Cel analizy

Celem tej analizy jest prognozowanie średnich miesięcznych temperatur w nowojorskim Central Parku na rok 2024 na podstawie danych historycznych z okresu 2000-2023. Badanie rozpocznie się wstępną transformacją danych, aby nadawały się do dalszej analizy oraz celem wizualizacji. Następnie wybrane zostaną metody odpowiednie do tematu i istoty analizy. Kolejnym krokiem jest przeprowadzenie analizy, jej interpretacja, ocena jakości oraz porównanie z drugą metodą. Etapem końcowym jest wyprowadzenie prognoz oraz zebranie wniosków z całego badania.

```r
#Wykorzystane biblioteki
library(tidyverse)
library(lubridate)
library(forecast)
library(tseries)
```

Źródło danych, opis i charakterystyka

Zbiór danych zawiera średnie miesięczne temperatury dla nowojorskiego Central Parku. Dane pochodzą z okresu 2000-2023. Zbiór danych zawiera dwie kolumny:

- data (w formacie RRRRMM),
- wartość (średnia temperatura w stopniach Fahrenheita).

Dane wykorzystane do tej analizy pochodzą z National Centers for Environmental Information (NCEI), amerykańskiej agencji rządowej, która zarządza jednym z największych na świecie archiwów danych m.in. atmosferycznych.

```r
dane <- read.csv("data.csv", skip = 3)
colnames(dane) <- c("Data", "Temperatura")
head(dane)
```

```
##     Data Temperatura
## 1 200001        31.5
## 2 200002        37.5
## 3 200003        47.4
## 4 200004        51.2
## 5 200005        63.8
## 6 200006        71.5
```

Przed rozpoczęciem analizy dane należy przystosować do analizy. Dokonano zmiany typu kolumny Data na *date*. Następnie przekonwertowano wartości temperatury ze stopni Fahrenheita na stopnie Celsjusza. Na koniec wartości zostały zaokrąglone.

```
dane <- dane %>%
  mutate(Data = as.Date(paste0(Data, "01"), format="%Y%m%d"))
dane[, 2] <- (dane[, 2] - 32) * 5 / 9
dane[, 2] <- ceiling((dane[, 2] * 10) / 10)
head(dane)
```

```
##         Data Temperatura
## 1 2000-01-01           0
## 2 2000-02-01           4
## 3 2000-03-01           9
## 4 2000-04-01          11
## 5 2000-05-01          18
## 6 2000-06-01          22
```

```
summary(dane$Temperatura)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   -4.00    6.00   14.00   13.85   22.00   28.00
```

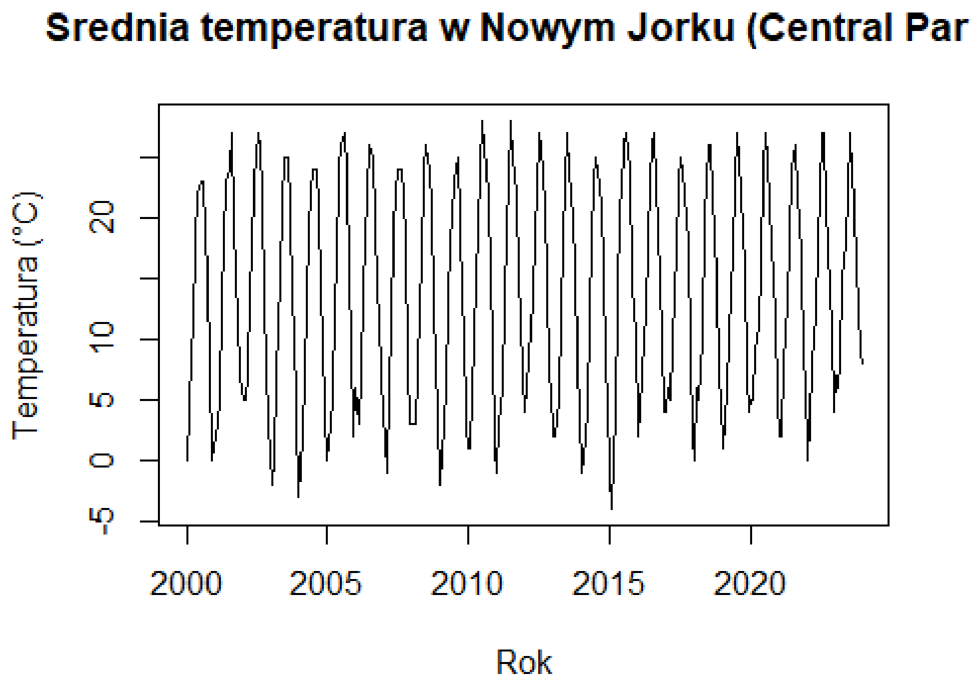
Minimalna średnia miesięczna temperatura odnotowana w tym okresie to -4°C, a maksymalna to 28°C. Wartość przeciętna średniej miesięcznej temperatury w badanym okresie to ok. 14°C. Warto dodać, że klimat panujący w Nowym Jorku jest określany jako umiarkowany kontynentalny. Charakteryzuje się on czterema wyraźnymi porami roku o temperaturach podobnych do tych występujących w Polsce. Jak widać, statystyki opisowe odzwierciedlają warunki klimatyczne tego miasta.

