Analiza danych rzeczywistych przy pomocy modelu ARMA

Spis treści:

1. Wstęp

Cel pracy

Niniejszy raport skupia się na zastosowaniu modelu ARMA do modelowania temperatury powietrza na wyspie Jan Mayen za ostatnie 10 lat. Celem analizy jest nie tylko identyfikacja wzorców i potencjalnych anomalii, ale także próba przewidzenia przyszłych trendów, które mogą mieć znaczący wpływ na gospodarkę, ekosystemy i życie codzienne społeczności lokalnych.

Informacja o danych oraz wizualizacja danych

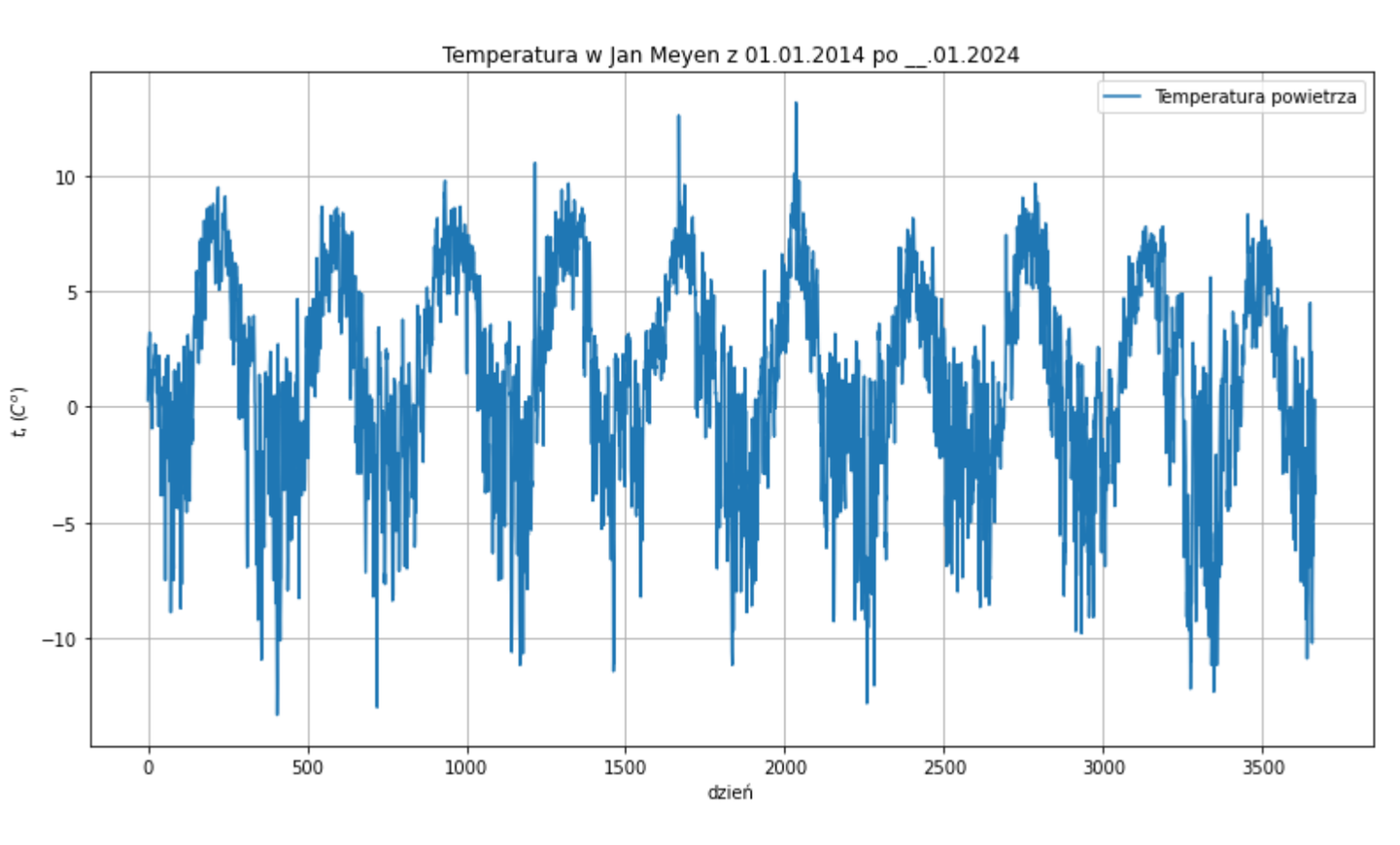
Dane są pobrane ze strony https://www.ncei.noaa.gov/ (NCEI - National Centers for Environmental Information).

> NCEI provides environmental data, products, and services covering the depths of the ocean to the surface of the sun to drive resilience, prosperity, and equity for current and future generations.

Dane dotyczą dziennej temperatury powietrza na wyspie Jan Mayen. Jan Mayen to wyspa wulkaniczna w Arktyce, ok. 500 km na wschód od Grenlandii, oblewają ją wody Oceanu Arktycznego: Morze Grenlandzkie od północy, Cieśnina Duńska od zachodu i Morze Norweskie od południa i wschodu. Od 1930 r. administracyjnie przynależy do Norwegii.

Probka zawiera dane za ostatnie 10 lat plus 23 dni 2024 roku.

Poniżej jest przedstawiony wykres danych, możemy zobaczyć wyraźną sezonowość:



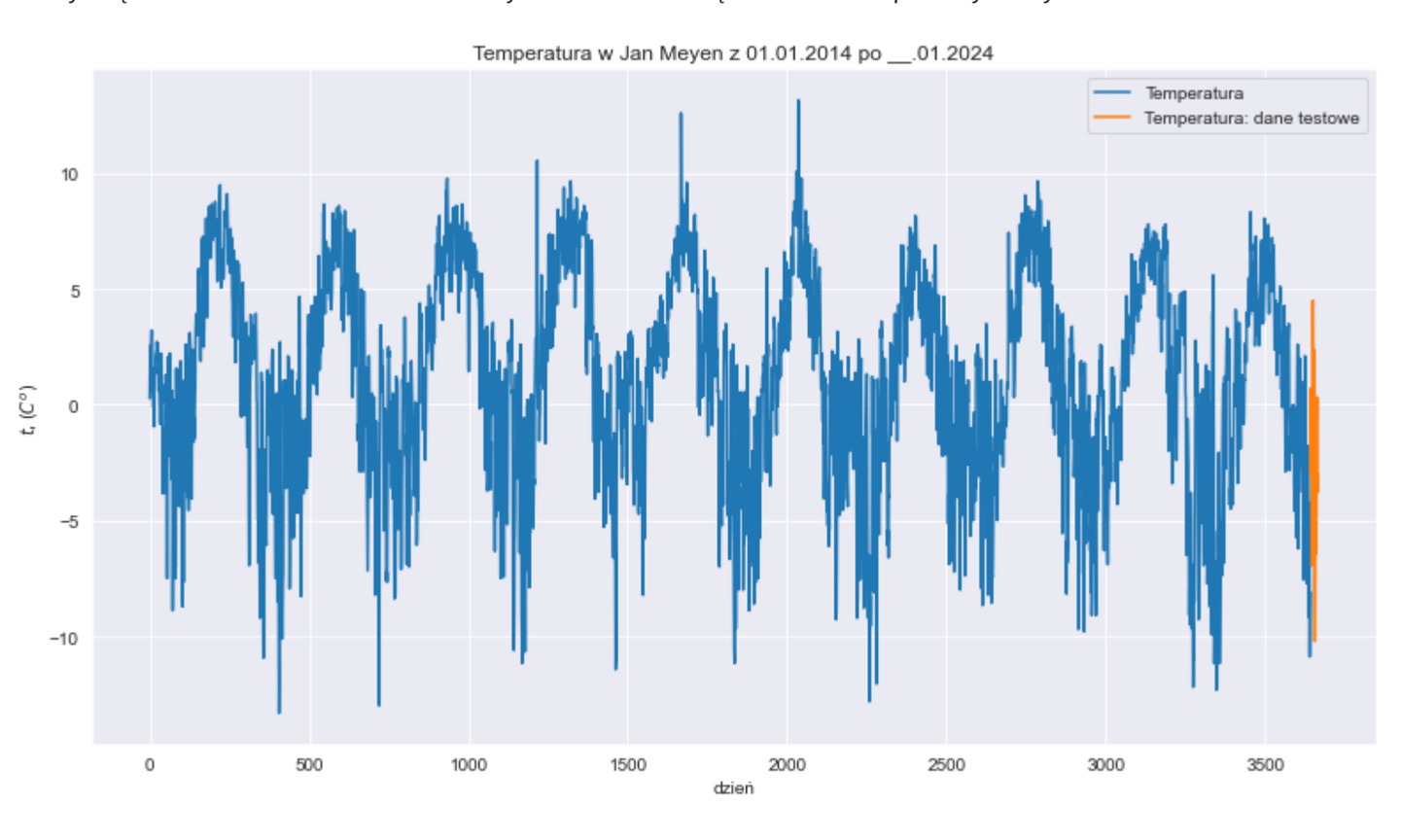
2. Przygotowanie danych do analizy

Zbadanie jakości danyh

W próbce jest 9 braków danych w dniach '2014-12-17', '2015-03-01', '2015-03-02', '2018-09-17', '2020-10-29', '2021-01-21', '2021-02-11', '2021-02-12', '2023-06-25'. Odnośnie 3666 to nie dużo, więc nie będzie miało dużego wpływu na modelowanie.

