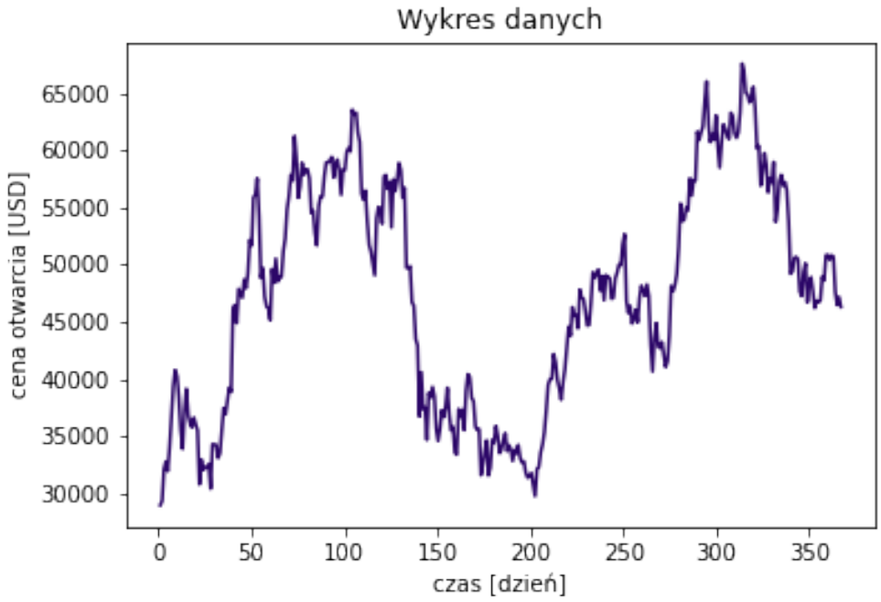
Aleksandra Janczewska

Analiza danych z wykorzystaniem modeli ARMA

1. Opis danych.

Dane, które będę analizować w moim sprawozdaniu pochodzą ze strony https://finance.yahoo.com i składają się z 366 obserwacji na przestrzenie czasu od 01.01.2021 do 01.01.2022 z częstotliwością jednego dnia. Dane dotyczą codziennej ceny otwarcia kursu Bitcoina (w USD) na przestrzeni roku.

1.1. Wykres danych.

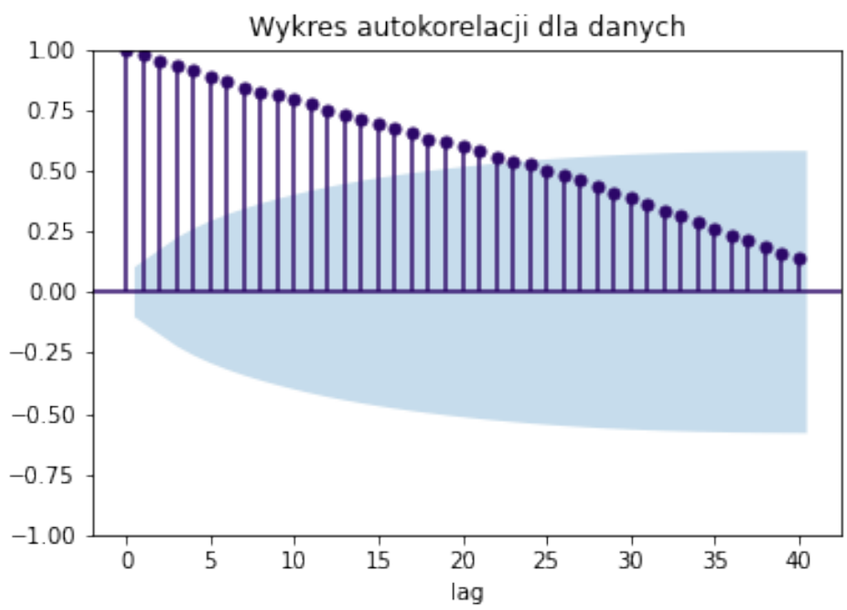


Rysunek 1.1: wykres danych

1.2. Wykres funkcji autokorelacji danych.

Autokorelację danych będę sprawdzać za pomocą empirycznej funkcji autokorelacji (ACF). Jak widać na wykresie 1.2, dane są skorelowane. Wartość korelacji maleje liniowo wraz ze zwiększaniem przesunięcia. Prowadzi to do stwierdzenia, że dane zawierają komponent deterministyczny zależny od czasu. Prawdopodobnie jest to trend liniowy.