Kolejną transformacją, jaką trzeba przeprowadzić, jest zamiana danych na szereg czasowy o frekwencji miesięcznej, a więc równej 12. Tuż po tym można zwizualizować zbiór danych.

```
ts_dane <- ts(dane$Temperatura, start = c(year(min(dane$Data)), month(min(dan
e$Data))), frequency = 12)
```

Prezentacja graficzna zbioru danych

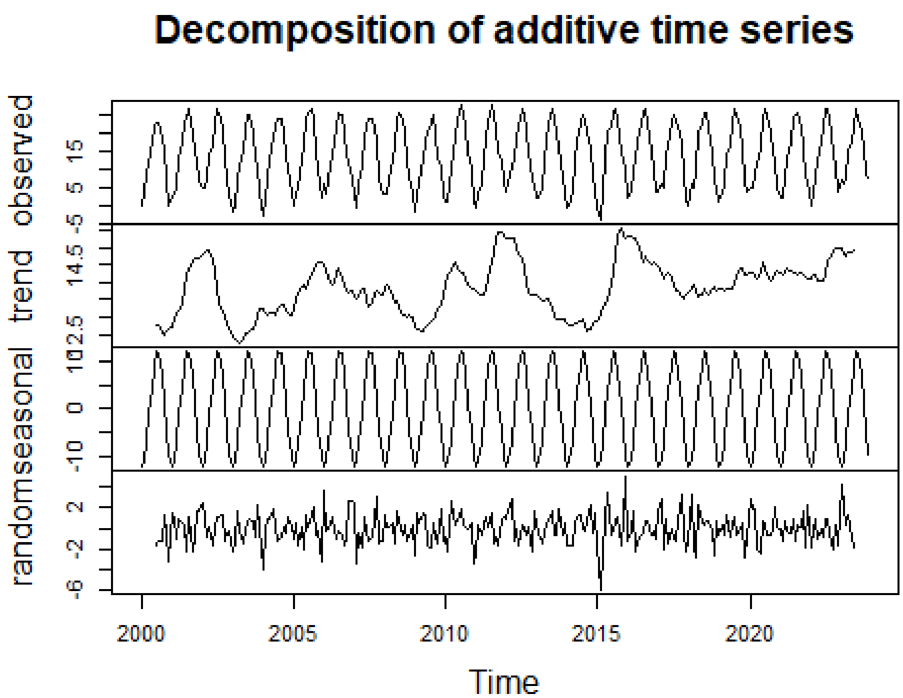
```
plot(ts_dane, main = "Srednia temperatura w Nowym Jorku (Central Park)", ylab
= "Temperatura (°C)", xlab = "Rok")
```



Wykres obrazuje, jakie wartości osiągały średnio w poszczególnych miesiącach temperatury w Nowym Jorku. Można wyraźnie zaobserwować sezonowość związaną z porami roku i jej cykliczność. Nie widać za to wyraźnego trendu; dopiero przy uważniejszym spojrzeniu uda się wychwycić nieregularny trend wzrostowy w badanym okresie.