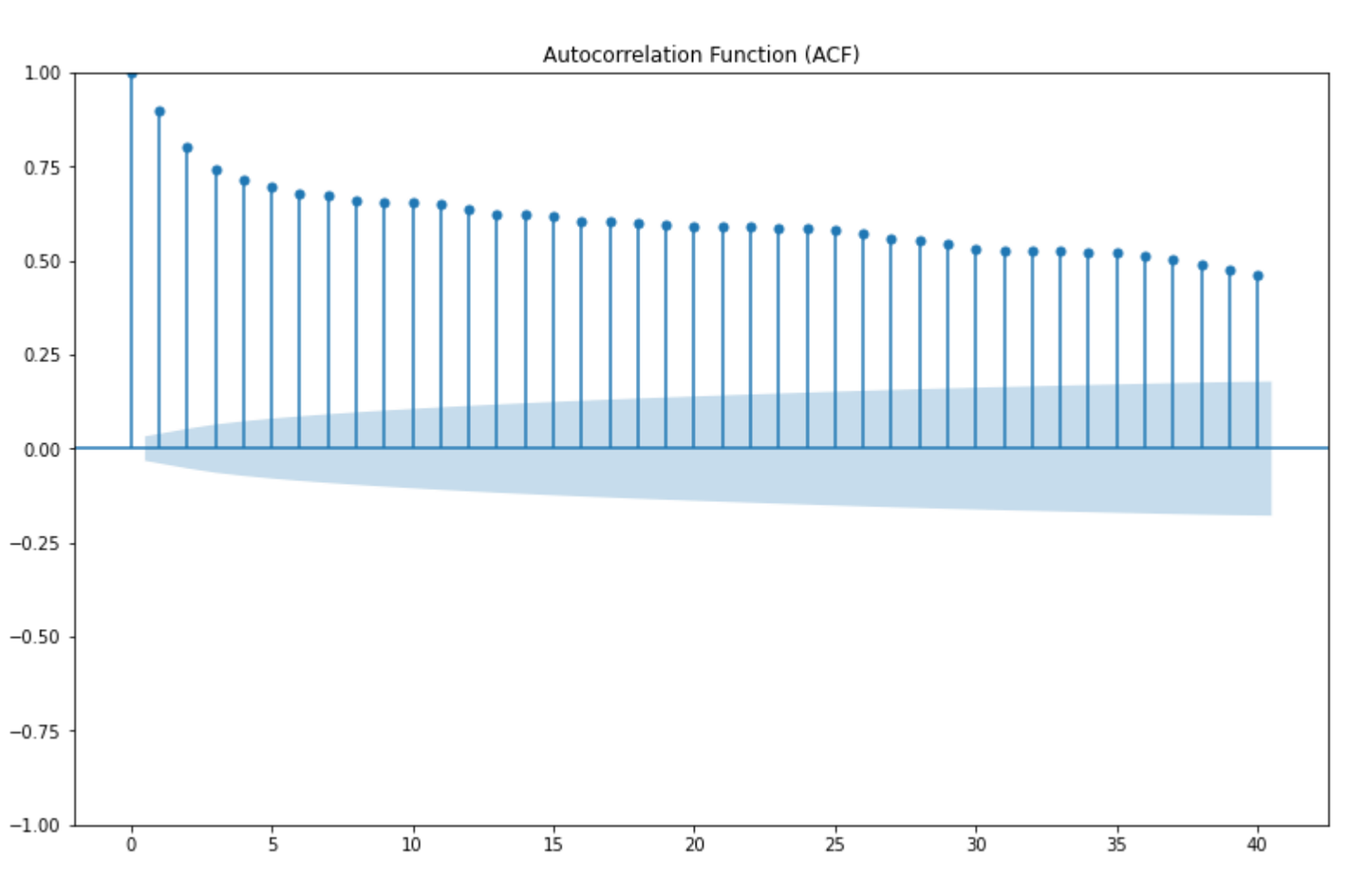
W danych nie ma odstających wartości. Liczyłam wartości za odstające jeżeli były z poza przedziału $[(x_{25} - 1{,}5 \cdot (x_{75} - x_{25})),\ (x_{75} + 1{,}5 \cdot (x_{75} - x_{25}))]$, gdzie x_{75} oraz x_{25} to kwantyli rzędów $0.75, 0.25$ odpowiednio.

Też wyodrębniłam dane testowe. Zbior testowy bedzie składać się z wartości temperatury za styczeń 2024 roku.

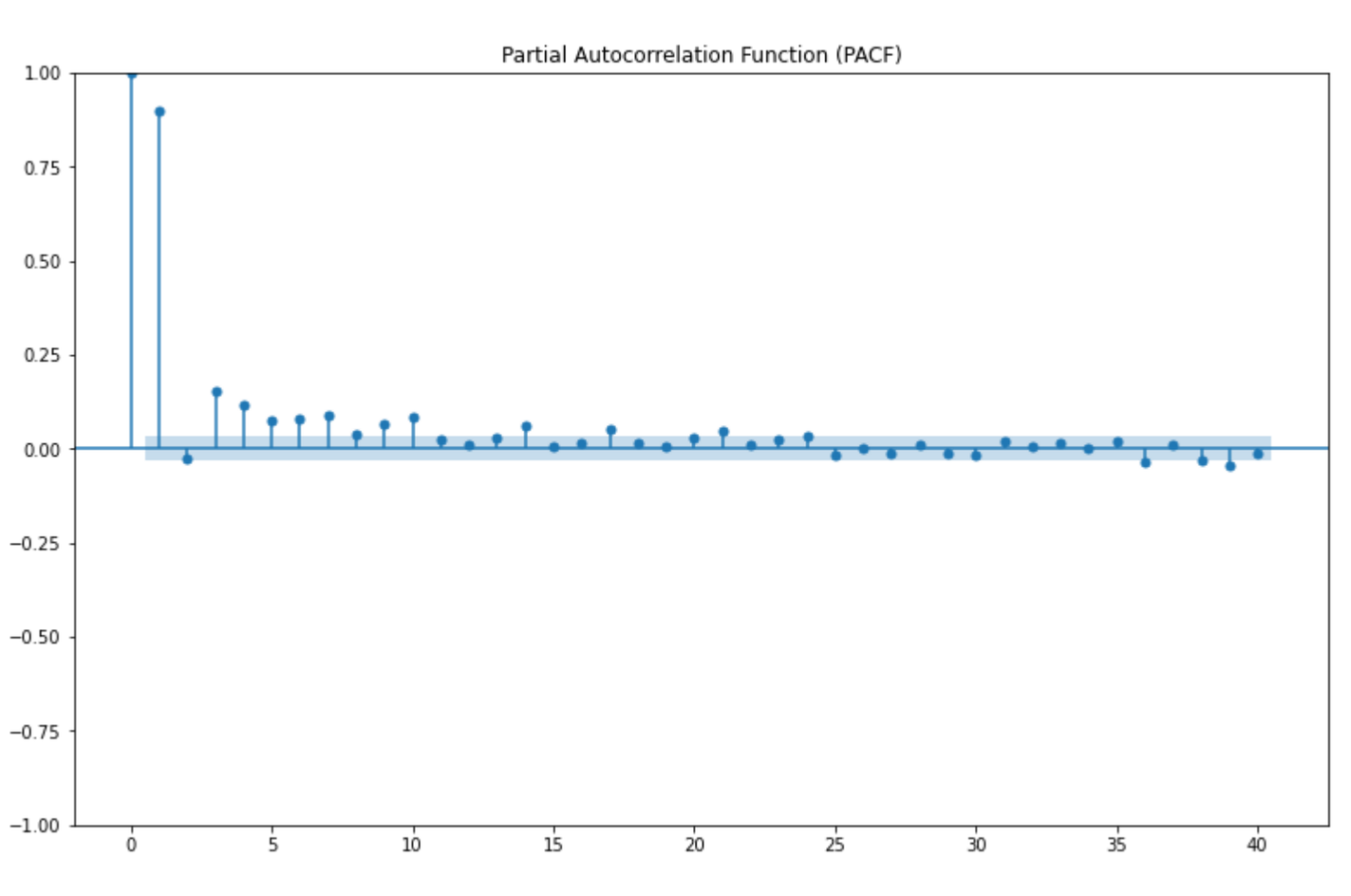


Dekompozycja szeregu czasowego

Wykres ACF oraz PACF



Wartości ACF stopniowo maleją, co może wskazywać na to, że szereg czasowy wykazuje pewien trend lub efekt autoregresyjny. Wartości są dodatnie i powoli spadają z każdym kolejnym opóźnieniem, co może sugerować, że proces ma naturę autoregresyjną. Brak gwałtownego spadku do zera po pierwszych kilku opóźnieniach sugeruje, że model ARMA może być bardziej odpowiedni niż czysty model AR lub MA. Wszystkie wartości ACF mieszczą się wewnątrz pasm ufności, co może wskazywać, że nie ma silnej autokorelacji w żadnym konkretnym opóźnieniu.



Pierwsze opóźnienie jest znaczące i wysokie, co wskazuje na silną autokorelację częściową w pierwszym opóźnieniu. Wartości PACF dla pozostałych opóźnień szybko spadają niemal do zera i oscylują wokół niego, pozostając wewnątrz pasm ufności. Oznacza to, że nie ma już dodatkowej autokorelacji częściowej po uwzględnieniu pierwszego opóźnienia.

