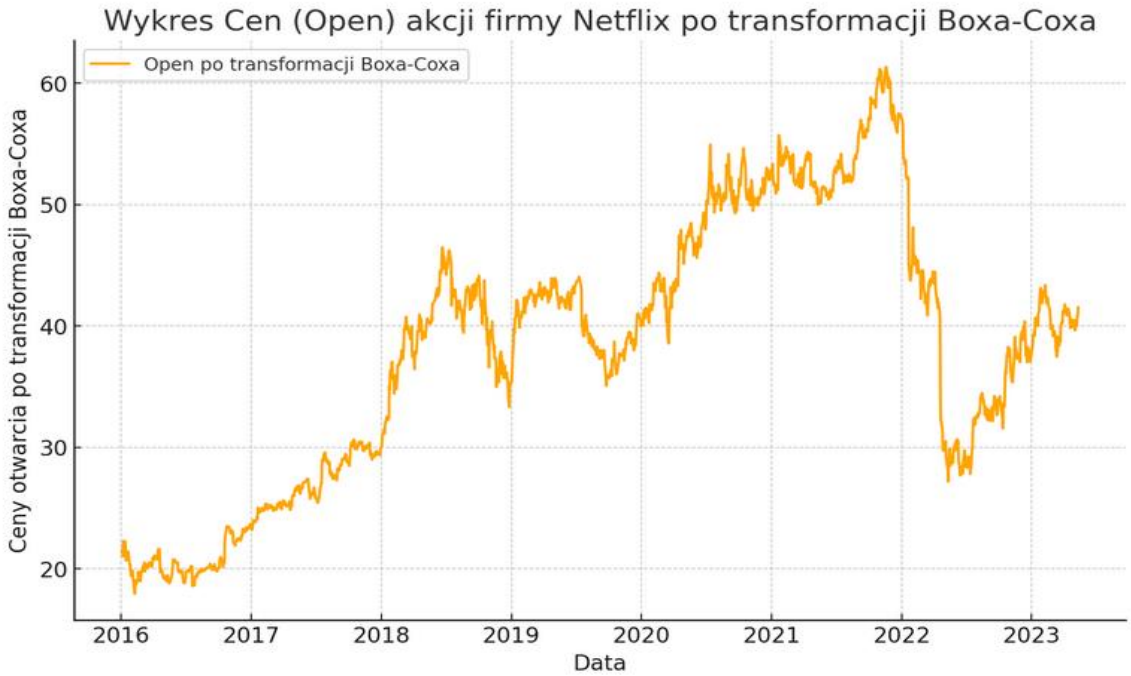
Statystyka testowa:	-1.668
Wartość p (p-value):	0.448
Liczba użytych obserwacji (Number of Observations Used):	1852

Wartość krytyczna dla 1% (Critical Value 1%):	−3.434
Wartość krytyczna dla 5% (Critical Value 5%):	−2.863
Wartość krytyczna dla 10% (Critical Value 10%):	−2.568

Rysunek 4 – Wyniki testu ADF dla danych surowych.

Z powyższej tabeli, widzimy, że statystyka testowa jest wyższa niż wszystkie wartości krytyczne, a p-value jest znacznie większa niż typowy próg 0.05, nie możemy zatem odrzucić hipotezy zerowej mówiącej o niestacjonarności szeregu czasowego. Co sugeruje, że surowe dane są niestacjonarne. Co potwierdza wcześniejsze wnioski, wyciągnięte z wykresów ACF i PACF (Rys.3).

Dążymy do otrzymania szeregu czasowego stacjonarnego w słabym sensie. Przeprowadzono w tym celu, transformacje stabilizująca wariancję, skorzystano z metody Boxa-Coxa. Dane po transformacji zaprezentowano na poniższym wykresie (Rys.5).



Rysunek 5 – Wykres danych po zastosowaniu metody Boxa-Coxa.