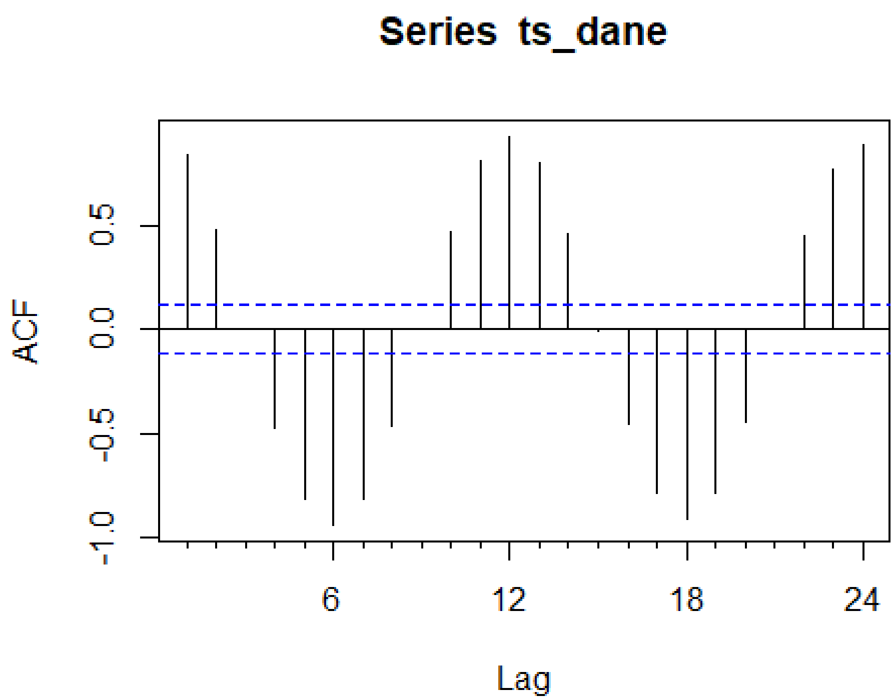
Dekompozycja badanego szeregu czasowego pozwoli na przejrzystszy wgląd w dane.

```
decomposed <- decompose(ts_dane)
plot(decomposed)
```



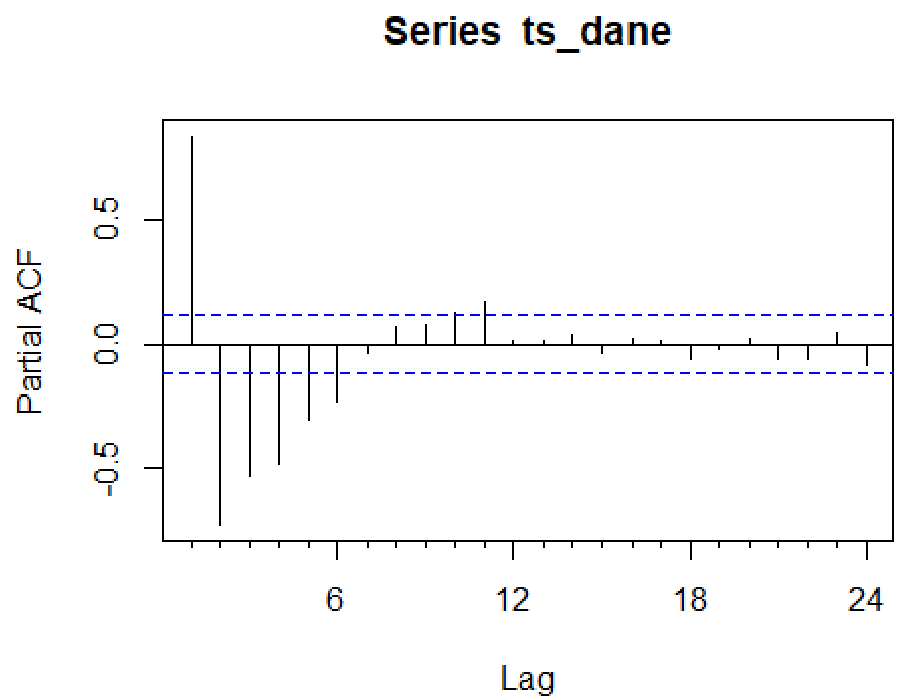
Dekompozycja rozkładu pokazuje, że średnie miesięczne temperatury mają wyraźny wzorzec sezonowy z roczna okresowością, długoterminowym trendem w tym okresie i przypadkowymi wahaniami. Sezonowość jest silna i wskazuje na znaczną regularną sezonową zmienność temperatur, podczas gdy trend na przestrzeni lat ogólnie jest dodatni, chociaż nie jest regularny i występują dość duże wahania. Reszty sugerują, że podczas gdy większość zmienności jest wychwytywana przez trend i składniki sezonowe, w danych nadal występuje pewien losowy szum.

```
Acf(ts_dane)
```