Na tym etapie można więc stwierdzić, że dane nie mają postaci stacjonarnej.



Rysunek 1.2: wykres ACF dla danych

2. Transformacja danych.

W kolejnym kroku będę transformować dane do postaci stacjonarnej, aby następnie móc dobrać do nich odpowiedni model ARMA(p, q).

Postać stacjonarna danych, to taka, w której:

⋆ dane nie zawierają komponentów deterministycznych - brak trendu i sezonowości,
⋆ średnia oraz wariancja są w przybliżeniu stałe,
⋆ funkcja autokorelacji zbiega do wartości 0, dla coraz większych lagów.

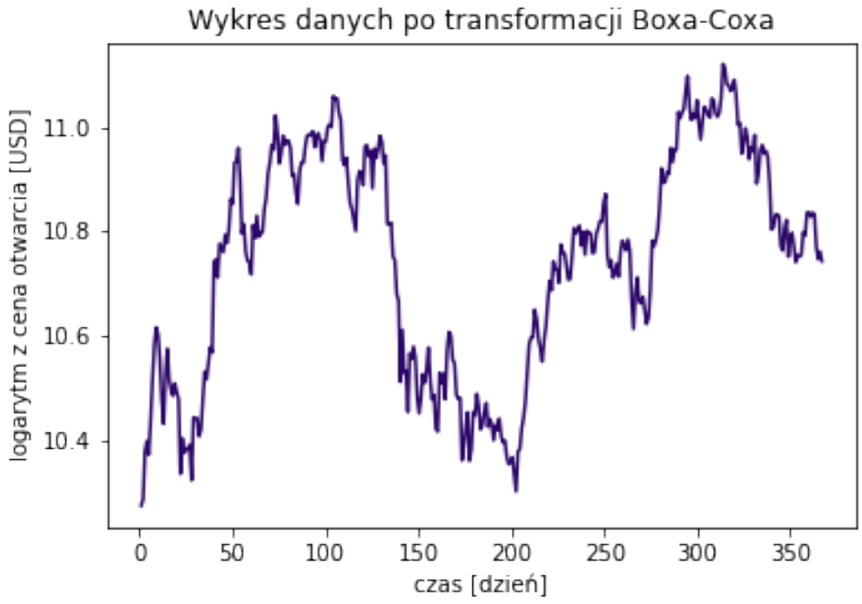
2.1. Transformacja Boxa-Coxa.

W celu stabilizacji wariancji danych, zastosuję transformatę Boxa-Coxa. Transformacja ta wyraża się następującym wzorem:

$$Y^{\lambda}(X) = \begin{cases} \frac{X^{\lambda}-1}{\lambda} & \text{dla } \lambda \neq 0 \\ ln(X) & \text{dla } \lambda = 0 \end{cases}$$

gdzie:

⋆ $Y^{\lambda}(X)$ to przekształcona obserwacja,
⋆ X to zmienna, którą przekształcamy,
⋆ λ to parametr, w naszym przypadku bierzemy $\lambda = 0$.



Rysunek 2.1: wykres danych po transformacji Boxa-Coxa

Na wykresie 2.1 przedstawione są dane po transformacji Boxa-Coxa. Skala na osi y zmieniła się z zakresu od około 30000 do 65000 (widoczne na rysunku 1.1), na zakres w przybliżeniu od 10 do 11. Dzięki temu znacznie ograniczamy wariancję danych.