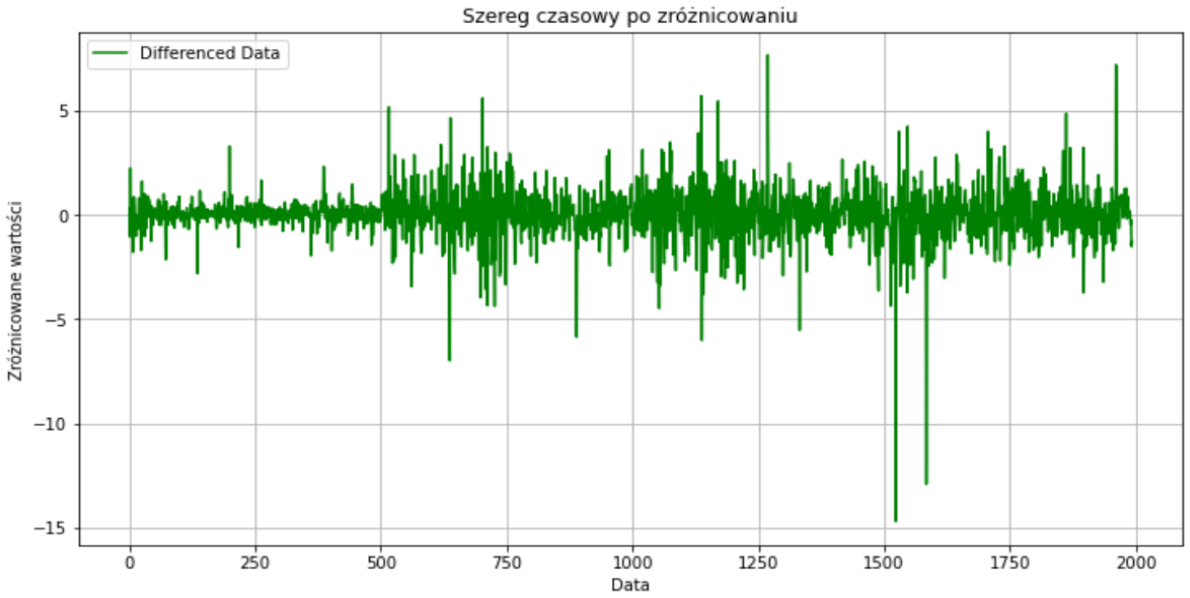
Z wykresu widać, że transformacja zmieniła rozkład wartości, ale nie wiadomo, czy poprawiła stacjonarność szeregu czasowego. Dlatego, ponownie przeprowadzono test ADF (Augmented Dickey-Fuller Test), weryfikujący hipotezę o niestacjonarności danych, po zastosowaniu transformacji Boxa-Coxa (na surowych danych cen otwarcia firmy Netflix).Wyniki przeprowadzonego testu zamieszczono w poniższej tabeli (Rys.6).

Statystyka testowa:	-1.662
Wartość p (p-value):	0.451
Liczba użytych obserwacji (Number of Observations Used):	1852
Wartość krytyczna dla 1% (Critical Value 1%):	-3.434
Wartość krytyczna dla 5% (Critical Value 5%):	-2.863
Wartość krytyczna dla 10% (Critical Value 10%):	-2.568

Rysunek 6 – Wyniki testu ADF dla danych po transformacji Boxa-Coxa.

Podobnie jak w poprzednim przypadku (dla danych surowych Rys.4), statystyka testowa jest wyższa niż wszystkie wartości krytyczne, a p-value jest nadal znacznie wyższa niż typowy próg 0.05. Świadczy to, że nawet po transformacji Boxa-Coxa szereg czasowy pozostaje niestacjonarny.

Dalej przeprowadzono zróżnicowanie danych już, po transformacji Boxa-Coxa, rzędem p=1. Dane po zróżnicowaniu ukazano, na poniższym wykresie (Rys.7).



Rysunek 7 – Wykres danych po zróżnicowaniu.

Tak jak wyżej znów przeprowadzono test ADF (Augmented Dickey-Fuller Test), weryfikujący hipotezę o niestacjonarności danych, po zastosowaniu transformacji Boxa-Coxa i zróżnicowaniu, na surowych danych cen otwarcia firmy Netflix. Wyniki przeprowadzonego testu zamieszczono w poniższej tabeli (Rys.8).