Funkcja autokorelacji ujawnia spodziewany wzorzec. Obecność znaczących dodatnich autokorelacji przy opóźnieniach wynoszących 12 miesięcy sugeruje roczną sezonowość, która jest powszechna w danych dotyczących temperatury ze względu na roczny cykl pór roku. Z kolei znacząca autokorelacja przy opóźnieniu 6 sugeruje półroczny wzorzec z powodu zmiany pory roku z zimy na lato.

```r
Pacf(ts_dane)
```



Wykres PACF dla średnich miesięcznych temperatur wskazuje na istotne autokorelacje od opóźnienia 1 do opóźnienia 6. Sugeruje to, że temperatury z miesiąca na miesiąc są ze sobą ścisłe powiązane.

Stacjonarność badanego szeregu czasowego sprawdzono odpowiednim do tego testem Dickeya-Fullera w wersji rozszerzonej.

```r
adf.test(ts_dane)

## Warning in adf.test(ts_dane): p-value smaller than printed p-value

##
##  Augmented Dickey-Fuller Test
##
## data:  ts_dane
## Dickey-Fuller = -12.789, Lag order = 6, p-value = 0.01
## alternative hypothesis: stationary
```

Ponieważ wartość p wynosi mniej niż 0,01, czyli mniej niż poziom istotności, odrzucono hipotezę zerową. Oznacza to, że istnieje podstawa do stwierdzenia, że ten szereg czasowy jest stacjonarny i można przejść do tworzenia modelu SARIMA.

Dobór metody analizy odpowiadającej celowi badania

Zadecydowano, że do odpowiednią do tej analizy metodą będzie SARIMA. Model SARIMA jest bardzo skuteczny w analizie i prognozowaniu średnich miesięcznych temperatur ze względu na jego zdolność do radzenia sobie z sezonowością, trendami i autokorelacją w danych.

```
sarima <- auto.arima(ts_dane, ic="bic", seasonal = TRUE)
print(sarima)
```

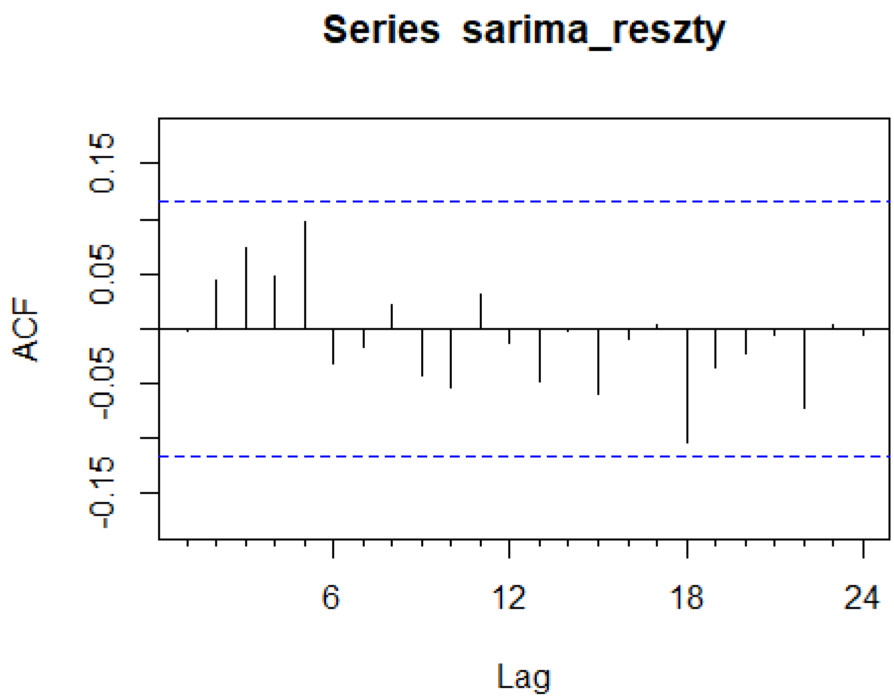
```
## Series: ts_dane
## ARIMA(0,0,1)(2,1,2)[12]
##
## Coefficients:
##          ma1    sar1     sar2     sma1    sma2
##       0.2114  0.4684  -0.1416  -1.4455  0.5349
## s.e.  0.0583  0.2347   0.0779   0.2335  0.2124
##
## sigma^2 = 3.132:  log likelihood = -557.73
## AIC=1127.47   AICc=1127.78   BIC=1149.19
```