2.2. Usunięcie trendu i sezonowości.

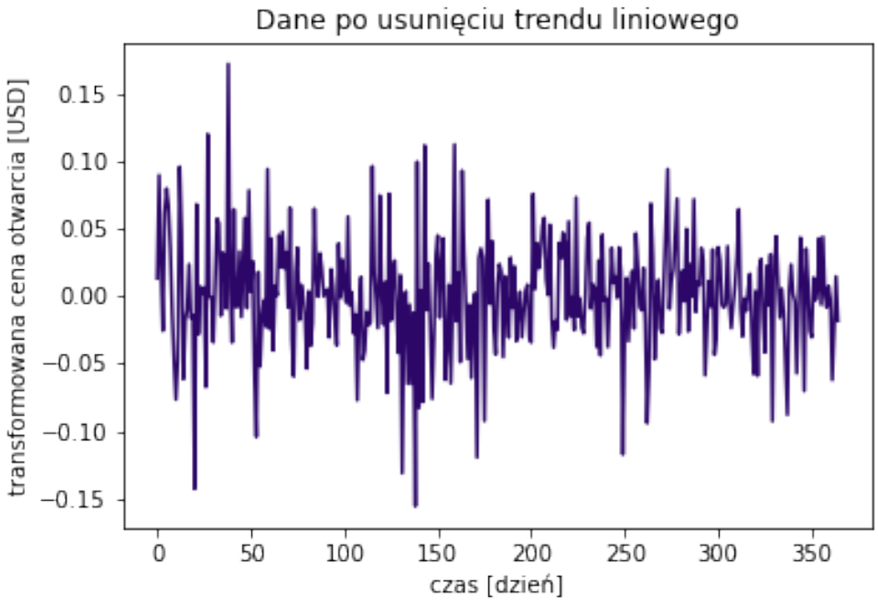
Trend liniowy z danych będę usuwać metodą różnicowania rzędu 1. Korzystam ze wzoru:

$$Y_t = X_t - X_{t-1},$$

gdzie:

⋆ X_t, X_{t-1} to obserwacje w poszczególnych momentach czasu t,
⋆ Y_t to dane po różnicowaniu.

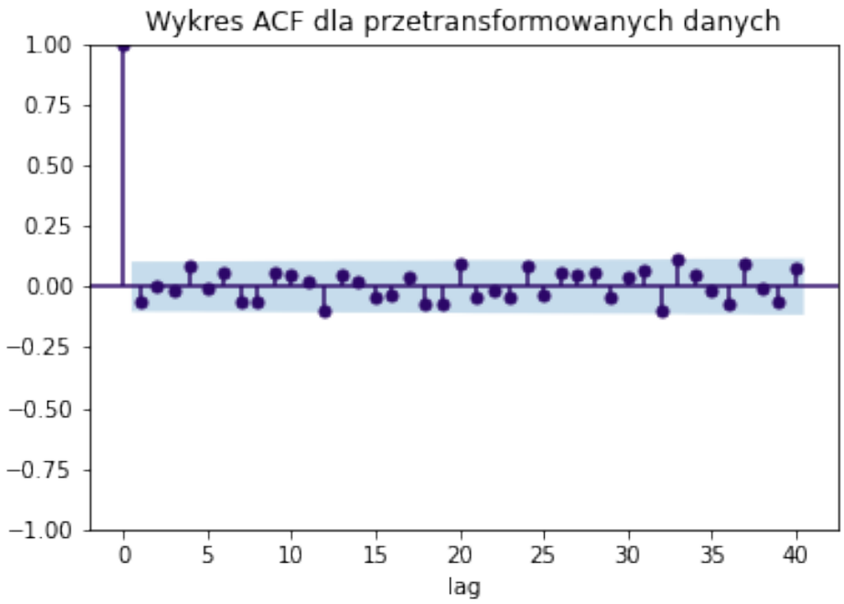
Wykres 2.2 przedstawia dane po usunięciu trendu liniowego. Na tym etapie można stwierdzić, że dane nie zawierają komponentu okresowego, dlatego też dalsza ich transformacja może prowadzić jedynie do nadmiernego dopasowania.



Rysunek 2.2: wykres danych po usunięciu trendu

2.3. Sprawdzenie stacjonarności danych po transformacji.

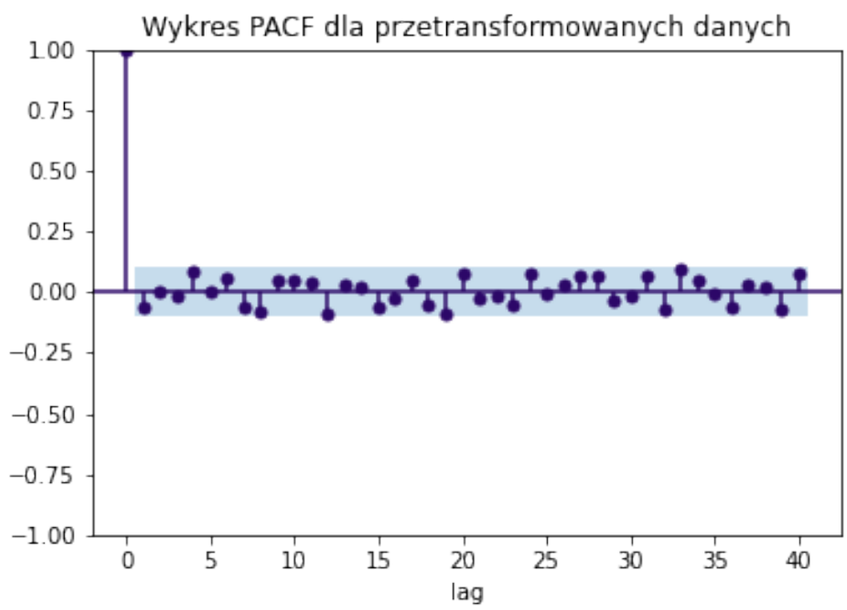
Po zastosowaniu transformacji Boxa-Coxa oraz metody różnicowania, dane nie zawierają komponentów deterministycznych. Średnia obserwacji oscyluje wokół wartości 0 (wynosi ona w przybliżeniu 0.00128). Wariancja natomiast, jak możemy zauważyć na wykresie 2.2 jest w przybliżeniu stała.