Statystyka testowa:	-42.598
Wartość p (p-value):	0.000
Liczba użytych obserwacji (Number of Observations Used):	1851
Wartość krytyczna dla 1% (Critical Value 1%):	-3.434
Wartość krytyczna dla 5% (Critical Value 5%):	-2.863
Wartość krytyczna dla 10% (Critical Value 10%):	-2.568

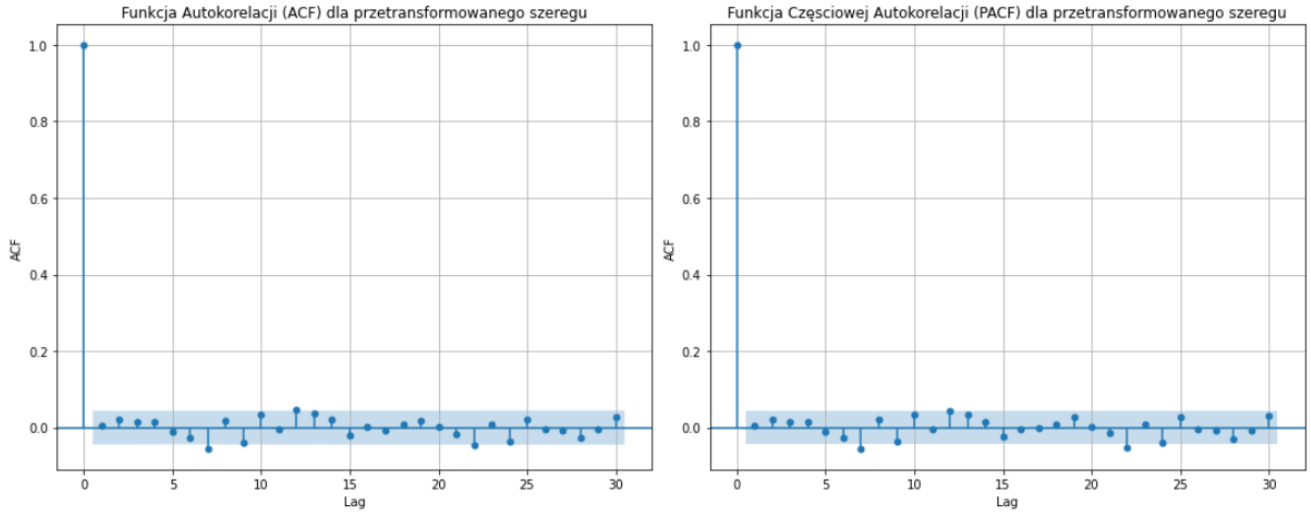
Rysunek 8 – Wyniki testu ADF dla danych po transformacji Boxa-Coxa i zróżnicowaniu.

Z powyższej tabeli, widzimy, że statystyka testowa jest znacznie niższa niż wszystkie wartości krytyczne dla ustalonych poziomów ufności, a p-value jest równa zero. Świadczy to o tym, że możemy odrzucić hipotezę zerową, na rzecz hipotezy alternatywnej, zróżnicowany szereg czasowy po transformacji Boxa-Coxa jest stacjonarny.

Dodatkowo dla uproszczenia dalszej analizy, odjęto od zróżnicowanych danych po transformacji Boxa-Coxa, średnią próbkową. Nie wpływa to na stacjonarność szeregu czasowego, ponieważ działanie to nie zmienia fundamentalnych właściwości stacjonarności szeregu czasowego. Poniższy wykres przedstawia szereg czasowy po wycentrowaniu (Rys.9).

Rysunek 9 – Wykres wycentrowanych, zróżnicowanych danych po transformacji Boxa-Coxa.

Dla wycentrowanych, zróżnicowanych danych po transformacji Boxa-Coxa, wykonano wykresy (Rys.10) autokorelacji (ACF) i częściowej autokorelacji (PACF).