Wybrany automatycznie model potwierdza wpływ temperatury z danego miesiąca na temperaturę tego samego miesiąca w następnych latach. Wskaźniki wydajności modelu sugerują, że równoważy on dopasowanie i złożoność.

Ocena jakości analizy

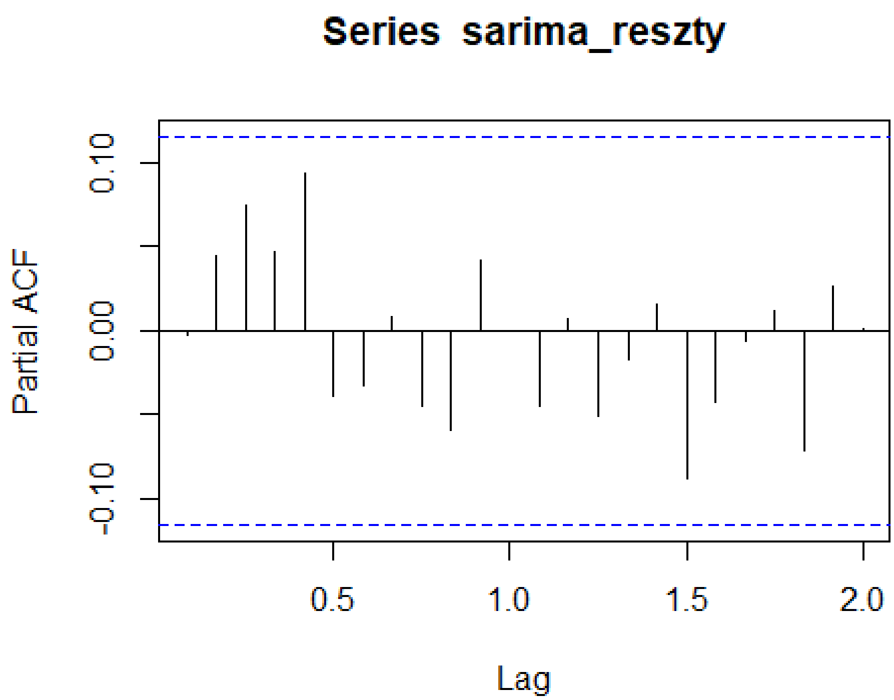
W celu oceny jakości modelu należy dokonać analizy reszt. Ponownie wykorzystane zostaną funkcje autokorelacji i autokorelacji częściowej.

```
sarima_reszty <- sarima$resid
Acf(sarima_reszty)
```



Na podstawie tego wykresu ACF można stwierdzić, że model SARIMA wydaje się dość dobrze pasować do danych, ponieważ reszty nie wykazują silnej autokorelacji. Na potwierdzenie tej tezy utworzono również wykres częściowej autokorelacji.

```r
pacf(sarima_reszty)
```



Wygląda na to, że model SARIMA dobrze poradził sobie z uchwyceniem zależności w danych, ponieważ większość opóźnień mieści się w przedziałach ufności. Wykres PACF sugeruje, że model SARIMA jest poprawnie dopasowany do danych.