Rysunek 2.3: wykres ACF przetransformowanych danych

Na rysunku 2.3 oraz 2.4 widać, że wartości funkcji autokorelacji oraz częściowej funkcji autokorelacji dla poszczególnych lagów oscylują wokół wartości 0, co świadczy o tym, że dane po transformacji są nieskorelowane.

Podsumowując, można stwierdzić, że skoro dane nie zawierają już trendu ani sezonowości, ich średnia jest równa około 0, wariancja jest w przybliżeniu stała, a funkcje ACF oraz PACF oscylują wokół 0, zostały one przekształcone do postaci stacjonarnej.



Rysunek 2.4: wykres PACF przetransformowanych danych

3. Dopasowanie modelu ARMA(p, q).

3.1. Teoretyczny model ARMA(p, q).

Model ARMA(p, q) jest to model autoregresyjny (z rzędem p) średniej ruchomej (z rzędem q). Szereg czasowy $\{X_t\}$ jest szeregiem ARMA(p, q), jeżeli spełnia następujące założenia:

1° jest szeregiem stacjonarnym w słabym sensie ($EX_t = \text{const}$ oraz jego funkcja autokowariancji nie zależy od czasu),

2° spełnia równanie:

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

gdzie $\{Z_t\} \sim \text{WN}(0, \sigma^2)$ oraz wielomiany:
$\phi(z) = z_t - \phi_1 \cdot z_{t-1} - \ldots - \phi_p \cdot z_{t-p}$ i $\theta(z) = z_t + \theta_1 \cdot z_{t-1} + \ldots + \theta_q \cdot z_{t-q}$ nie mają wspólnych pierwiastków.

3.2. Kryteria informacyjne.

W celu dobrania rzędów modelu $p, q \in \mathbf{N}$, będę korzystać z kryterium informacyjnego AICc. Kryterium AICc jest to skorygowane kryterium AIC używane dla małych próbek.

Kryterium informacyjne Akaike (AIC) jest estymatorem błędu predykcji doboru modelu do danych. AIC szacuje jakość modelu. Rozważając zbiór różnych modeli ARMA(p,q), im wartość statystyki AICc dla danego modelu jest mniejsza, tym lepiej dobrany jest ten model. Daje to podstawę do wyboru najlepszego ze sprawdzanych modeli.

3.3. Model dobrany do danych.

W celu dobrania modelu do danych oraz estymacji jego parametrów korzystam z programu ITSM. Sprawdzam jakość dopasowania modelu dla wszystkich konfiguracji rzędów p oraz q modelu ARMA(p, q) dla $p \in [0, 10]$ i $q \in [0, 10]$.