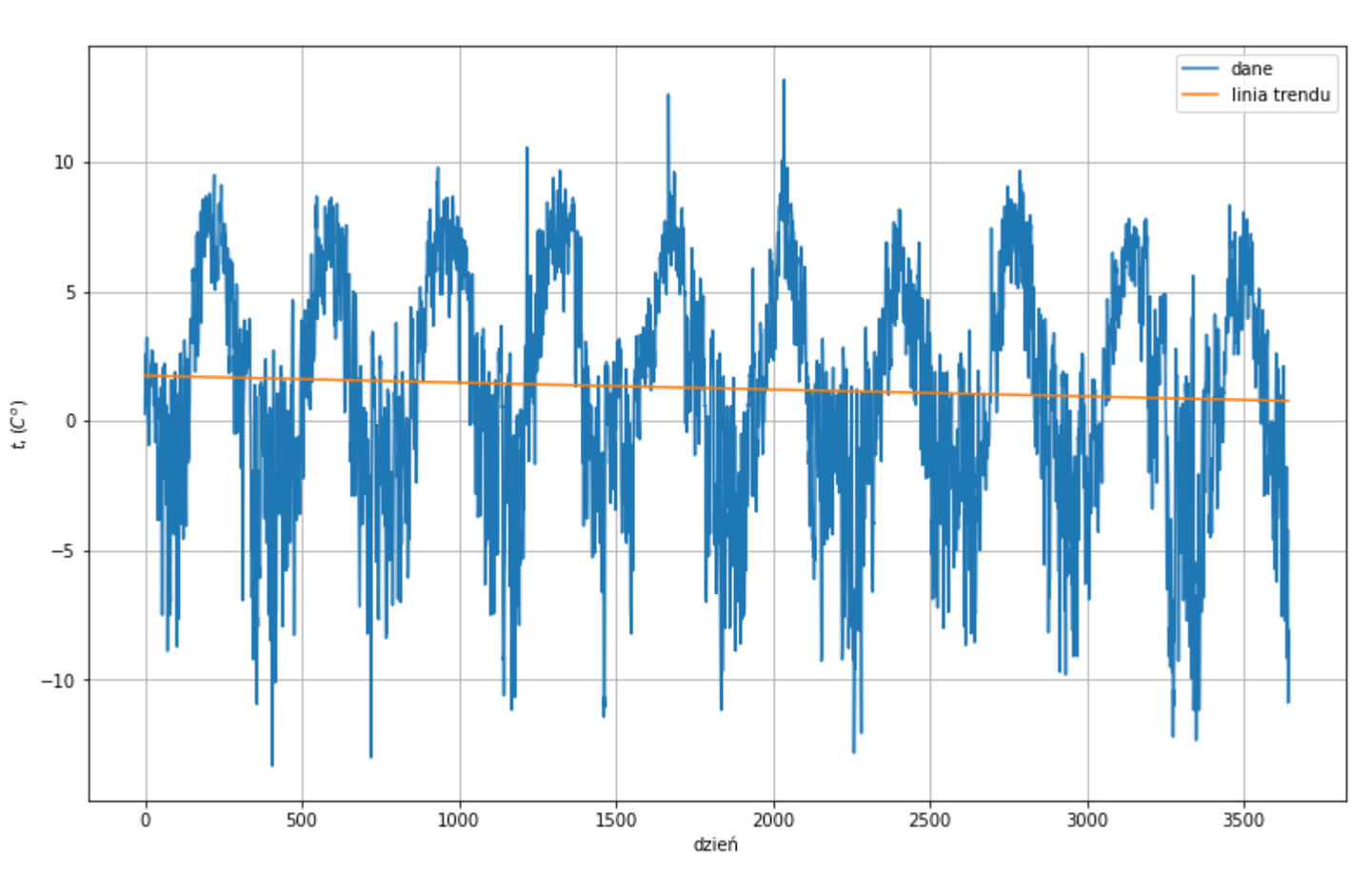
Dickey - Fuller test:

Statystyka testowa wynosi -3.896228, jest mniejsza niż wartości krytyczne, jakie wynoszą $-3.432, -2.862, -2.567$ dla $\alpha = \{0.01, 0.05, 0.1\}$, sugeruje to, że szereg czasowy jest stacjonarny. P-wartość wynosi 0.002063 jest mniejsza niż 0.05, sugeruje to, że szereg czasowy jest stacjonarny z poziomem ufności 95%.

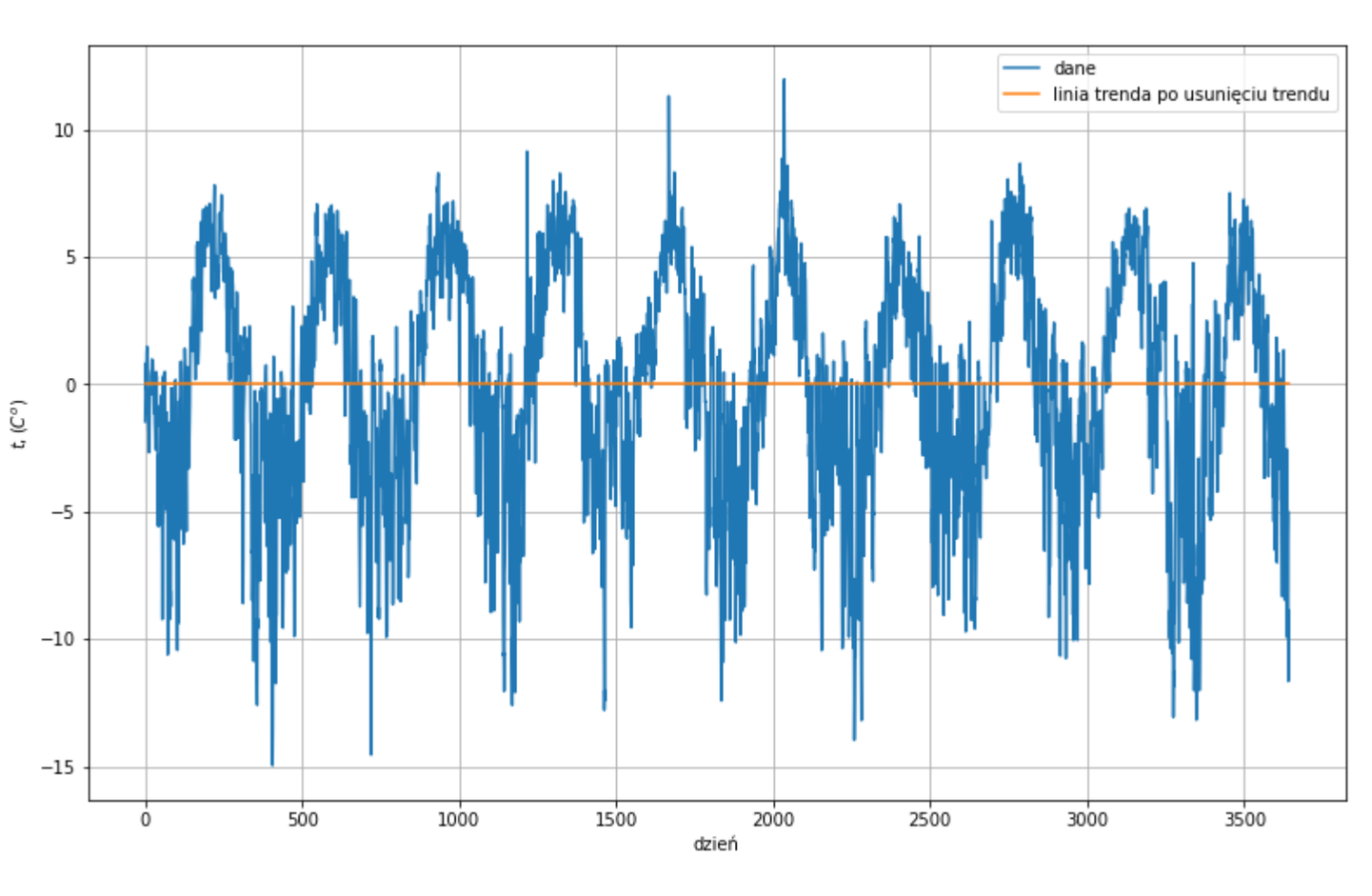
Identyfikacja trendów deterministycznych

Usunięcie trendu:

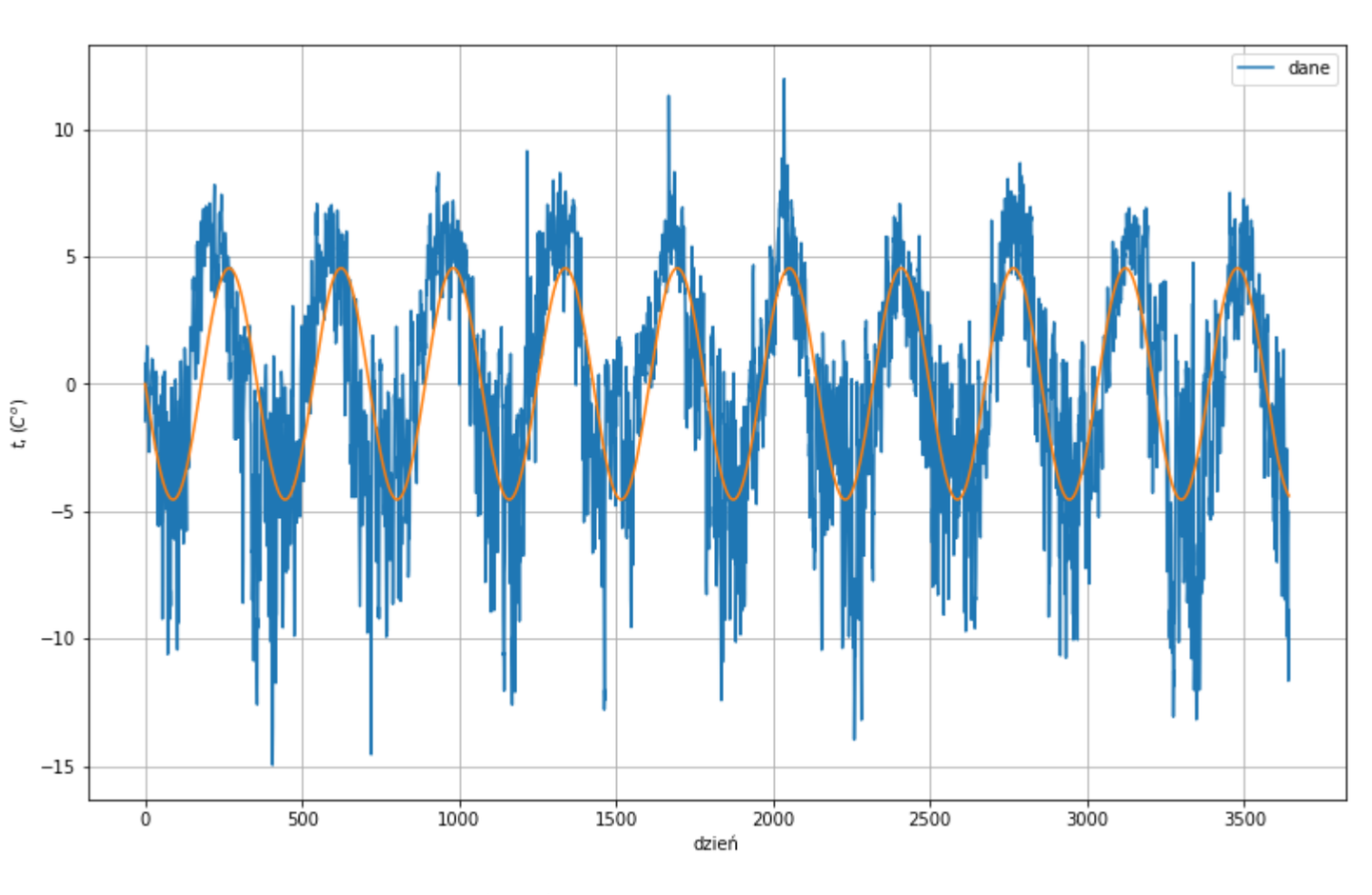
Narysujemy linię trendu



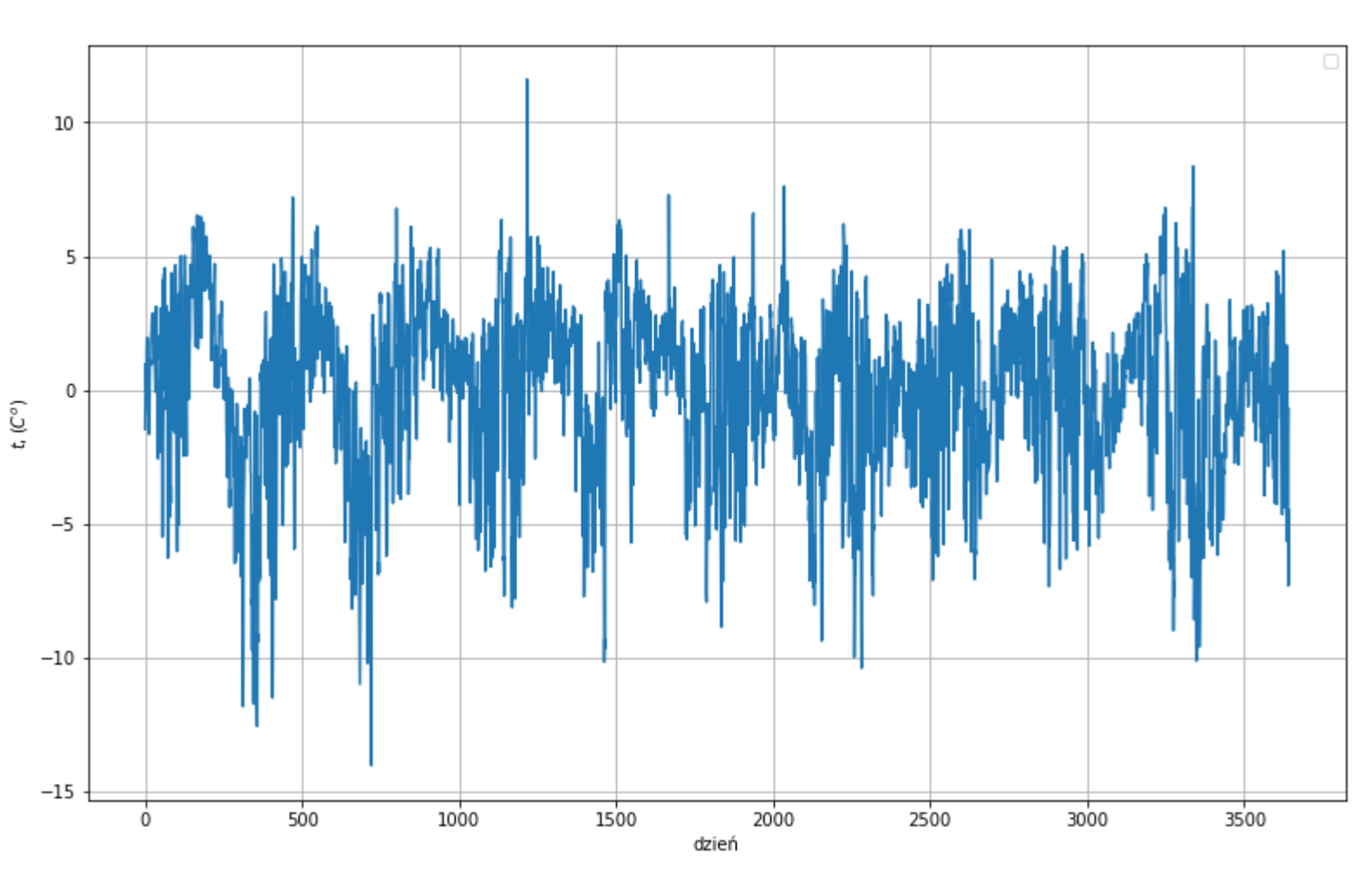
Do usunęcia trendu skorzystałam z regresji liniowej: dobrałam parametry β_0 oraz β_1 do równania $y = \beta_0 + \beta_1 x$. Linia trendu ma równanie $y = -0.000268x + 1.745$. Po usunięciu trendu wykres troche przeniósł się do dołu.



Narysujemy sezonowość:

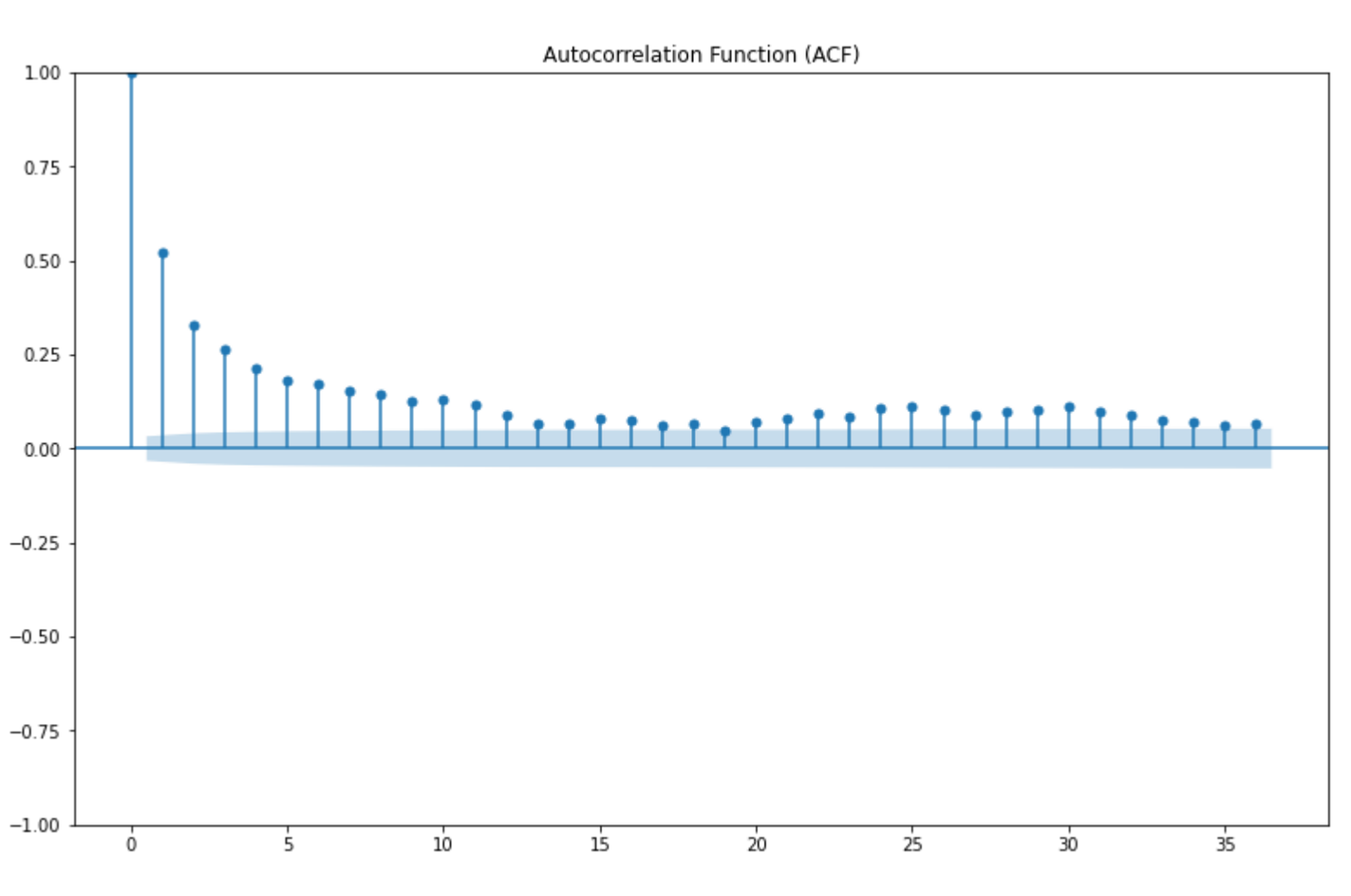


Dobrałam paramytry do równania $y = a\sin(b * x)$ za pomocą funkcji optimize.curve_fit() z biblioteki *scipy*. Linia sezonowości ma równanie $y = -4.54\sin(0.0176x)$. Wykres po usunięciu sezonowości:

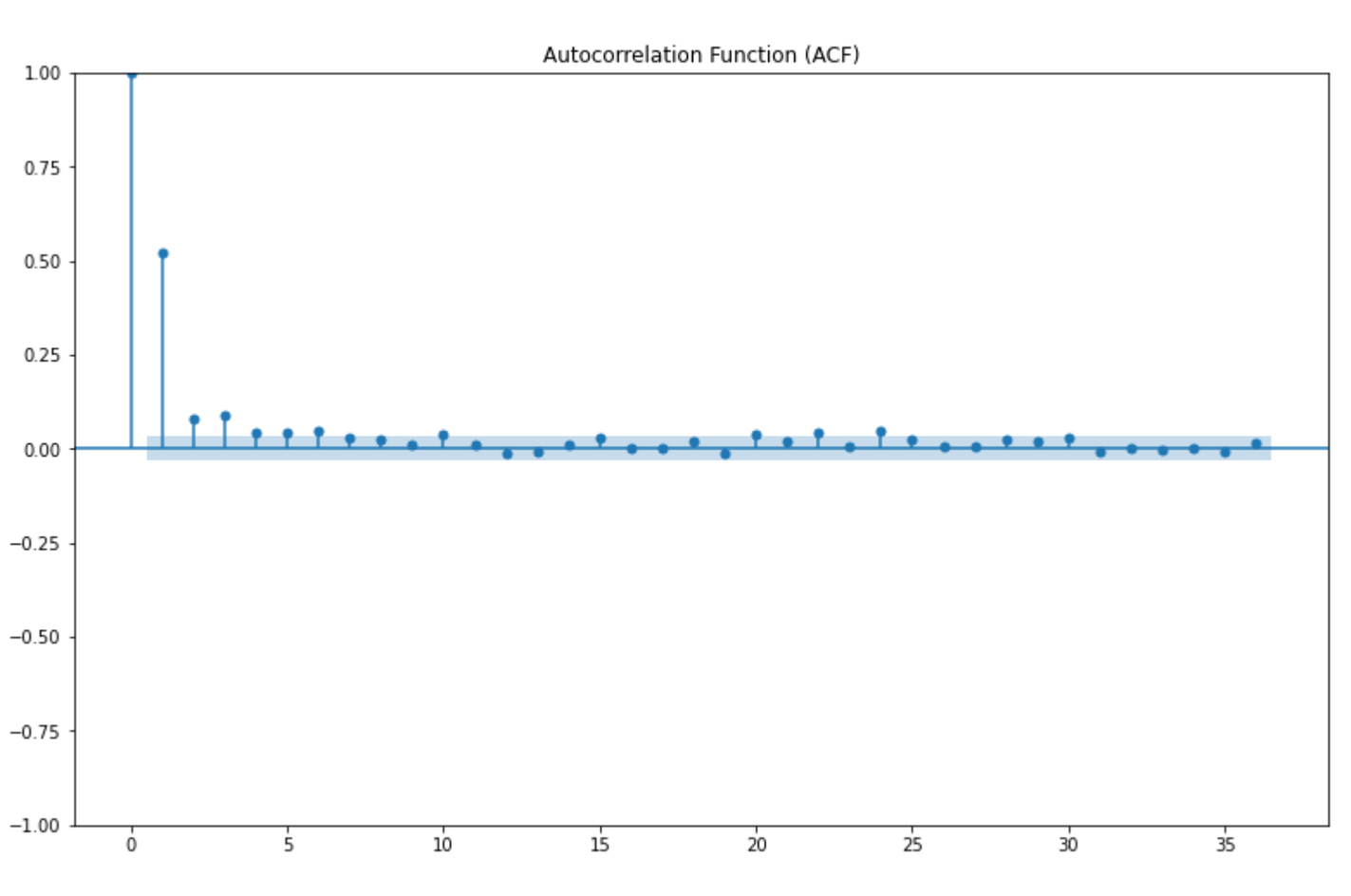


Wykres ACF oraz PACF dla uzyskanego szeregu:

Skorzystałam jeszcze z transformacji stabilizującej warjancje, żeby o ile to możliwe zmniejszyć stacjonarność. To jest wykres ACF uzyskanego szeregu, widzimy że on geometrycznie zanika.



Wykres PACF uzyskanego szeregu:

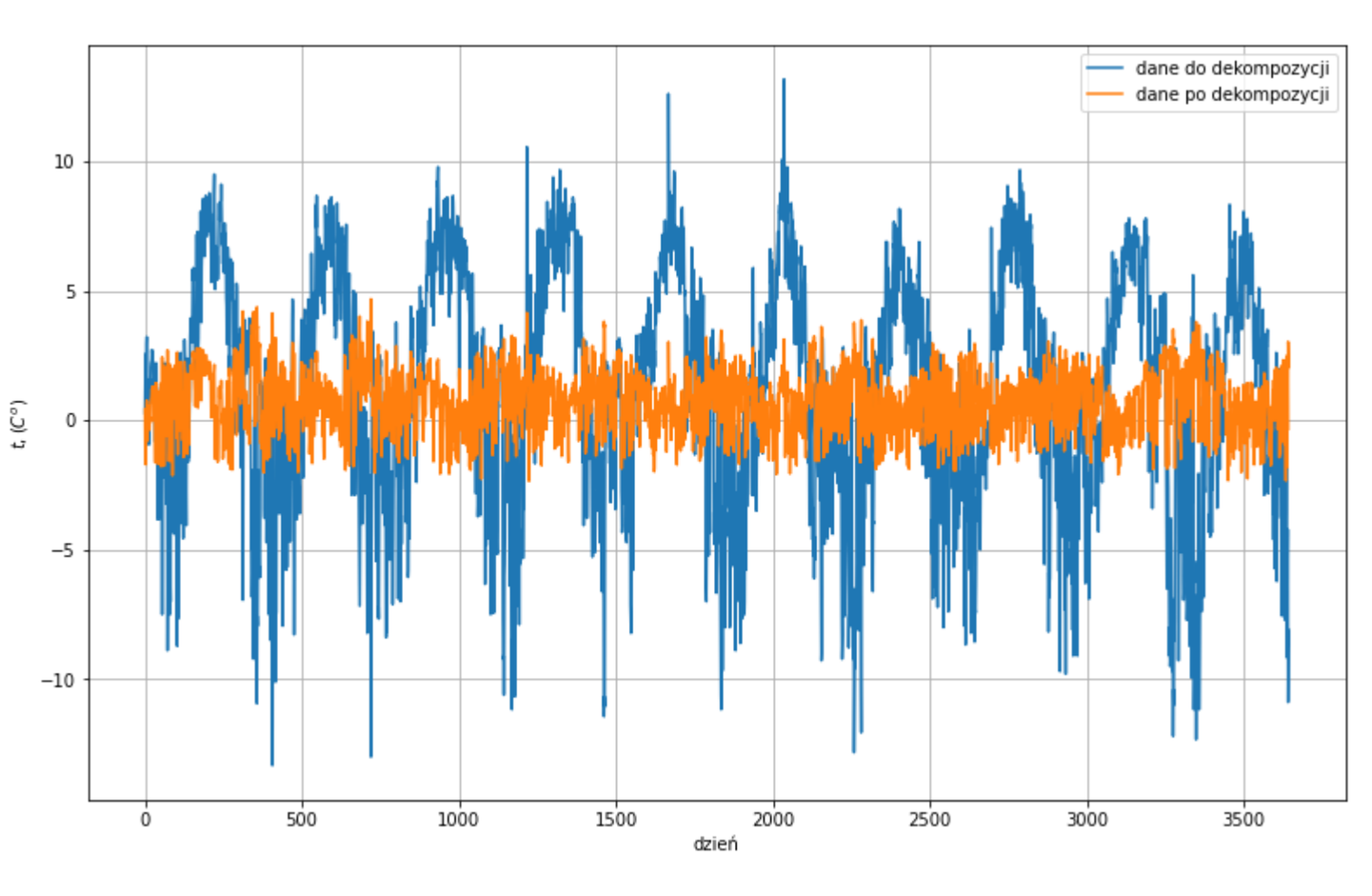


Też zanika co znaczy możemy skorzystać z modeli ARMA

Augmented Dickey-Fuller Test:

Statystyka testowa teraz wynosi $-8,614382$, jest znaczniej mniejsza niż wartości krytyczne, jakie wynoszą $-3.432, -2.862, -2.567$ dla $\alpha = \{0.01, 0.05, 0.1\}$, sugeruje to, że szereg czasowy jest stacjonarny. P-wartość wynosi 0.00

Jeszcze raz zobaczymy jak wyglądają teraz dane:



3. Modelowanie danych przy pomocy ARMA

Dobranie rzędu modelu na podstawie kryterium AIC

Żeby dobrać rzęd modeli skorzystałam z metody siatki (grid search) na podstawie kryterium AIC (Akaike Information Criterion).
Algorytm z którego korzystałam:

1. Zdefiniowanie zakresu parametrów.
2. Przeszukiwanie przestrzeni parametrów.
3. Obliczenie AIC dla każdego modelu.
4. Wyszukiwanie modeli z najniższym AIC.