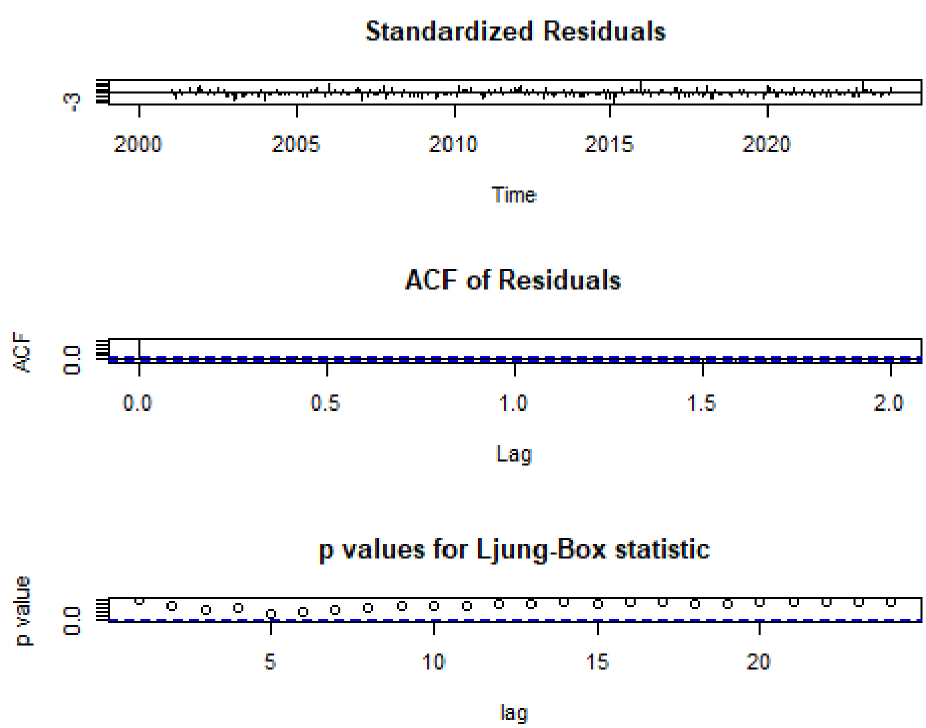
```r
shapiro.test(sarima_reszty)

## 
##  Shapiro-Wilk normality test
## 
## data:  sarima_reszty
## W = 0.98992, p-value = 0.04397
```

Przy wartości p równej 0,04397 odrzucamy hipotezę zerowa. W związku z tym reszty nie mają rozkładu normalnego, co świadczy o związanej z tym niedokładnością modelu.

```
Box.test(sarima_reszty, lag=1, type="Ljung-Box")

##
##  Box-Ljung test
##
## data:  sarima_reszty
## X-squared = 0.0022903, df = 1, p-value = 0.9618

Box.test(sarima_reszty, lag=12, type="Ljung-Box")

##
##  Box-Ljung test
##
## data:  sarima_reszty
## X-squared = 7.9675, df = 12, p-value = 0.7877

Box.test(sarima_reszty, lag=24, type="Ljung-Box")

##
##  Box-Ljung test
##
## data:  sarima_reszty
## X-squared = 15.339, df = 24, p-value = 0.9105
```