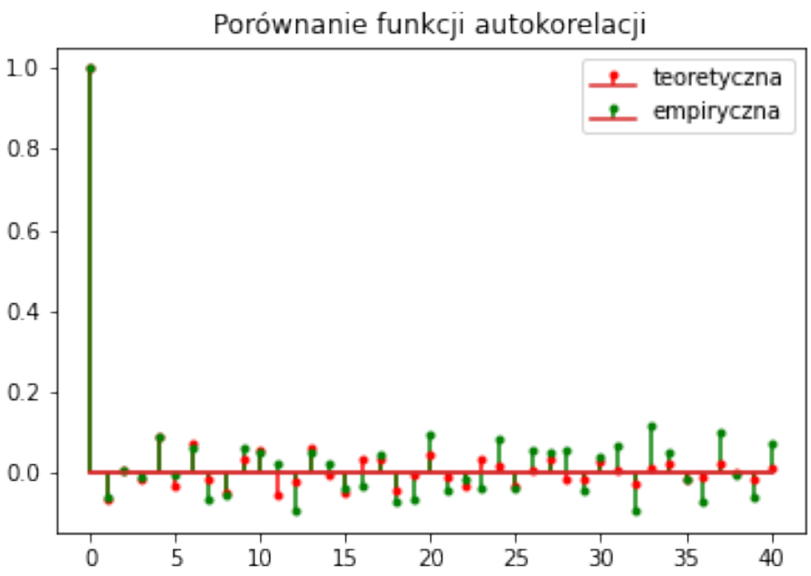
Program wyliczył, że najlepiej dopasowanym modelem jest ARMA(6, 2). Wartość statystyki AICc dla tego modelu wynosi -1272.47. Model jest opisany równaniem:

$$\begin{aligned} X_t = &-0.4673X_{t-1} - 0.9495X_{t-2} - 0.07505X_{t-3} + 0.07407X_{t-4} + \\ &0.0005091X_{t-5} + 0.1367X_{t-6} + Z_t + 0.4121Z_{t-1} + 0.95Z_{t-2}, \end{aligned}$$

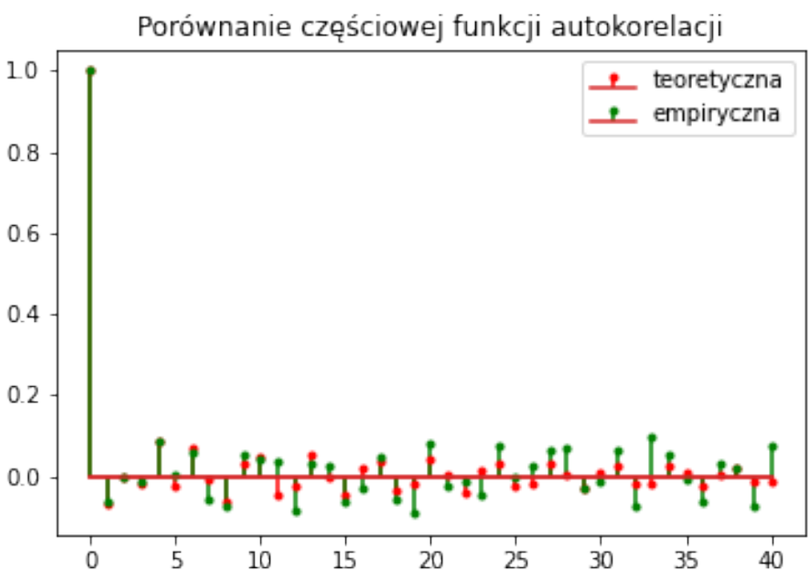
gdzie $\{Z_t\} \sim \text{WN}(0, \sigma^2 = 0.001694)$.

4. Sprawdzenie dopasowania modelu do danych.

4.1. Porównanie empirycznych oraz teoretycznych ACF i PACF.



Rysunek 4.1: porównanie ACF



Rysunek 4.2: porównanie PACF

Wykresy funkcji ACF (rysunek 4.1) oraz PACF (rysunek 4.2) są do siebie w znacznym stopniu podobne. Na obu wykresach widać, że funkcje teoretyczne dla modelu ARMA(6, 2) oraz empiryczne, to znaczy te wyznaczone z danych, pokrywają się tylko w niektórych punktach. Na podstawie tych wykresów można więc wywnioskować, że model ARMA(6, 2) nie jest najlepiej dobranym modelem dla analizowanych danych.