Otrzymałam że $p = 2$ oraz $q = 3$, kryterium AIC wyniosł 10143.966652729214

Estymacja parametrów modelu

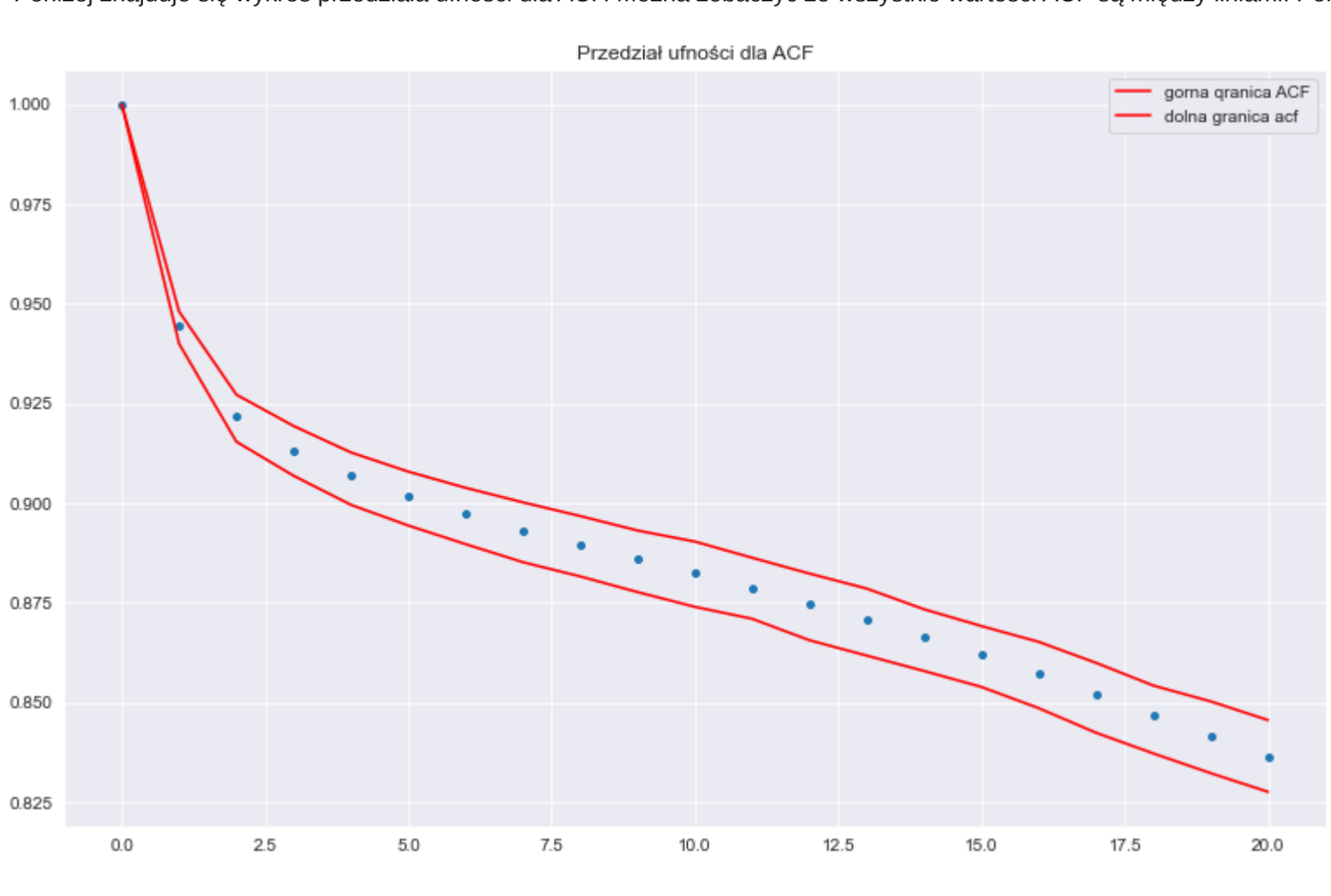
Dla estymacji parametrów skorzystałam z biblioteki statsmodels.tsa.arima.model w Pythonie. Otrzymałam szereg czasowy z parametrami: $c = 0.7854$, $\phi = \{1.6677, -0.6746\}$, $\theta = \{-1.2099, 0.1482, 0.0932\}$, $\sigma = 0.9442$

```
                               SARIMAX Results
==============================================================================
Dep. Variable:                      y   No. Observations:                 3643
Model:                 ARIMA(2, 0, 3)   Log Likelihood               -5064.983
Date:                Wed, 07 Feb 2024   AIC                          10143.967
Time:                        13:29:10   BIC                          10187.371
Sample:                             0   HQIC                         10159.425
                               - 3643
Covariance Type:                  opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
const          0.7854      0.077     10.192      0.000       0.634       0.936
ar.L1          1.6677      0.072     23.302      0.000       1.527       1.808
ar.L2         -0.6746      0.068     -9.898      0.000      -0.808      -0.541
ma.L1         -1.2099      0.074    -16.261      0.000      -1.356      -1.064
ma.L2          0.1482      0.041      3.646      0.000       0.069       0.228
ma.L3          0.0932      0.030      3.063      0.002       0.034       0.153
sigma2         0.9442      0.022     42.537      0.000       0.901       0.988
===================================================================================
Ljung-Box (L1) (Q):                   0.00   Jarque-Bera (JB):               120.34
Prob(Q):                              0.99   Prob(JB):                         0.00
Heteroskedasticity (H):               0.92   Skew:                            -0.41
Prob(H) (two-sided):                  0.17   Kurtosis:                         3.33
===================================================================================
```

4. Ocena dopasowania modelu

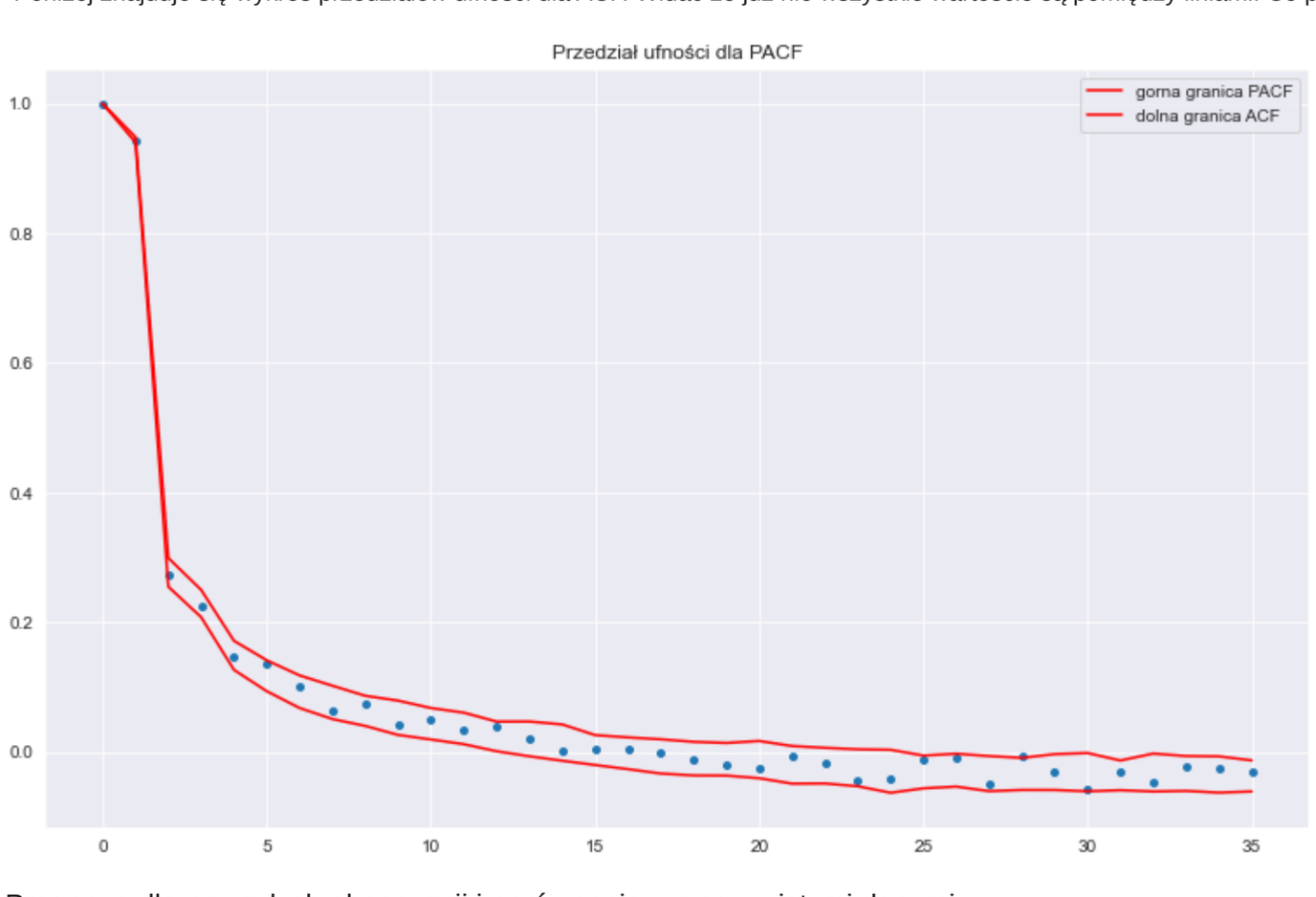
Przedziały ufności dla ACF

Poniżej znajduje się wykres przedziała ufności dla ACF. Można zobaczyć że wszystkie wartości ACF są między liniami. Policzyłam jeszcze procent wartości ACF poza przedziałem ufności dla każdego h, średnie on wyniósł 0.095.



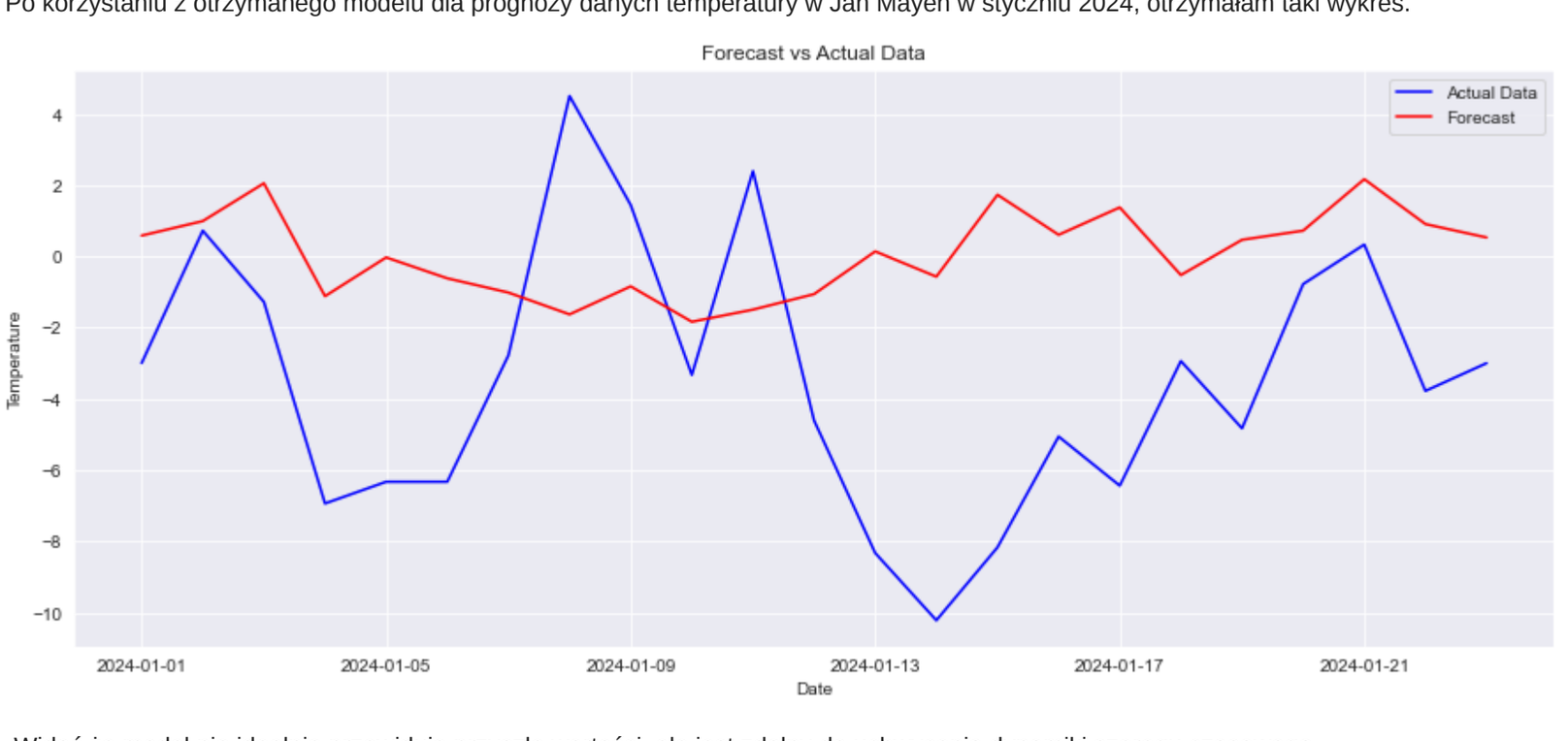
Przedziały ufności dla PACF

Poniżej znajduje się wykres przedziałow ufności dla ACF. Widać że już nie wszystkie wartoście są pomiędzy liniami. Co potwierdza ten fakt, że procent wartości PACF poza przedziałem ufności jest dużej niż dla ACF, wyniósł on 0.245.