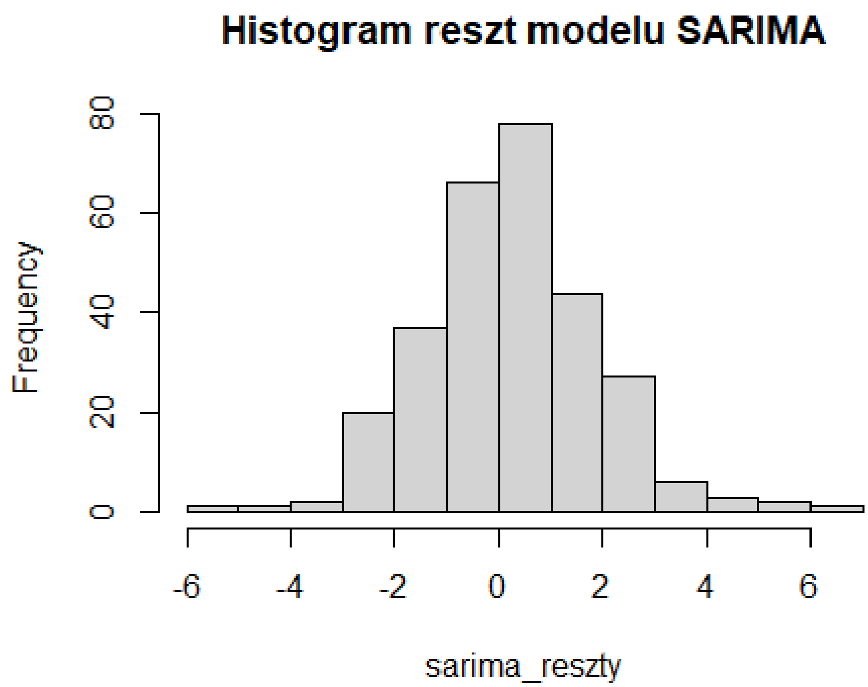
Wyniki testów Boxa-Ljunga stwierdzają brak znaczącej autokorelacji w resztach modelu SARIMA. Model odpowiednio odzwierciedla czasowe zależności w danych.

```
tsdiag(sarima, gof.lag=24)
```



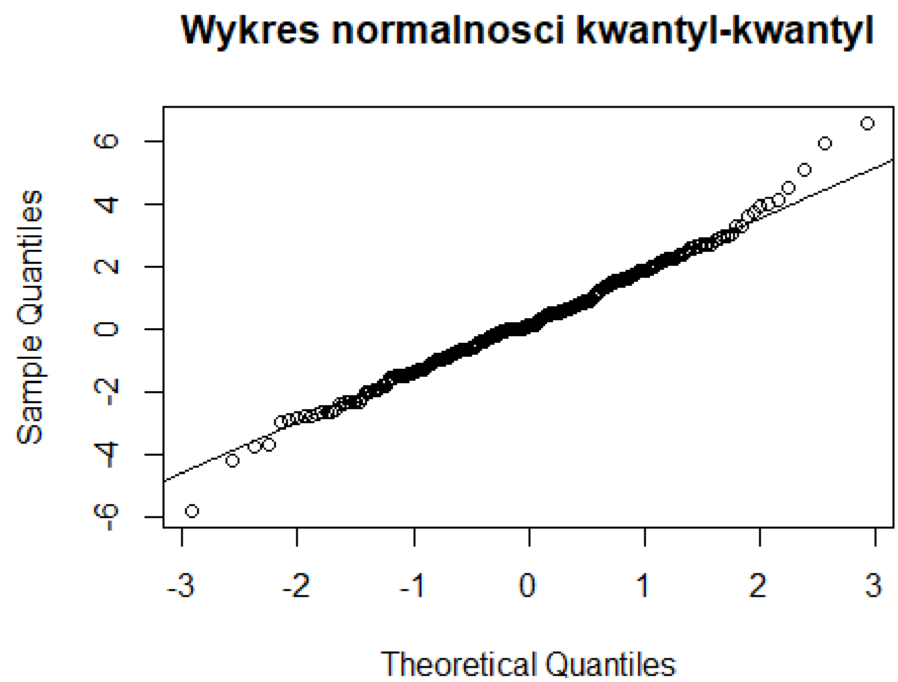
Kolejne spojrzenie na wykresy diagnostyczne dla modelu SARIMA potwierdza, że reszty nie są autoskorelowane i rzeczywiście są losowe. Świadczy to o dobrym dopasowaniu modelu.

```
hist(sarima_reszty, main="Histogram reszt modelu SARIMA")
```



Histogram reszt modelu SARIMA wyglądem przypomina rozkład normalny, ale, jak wiadomo po przeprowadzeniu testu Shapiro-Wilka, jest to jedynie rozkład zbliżony do rozkładu normalnego.

```r
qqnorm(sarima_reszty, main="Wykres normalnosci kwantyl-kwantyl")
qqline(sarima_reszty)
```