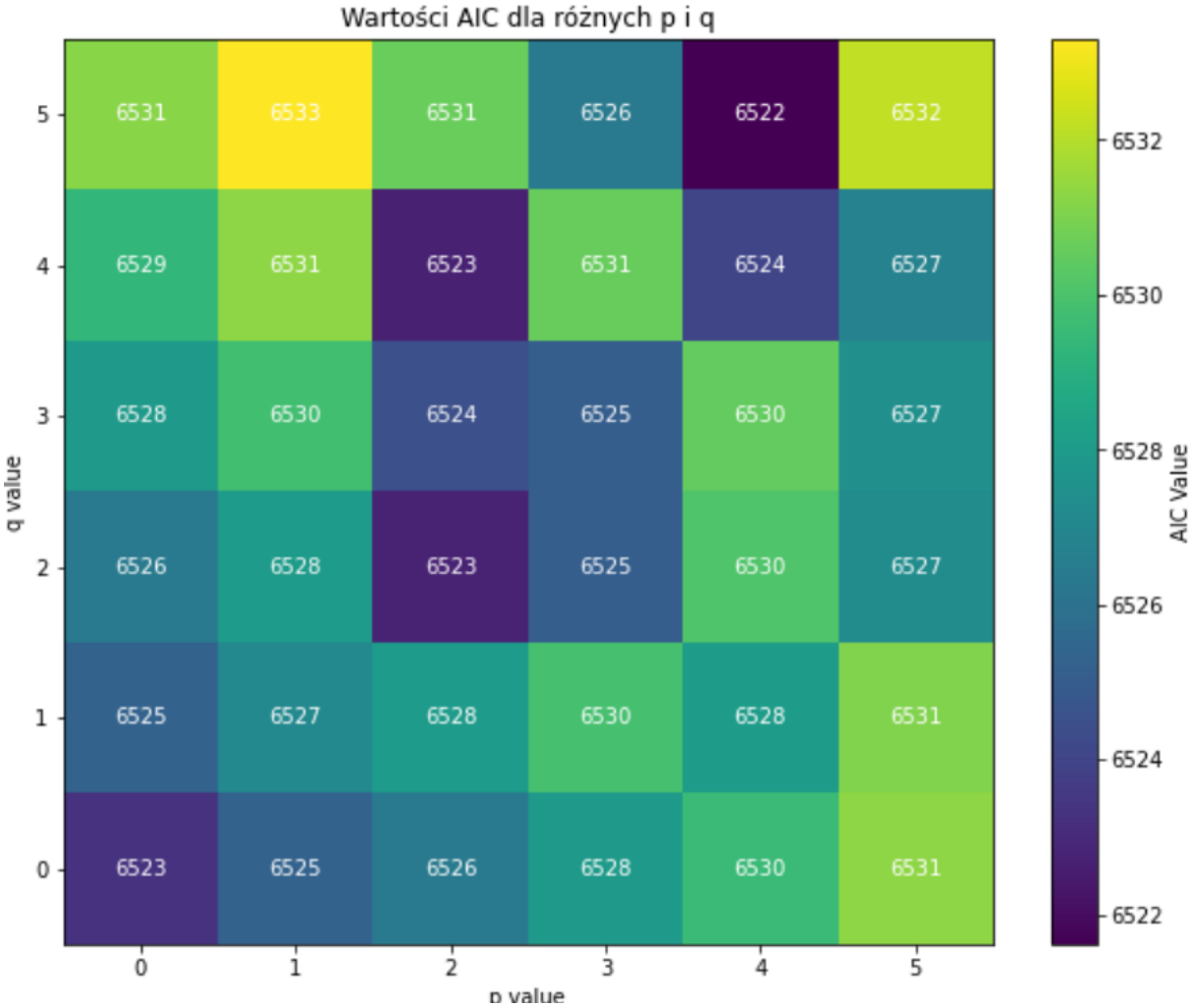
Rysunek 10 – Wykresy ACF i PACF dla danych po wszystkich transformacjach.

Z wykresu funkcji autokorelacji (ACF), widać, silną korelację w lagu zero, ponieważ każdy szereg jest idealnie skorelowany sam ze sobą. Bardzo szybki spadek korelacji po pierwszym lagu wskazuje na to, że szereg czasowy nie wykazuje znaczącej autokorelacji dla wyższych lagów, co jest typowe dla stacjonarnego szeregu czasowego. Dodatkowo z wykresu funkcji częściowej autokorelacji (PACF), widzimy, że wszystkie wartości „chowają" się w przedział ufności, co świadczyć może, o niezależności danych. Wnioskować, zatem można, że szereg czasowy jest prawdopodobnie stacjonarny po zastosowanych transformacjach. Szybki spadek korelacji po pierwszym lagu na obu wykresach może sugerować, że dla modelowania tego szeregu czasowego, odpowiedni byłby model AR(1) lub MA(1). W ARIMA(p, d, q), p odpowiadałby wartości 1 ze względu na obserwację na wykresie częściowej autokorelacji(PACF), podczas gdy q również może być równe 1 na podstawie wykresu korelacji (ACF).