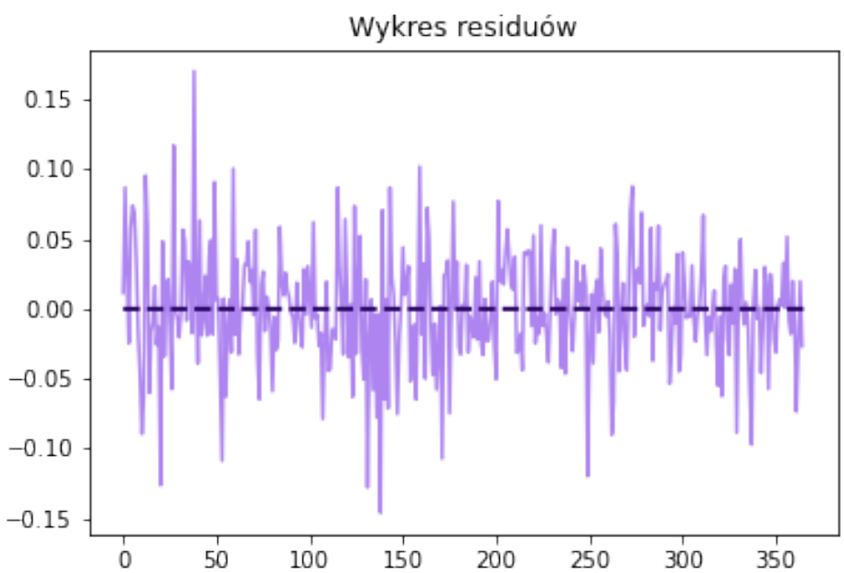
4.2. Analiza residuów.

W poprawnie dobranym modelu muszą być spełnione następujące założenia dotyczące residuów (ozn. ε_i) $\forall i = 1, ..., n$:

1° $E(\varepsilon_i) = 0$,
2° $var(\varepsilon_i) = \sigma^2 = const$,
3° ε_i są niezależne,
4° $\varepsilon_i \sim N(0, \sigma^2)$.

Analiza średniej oraz wariancji (warunek 1° oraz 2°).

Do ogólnej analizy średniej oraz wariancji będę korzystać z wykresu residuów (4.3).



Rysunek 4.3: wykres residuów modelu

Na wykresie 4.3 znajdują się wyplotowane residua i jest zaznaczona ich średnia, która oscyluje wokół wartości 0 ($E(\varepsilon_i) \approx -2.4 \cdot 10^{-5} \approx 0$). Wartości residuów mieszczą się w przedziale od -0.15 do 0.15. Na podstawie wykresu możemy więc stwierdzić, że wariancja jest w przybliżeniu stała.

Analiza niezależności (warunek 3°).

W celu analizy niezależności residuów, sprawdzę ich korelację za pomocą funkcji autokorelacji (ACF) oraz częściowej funkcji autokorelacji (PACF). Jeżeli zmienne są nieskorelowane to są również niezależne.

Rysunek 4.4 prezentuje funkcję autokorelacji, zaś wykres 4.5 częściową funkcję autokorelacji dla residuów. Na obu wykresach widać, że słupki należą do obszarów na nich zaznaczonych. Świadczy to o tym, że residua nie są skorelowane, a co się z tym wiąże, są niezależne.

Analiza rozkładu residuów (warunek 4°).

Sprawdzę teraz, czy residua mają rozkład $N(0, \sigma^2)$. W tym celu wykonam testy statystyczne oraz sprawdzę pokrycie danych z rozkładem teoretycznym na wykresach.

Testy statystyczne.

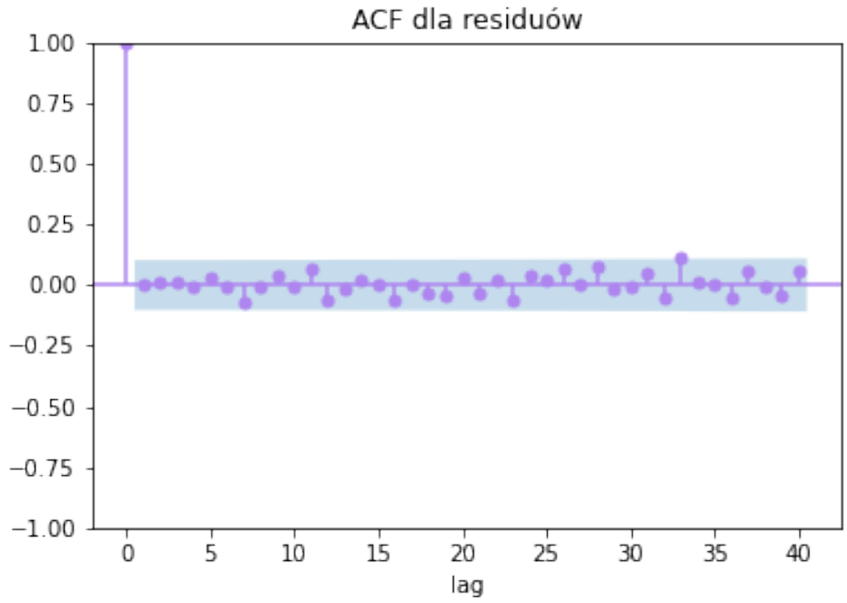
Jeżeli p-value w testach jest większa od poziomu istotności α, wtedy próba ma rozkład normalny. Ustalam poziom istotności $\alpha = 0.05$ i wykonuję testy.