Prognoza dla przyszłych obserwacji i porównanie z rzeczywistymi danymi

Po korzystaniu z otrzymanego modelu dla prognozy danych temperatury w Jan Mayen w styczniu 2024, otrzymałam taki wykres.

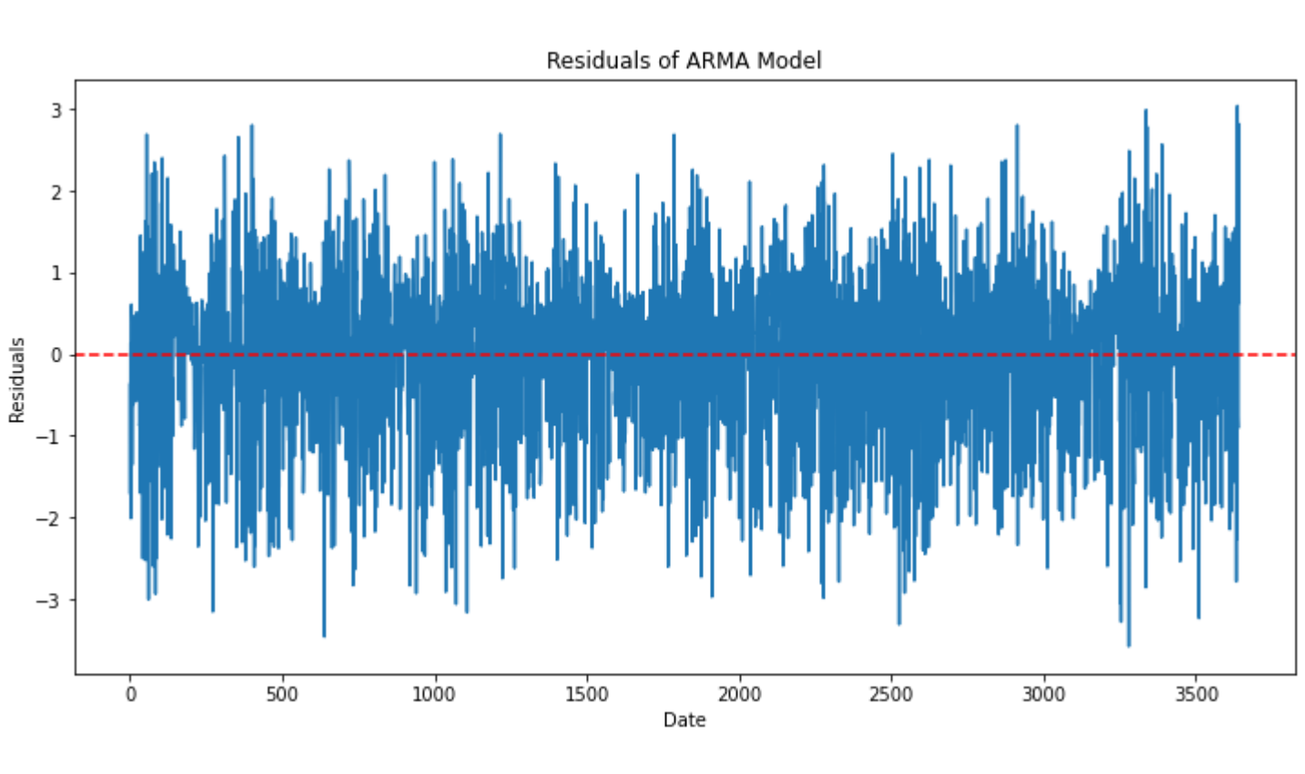


Widać że model nie idealnie przewiduje przyszłe wartości, ale jest zdolny do uchwycenia dynamiki szeregu czasowego.

5. Weryfikacja założeń dotyczących szumu

Założenie dot. średniej

Poniżej jest przedstawiony wykres wartości resztowych. Wykres nie pokazuje oczywistych wzorców czy trendów, co jest dobrym znakiem i sugeruje, że model dobrze dopasowuje się do danych. Istnieje jednak pojedyncze, znaczne wypiętrzenie, które może wskazywać na wartość odstającą lub inny problem w danych, który nie został uwzględniony w modelu.



Wyniki testu t:

t-statistic: -0.007734570314583497 p-value: 0.9938292379486273

Wartość statystyki testowej jest bardzo bliska zeru, co sugeruje, że średnia reszt jest bliska zeru. Wartość p jest znacznie wyższa niż standardowy poziom istotności (zazwyczaj 0.05), co oznacza, że nie ma podstaw do odrzucenia hipotezy zerowej mówiącej o braku różnicy między średnią reszt a zerem.

Założenie dot. wariancji

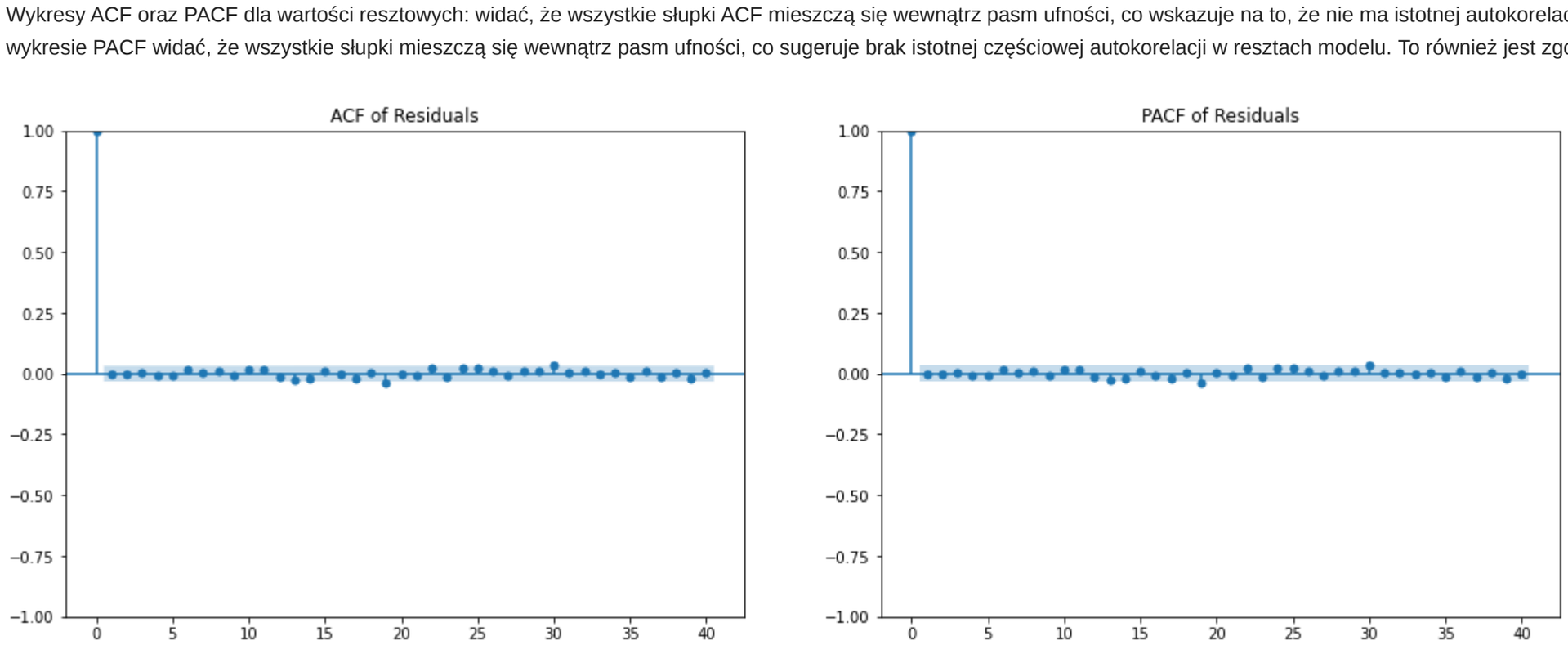
Wyniki testu Levene'a:

Levene Test Statistic: 7.1463475333135715 p-value: 0.007548926471167769

Statystyka testu Levene'a jest stosunkowo wysoka. Wartość p jest poniżej 0.10, co może wskazywać na to, że wariancje są nierówne (heteroskedastyczność), jednak wartość p jest wyższa niż standardowy próg 0.05, więc wynik ten może być na granicy istotności statystycznej.

Założenie dot. niezależności

Wykresy ACF oraz PACF dla wartości resztowych: widać, że wszystkie słupki ACF mieszczą się wewnątrz pasm ufności, co wskazuje na to, że nie ma istotnej autokorelacji na żadnym z opóźnień. Jest to dobra oznaka i wskazuje na to, że reszty są bliskie niezależności. Na wykresie PACF widać, że wszystkie słupki mieszczą się wewnątrz pasm ufności, co sugeruje brak istotnej częściowej autokorelacji w resztach modelu. To również jest zgodne z oczekiwaniem, że reszty są niezależne.