3. Modelowanie danych przy pomocy ARMA(p,q).

3.1 Dobranie rzędu modelu.

Do dobrania rzędu modelu, wykorzystano kryterium informacyjne Akaikego (AIC). Sprawdzono kilka kombinacji p i q, z których wybrano, tą parę, która miała najmniejsza wartość współczynnika AIC. Na poniższej mapie cieplnej, zaprezentowano kilka sprawdzonych kombinacji (Rys.11). Dla p=4 i q=5 otrzymano najniższą wartość współczynnika AIC.



Rysunek 11 – Wykres ciepła współczynnika AIC dla różnych wartości p i q.

Do zweryfikowania rzędu modelu, wykorzystano również Bayesowskie kryterium informacyjne Schwarza (BIC) oraz kryterium informacyjne Hannana-Quinna (HQIC), które potwierdziły poprawność dobrania p i q, jako odpowiednio 4 i 5.

3.2 Estymacja parametrów modelu.

Do estymacji parametrów modelu ARMA(4,5) skorzystano z metody największej wiarygodności. Model ARMA(4,5) prezentuje się następująco:

$$X_t = \phi_1 X_{t-1} + \phi_2 X_{t-2} + \phi_3 X_{t-3} + \phi_4 X_{t-4} + Z_t + \theta_1 Z_{t-1} + \theta_2 Z_{t-2} + \theta_3 Z_{t-3} + \theta_4 Z_{t-4} + \theta_5 Z_{t-5};$$

Gdzie:

- $\phi_1, \phi_2, \phi_3, \phi_4$ – to parametry autoregresyjne,
- $\theta_1, \theta_2, \theta_3, \theta_4, \theta_5$ – to parametry średniej ruchomej,
- Z_t–biały szum (white noise) z rozkładem $N(0, \sigma^2)$.

Estymacji parametrów modelu ARMA(4,5) przeprowadzana analitycznie, jest bardzo trudna, więc skorzystano z metody numerycznej do wyznaczenia estymatorów. Wyniki zaprezentowano w poniższej tabeli (Rys.12).

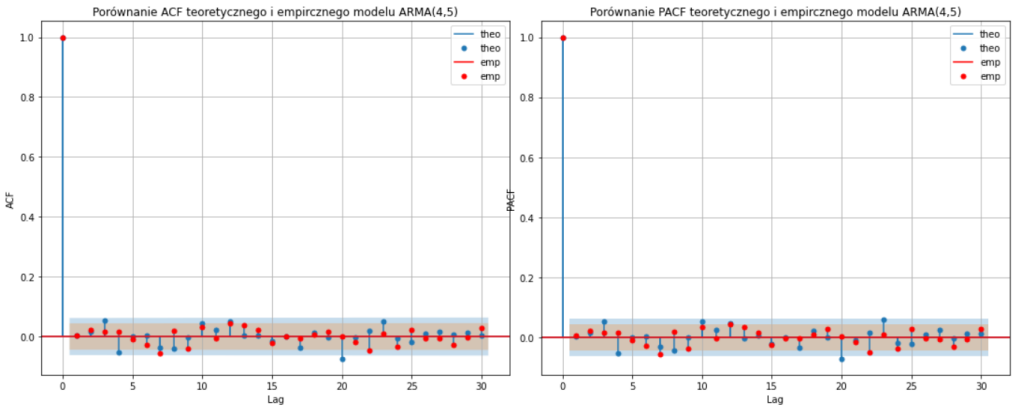
ar.L1	-0.2654
ar.L2	1.0378
ar.L3	-0.2244
ar.L4	-0.7190
ma.L1	0.2737
ma.L2	-1.0205
ma.L3	0.2566
ma.L4	0.7224
ma.L5	-0.0483
sigma2	1.5234

Rysunek 12 – Wartości estymatorów parametrów autoregresyjnych, średniej ruchomej i wariancji.

Wartości estymatorów ϕ_t na powyższej grafice jest przedstawiona jako ar.L(t), a wartości estymatorów θ_t, jako ma.L(t). Estymator wariancji, przedstawiono jako sigma2.

4. Ocena dopasowania modelu.

W celu oceny dopasowani modelu porównano, wykresy funkcji autokorelacji (ACF) i funkcji częściowej autokorelacji (PACF), teoretycznego modelu ARMA(4,5) z empirycznym modelem (Rys.13).