Test Kołmogorowa-Smirnowa: p-value w teście wynosi p-value $\approx 3.303 \cdot 10^{-69} < 0.05$. Oznacza to, że według testu Kołmogorowa-Smirnowa residua nie mają rozkładu normalnego.

Test Jarque-Bera: p-value w teście wynosi p-value $\approx 2.594 \cdot 10^{-6} < 0.05$. Według tego testu residua również nie mają rozkładu normalnego.

Wykresy.

Na rysunku 4.6 pokazane jest porównanie dystrybuanty teoretycznej z rozkładu normalnego z parametrami średniej $\mu = E(\varepsilon_i) \approx 0$ oraz wariancji $var(\varepsilon_i) \approx 0.0017$ ($N(\mu = 0, \sigma^2 = 0.0017)$) z dystrybuantą empiryczną z próby residuów. Dystrybuanty te pokrywają się w większości punktów. Są jednak miejsca, w których na siebie nie nachodzą.

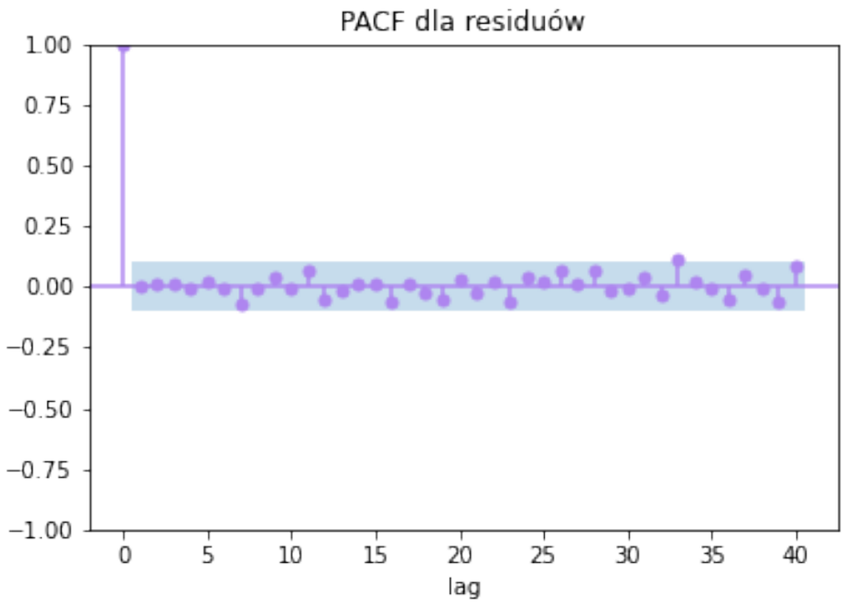


Rysunek 4.4: wykres ACF residuów

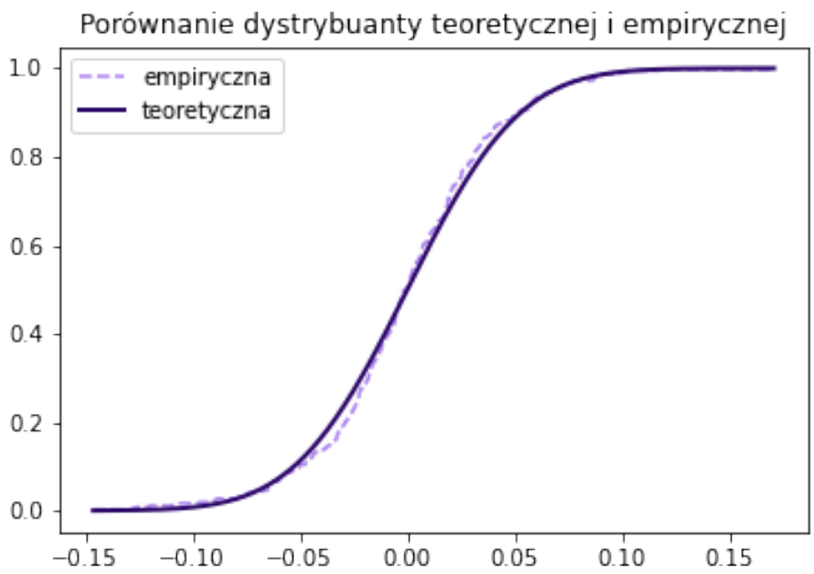
Na wykresie 4.7 prezentowany jest histogram residuów z nałożoną na niego gęstości teoretyczną z rozkładu normalnego ($N(\mu = 0, \sigma^2 = 0.0017)$). Jak można zauważyć gęstość i histogram w pełni się nie pokrywają. Jednak kształtem histogram jest zbliżony do rozkładu normalnego.