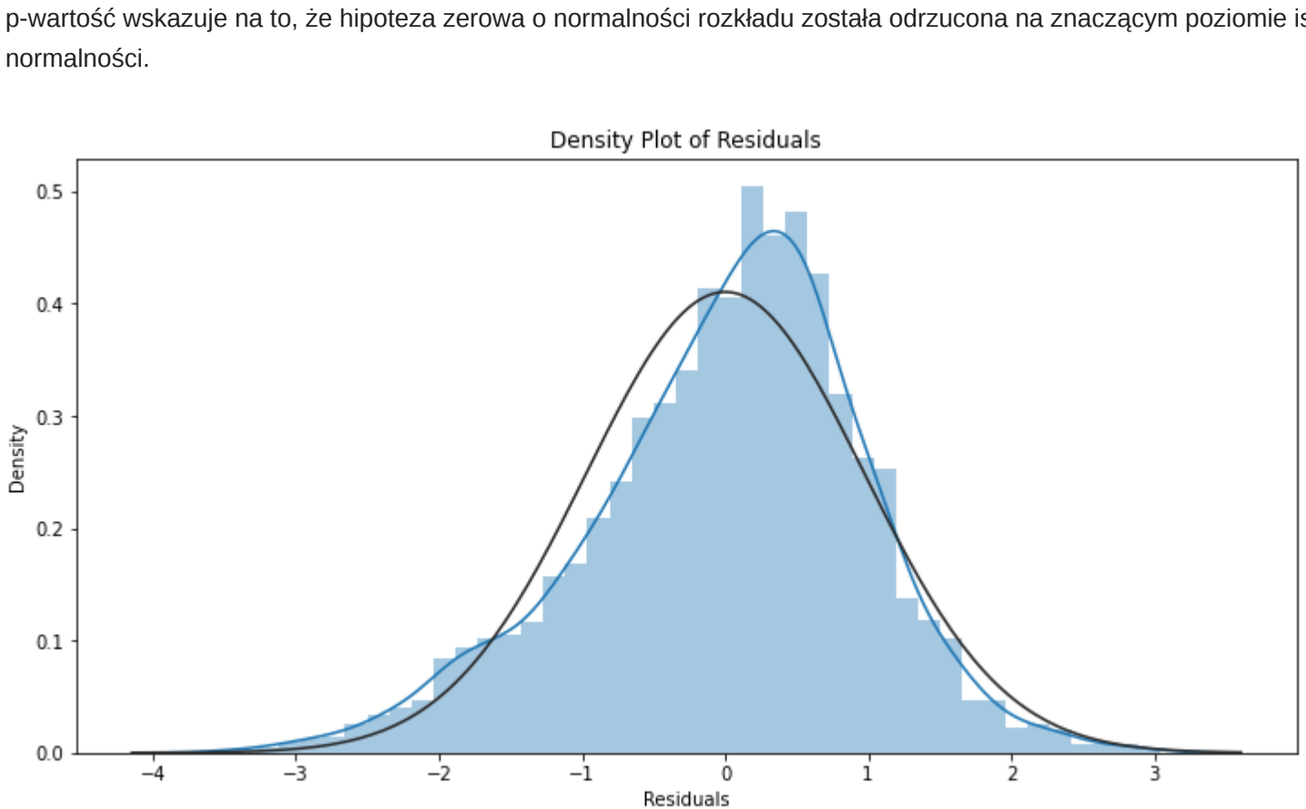
Założenie dotyczące normalności rozkładu

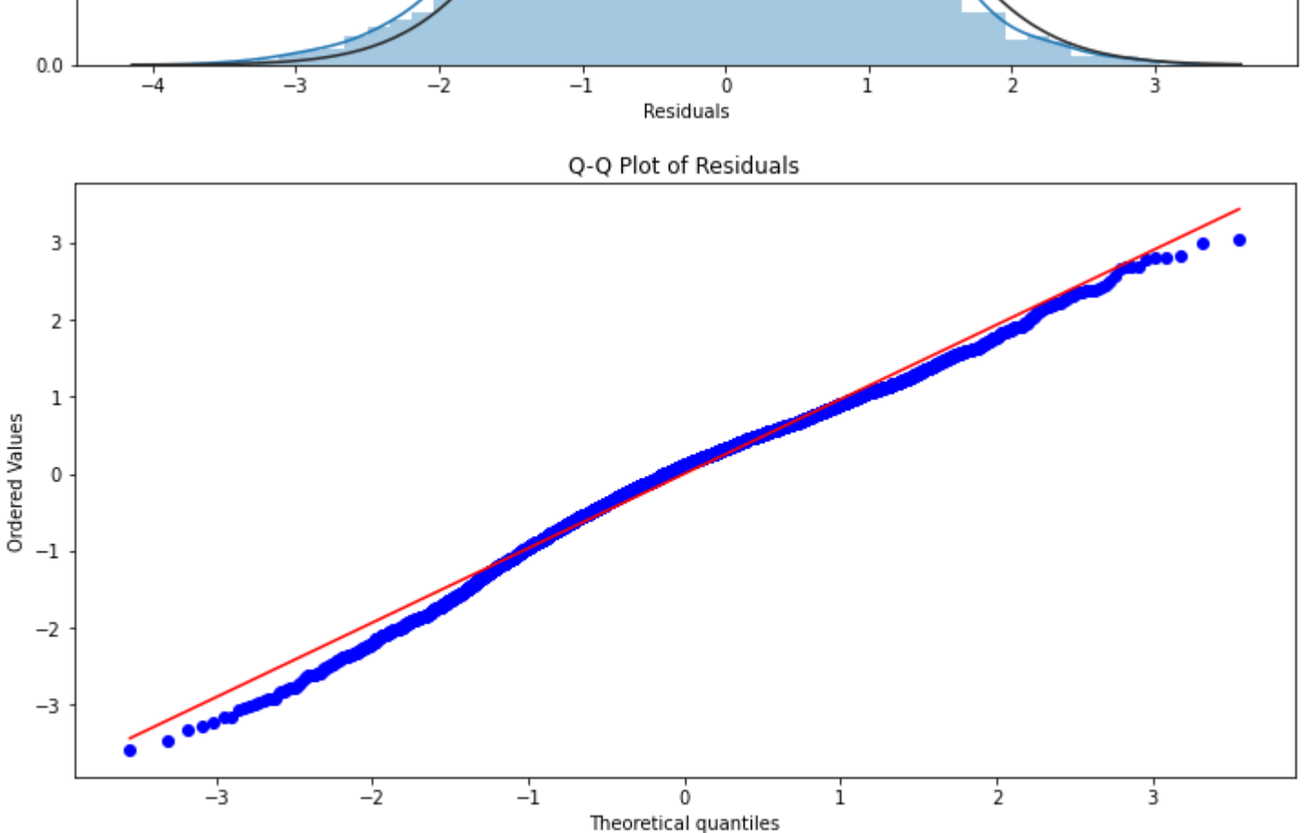
Wyniki testu Shapiro-Wilka:

Shapiro-Wilka Test: Statistics=0.984, p=0.000

Wartość statystyki testu Shapiro-Wilka, jest bliska 1, sugerując, że dane mogą mieć rozkład bliski normalnemu.

p-wartość wskazuje na to, że hipoteza zerowa o normalności rozkładu została odrzucona na znaczącym poziomie istotności. Prawdopodobnie ze względu na dużą liczbę próbek nawet niewielkie odchylenia od normalności mogą prowadzić do odrzucenia hipotezy o normalności.





Na podstawie wykresu gęstości, reszty wydają się mieć kształt zbliżony do normalnego, ale ze względu na wyniki testów statystycznych, mogą występować pewne odchylenia od normalności.

Na podstawie wykresu kwantylowego, większość punktów leży blisko linii, ale widoczne są odstępstwa na końcach rozkładu, co wskazuje na obecność "ciężkich ogonów" - więcej ekstremalnych wartości niż oczekuje się w przypadku normalnego rozkładu.

6. Wnioski

Podsumowując przeprowadzoną analizę, model ARMA wydaje się być skutecznym narzędziem do modelowania i analizy wybranego szeregu czasowego. Na podstawie uzyskanych wyników, wykresów oraz przeprowadzonych testów statystycznych, możemy sformułować następujące wnioski:

1. Udało się z powodzeniem zastosować model ARMA do analizy szeregów czasowych, co wskazuje na adekwatność tego modelu do charakterystyki danych. Dobra zdolność przewidywania przyszłych wartości potwierdza, że model dobrze uchwycił zależności w danych historycznych.
2. Proces dopasowania modelu, w tym wybór odpowiednich rzędów autoregresji (AR) i średniej ruchomej (MA), został przeprowadzony efektywnie. Odpowiednia konfiguracja parametrów jest kluczowa dla osiągnięcia wiarygodnych prognoz i wskazuje na dokładną analizę właściwości danych oraz kompetencje w zakresie modelowania statystycznego.
3. Dobre wyniki prognozowania przez model ARMA podkreślają jego przydatność w planowaniu i podejmowaniu decyzji.
4. Dopasowanie Modelu: Model ARMA został z powodzeniem dopasowany do danych, co wskazują analizy reszt. Brak wyraźnych wzorców na wykresach ACF i PACF reszt sugeruje, że model skutecznie wyłapał informacje z danych szeregu czasowego. Reszty wydają się być niezależne, co jest kluczowym założeniem w efektywnym modelowaniu szeregów czasowych.
5. Analiza Reszt: Analiza reszt wskazała na brak istotnych autokorelacji, zarówno na podstawie ACF, jak i PACF, co sugeruje dobrą specyfikację modelu. Jednakże, wykryto potencjalne odchylenia od normalności rozkładu, co może wskazywać na konieczność dalszej optymalizacji modelu lub zastosowania alternatywnych podejść do modelowania reszt, które lepiej radzą sobie z ewentualnymi nieregularnościami.
6. Testy na Normalność: Wyniki testów Shapiro-Wilka wskazują na odrzucenie hipotezy o normalności rozkładu reszt. To może sugerować, że warto rozważyć transformacje danych lub zastosowanie bardziej zaawansowanych modeli, takich jak ARMA z niestandardowymi rozkładami reszt.

Dobrze dopasowany model ARMA może stanowić podstawę do dalszych badań i eksploracji innych modeli szeregów czasowych, takich jak ARIMA, SARIMA czy nawet bardziej zaawansowane metody oparte na uczeniu maszynowym. Może to prowadzić do jeszcze lepszego zrozumienia danych i dalszego zwiększenia dokładności prognoz.

Podsumowując, skuteczne dopasowanie modelu ARMA i generowanie przez niego dobrych prognoz to znaczący sukces, który otwiera możliwości wykorzystania tych technik w praktycznych zastosowaniach oraz zachęca do dalszego pogłębiania wiedzy i umiejętności w dziedzinie modelowania szeregów czasowych.