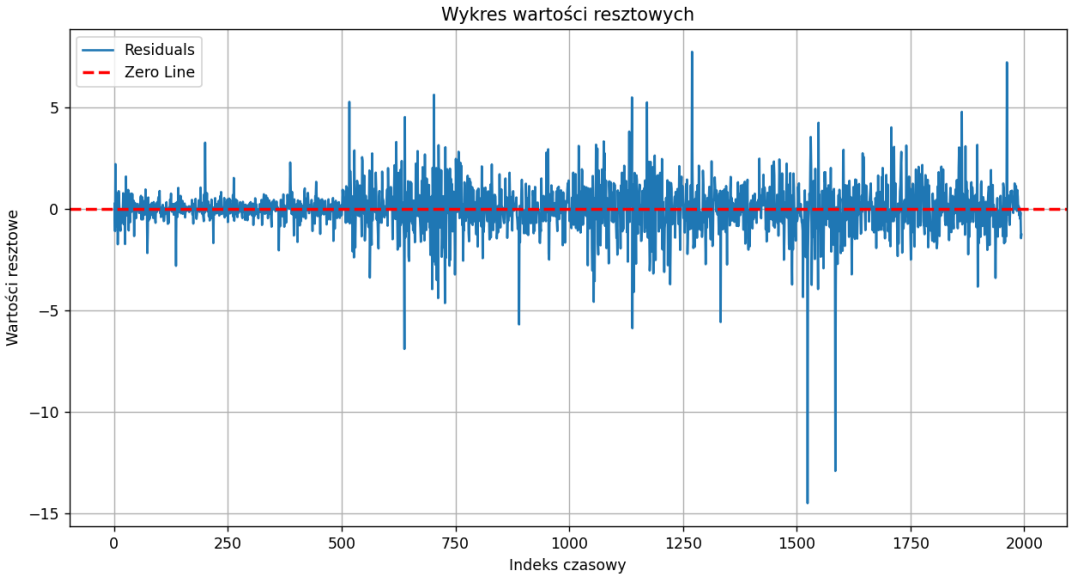


Rysunek 13– Porównanie wykresów ACF i PACF modelu empirycznego i teoretycznego.

Z poprzedniego wykresu, widać, że empiryczne wartości ACF i PACF pasują do teoretycznych wartości wygenerowanych przez model. Empiryczne wartości znajdują się w przedziałach ufności, sugeruje to, że model może być odpowiedni do opisu danych.

5. Weryfikacja założeń dotyczących szumu.

Dla weryfikacji założeń dotyczących białego szumu, wykonano wykres wartości resztowcyh (residuów) (Rys.14). Następnie zaczęto sprawdzać kolejno założenia, dotyczące: wartości średniej, wariancji, niezależności rozkładu.



Rysunek 14 – Wykres residuów.

5.1 Założenie o stałej zerowej wartości średniej.

Do weryfikacji stałej zerowej wartości średniej, wykonano t-test (t-Student test). Wyniki przeprowadzonego testu zapisano w poniższej tabelce (Rys.15).

Statystyka testowa:	-0.019
Wartość p (p-value):	0.985

Rysunek 15 – Wynik t-testu.

Z powyższej tabeli, widzimy, że statystyka testowa jest w przybliżeniu równa zero, a p-value jest w przybliżeniu równe 1. Świadczy, to o tym, że średnia próbki jest bardzo zbliżona do wartości porównawczej (porównywano do zera) i nie ma wystarczających dowodów do odrzucenia hipotezy zerowej, o stałej średniej, równej zero.

5.2 Założenie o stałej skończonej wariancji.

Do weryfikacji stałej skończonej (niezerowej) wariancji, wykonano Arch test (Auto-Regressive Conditional Heteroskedasticity test). Wyniki przeprowadzonego testu zapisano w poniższej tabelce (Rys.16).

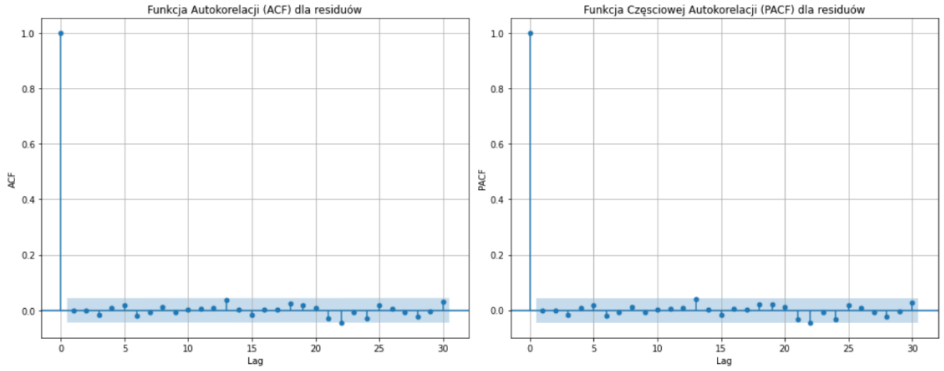
Statystyka LM:	23.613
Wartość p (p-value):	0.009
F-statystyka:	2.3764
Wartość p (p-value):	0.009