Wnioski.

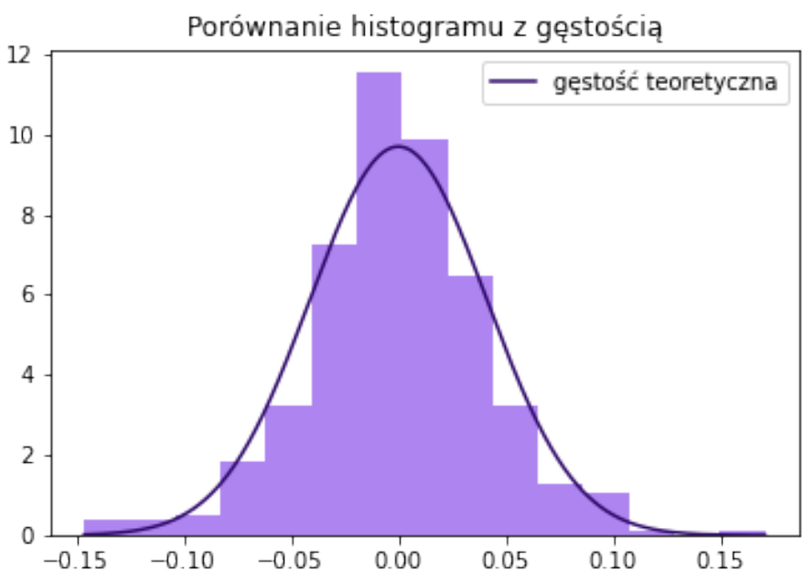
Na podstawie wykresów możemy stwierdzić, że residua spełniają warunek 1° oraz 2°. Ich średnia oscyluje w okolicach wartości 0, a ich wariancja jest w przybliżeniu stała. Spełniony jest również warunek 3° dotyczący tego, że residua są niezależne. Warunek 4° sprawdzający normalność rozkładu residuów nie jest jednak do końca spełniony. Według testów statystycznych residua nie mają rozkładu normalnego. Porównanie histogramu z gęstością teoretyczną oraz dystrybuant wskazuje na to, że residua mogą mieć rozkład normalny. Jednakże na podstawie tych wykresów nie jesteśmy w stanie stwierdzić, że tak jest na pewno.



Rysunek 4.5: wykres PACF residuów



Rysunek 4.6: dystrybuanty



Rysunek 4.7: histogram i gęstość

5. Podsumowanie oraz wnioski.

W sprawozdaniu przeanalizowałam dane dotyczące codziennego kursu Bitcoina (w USD) na przestrzeni roku, dobierając do nich model ARMA. Na początku dane wejściowe wymagały transformacji do postaci stacjonarnej, aby spełniały one założenia modelu. Następnie za pomocą programu ITSM, sprawdzając różne kombinacje rzędów p oraz q modeli ARMA(p, q), dobrałam model ARMA(6, 2). W kolejnym kroku sprawdzałam jakość dopasowania modelu ARMA(6, 2) do danych. Porównując empiryczne oraz teoretyczne funkcje ACF i PACF wyszło, że na ich podstawie model nie jest najlepiej dopasowany do danych. Analizując residua modelu mogę stwierdzić, że spełniają one prawie wszystkie założenia (z wyjątkiem tego o rozkładzie normalnym).

Podsumowując, na podstawie powyższych analiz stwierdzam, że model ARMA(6, 2) nie jest najlepszym modelem do prognozowania wybranych przeze mnie danych. Mimo, że częściowo model ten pasuje do danych, nie jest to idealne dopasowanie. Jednak biorąc pod uwagę fakt, że modelowane dane są rzeczywiste, model do nich dobrany prawie nigdy nie będzie idealnie dopasowany.

6. Źródła.

* https://pl.wikipedia.org/wiki/Przekszta%C5%82cenie_Boxa-Coxa
* https://en.wikipedia.org/wiki/Akaike_information_criterion