Rysunek 16 – Wynik Arch testu.

Wysoka wartość statystyki LM i niska p-wartość wskazują na odrzucenie hipotezy zerowej (posiadanie własności homoskedastyczności), sugerując obecność heteroskedastyczności. Podobnie jak w przypadku statystyki LM, wysoka wartość F-statystyki i niska p-wartość wskazują na to, że modele zmienności reszt są statystycznie znaczące, co oznacza, że reszty nie są jednorodne. Obie te statystyki świadczą silnie na heteroskedastyczności warunkowej w modelu. Świadczy to o tym, że badane wartości resztowe, nie posiadają własności homoskedastyczności, zmienne nie posiadają tej samej, skończonej wariancji.

5.3 Założenie o niezależności.

Do weryfikacji niezależności, wykonano wykresy funkcji autokorelacji (ACF) i funkcji częściowej autokorelacji (PACF), oraz przeprowadzono test Ljunga-Boxa. Wykresy jak i wynik testu zaprezentowano poniżej (Rys.17-18).



Rysunek 17 – Wykresy ACF i PACF dla residuów.

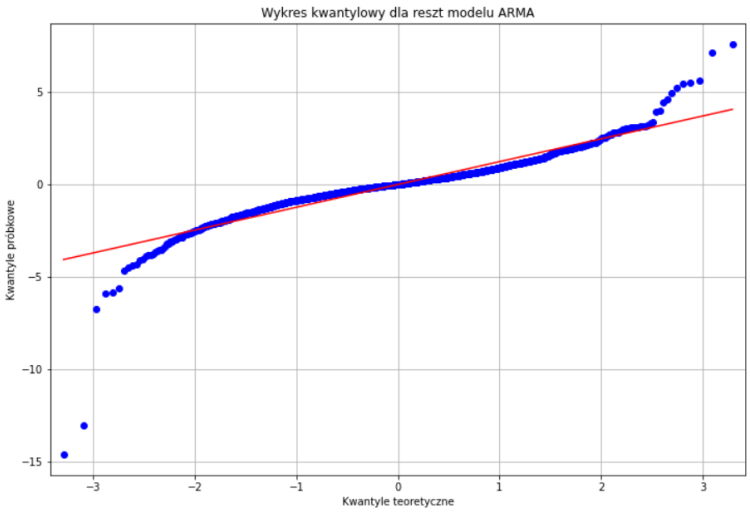
Statystyka testowa:	7.036
Wartość p (p-value):	0.722

Rysunek 18– Wynik testu Ljunga-Boxa.

Z wykresu funkcji autokorelacji, szybka zbieżność do zera wartości tej funkcji, sugerować, że dobrany model, dobrze uchwycił informacje zawarte w danych. Natomiast z wykresu częściowej autokorelacji, również widać, że wartości tej funkcji szybo zbiegają do zera, oraz „chowają" się (zawierają) w przedziale ufności, świadczy to o niezależności danych. Dodatkowo wysoka wartość statystyki, oraz wysoka wartość p-value, z testu Ljunga-Boxa, wskazuje na brak podstaw do odrzucenia hipotezy zerowej (braku autokorelacji w danych). Stąd na podstawie testu Ljunga-Boxa i wykresu funkcji częściowej autokorelacji (PACF), wnioskować możemy o niezależności danych.

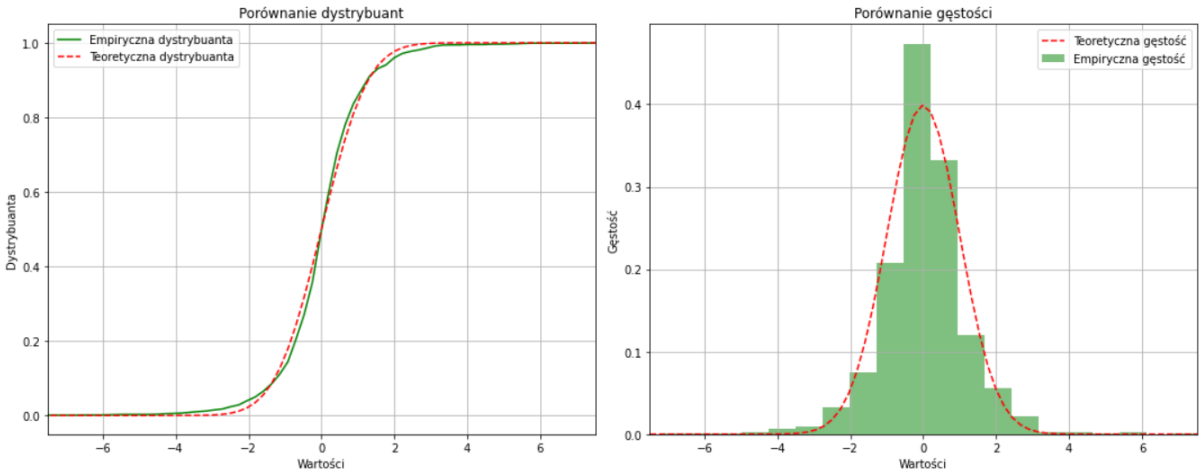
5.4 Założenie o rozkładzie normalnym $N(0, \sigma^2)$.

Do weryfikacji założenia o normalnej dystrybucji, wykonano porównanie, wykresów (Rys.19-20): dystrybuanty empirycznej z dystrybuantą teoretyczną, gęstości empiryczną z gęstości teoretyczną oraz kwantylowych empirycznych i kwantylowych teoretycznych. Dodatkowo przeprowadzono test Shapiro-Wilka. Wykresy jak i wynik testu zaprezentowano poniżej (Rys.19-21).



Rysunek 19 –Wykres kwantylowy teoretyczny i empiryczny.

Rysunek 20 – Wykresy dystrybuant i gęstości, teoretycznych i empirycznych.



Statystyka testowa:	0.880
Wartość p (p-value):	0.000

Rysunek 21 – Wynik testu Shapiro-Wilka.

Z porównania wykresu kwantylowego dla wartości teoretycznych i empirycznych, widać, że większość punktów znajduje się blisko linii, co wskazuje na to, że reszty modelu są zbliżone do rozkładu normalnego, ale są pewne odstające wartości. Punkty oddalone od linii wskazują na obserwacje, które są nietypowe lub odstające od oczekiwanego rozkładu. Z porównania dystrybuant i gęstości, teoretycznych i empirycznych wartości, również widzimy, że empiryczny rozkład jest podobny do rozkładu normalnego, ale również występują w nich różnice. Wątpliwości rozwiewa, wynik testu Shapiro-Wilka, wysoka wartość statystyk (bliska jedynce) sugeruje wysoki poziom podobieństwo rozkładu empirycznego do rozkładu normlanego, ale p-value na poziomie zero, świadczy o odrzucenia hipotezy zerowej, że badane dane, mają rozkład normlany i przyjęciu hipotezy alternatywnej o tym, że dane mają rozkład inny od normlanego.

6. Wnioski końcowe.

Głównym celem raportu, było dobranie do danych rzeczywistych modelu ARMA(p,q), przeprowadzenie oceny wytypowanego modelu oraz przeprowadzenie weryfikacji założeń białego szumu (whitenoise). Wybrano model ARMA(4,5), co kolejno oceniono jako w miarę dobre dopasowanie modelu, analizując porównanie wykresów funkcji autokorelacji (ACF) i funkcji częściowej autokorelacji (PACF) dla danych teoretycznych i empirycznych (Rys.13). Jednakowoż, pokazano, przez analizę wykresów i wyniki przeprowadzanych testów, że residua analizowanego modelu ARMA(4,5) nie spełniają założeń odnośnie stałej skończonej wariancji(Rys.16) i normlanej dystrybucji (Rys.19-21). Model ARMA(4,5) nie jest zatem dobrym modelem do modelowania danych cen otwarcia firmy Netflix. Prawdopodobnie bardziej zaawansowany model, z możliwymi innymi transformacjami danych do postaci szeregu stacjonarnego w słabym sensie.