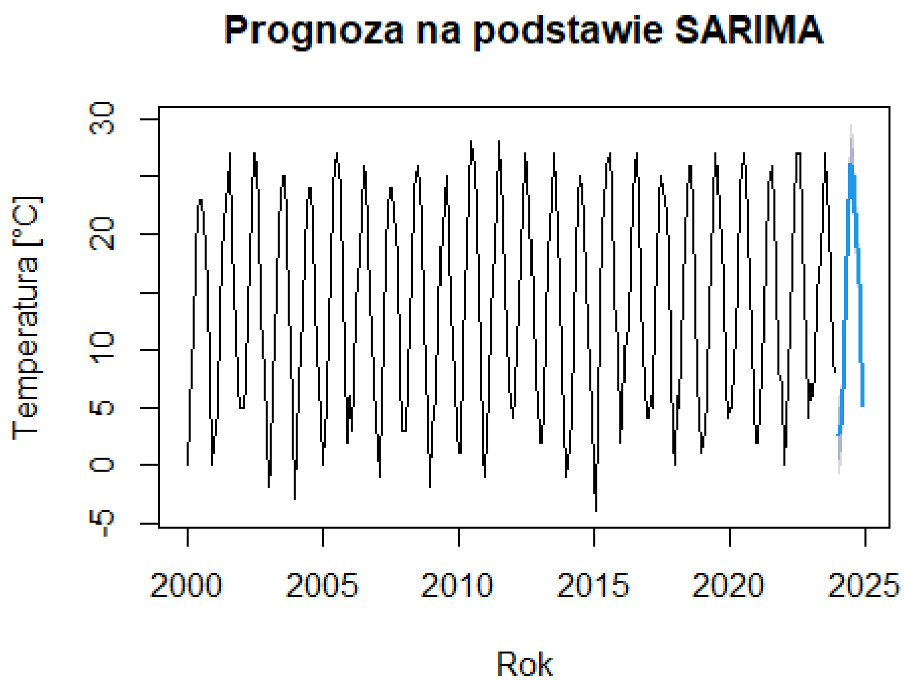
Wykres normalności kwantyl-kwantyl potwierdza, że rozkład teoretyczny jest bardzo zbliżony do rozkładu normalnego.

Po przeprowadzeniu analizy można przejść do etapu prognozowania temperatury. Prognozę przeprowadzono na kolejnych 12 miesięcy, a wiec na cały rok 2024.

```r
prognoza_sarima <- forecast(sarima, h = 12)
prognoza_sarima
```

```
##          Point Forecast      Lo 80     Hi 80      Lo 95     Hi 95
## Jan 2024       2.584960  0.3169168  4.853003 -0.8837126  6.053632
## Feb 2024       3.324671  1.0065199  5.642822 -0.2206351  6.869977
## Mar 2024       6.552047  4.2338963  8.870198  3.0067413 10.097353
## Apr 2024      12.506505 10.1883545 14.824656  8.9611995 16.051811
## May 2024      18.091455 15.7733040 20.409606 14.5461490 21.636761
## Jun 2024      23.064244 20.7460929 25.382395 19.5189379 26.609550
## Jul 2024      26.115085 23.7969337 28.433236 22.5697787 29.660391
## Aug 2024      25.281363 22.9632124 27.599514 21.7360575 28.826669
## Sep 2024      21.921424 19.6032730 24.239575 18.3761180 25.466730
## Oct 2024      15.792831 13.4746803 18.110982 12.2475253 19.338137
## Nov 2024       8.957083  6.6389323 11.275234  5.4117774 12.502389
## Dec 2024       5.061136  2.7429850  7.379287  1.5158300  8.606442
```

```r
plot(prognoza_sarima, main = "Prognoza na podstawie SARIMA", ylab = "Temperat
ura [°C]", xlab = "Rok")
```



Prognoza utworzona przy pomocy modelu SARIMA przewiduje temperatury na podobnym poziomie, co w poprzednich latach. Dla wysokich poziomów ufności prognoza bierze pod uwagę scenariusz, w którym lato 2024 jest najcieplejsze w badanym okresie, jednakże zima 2024 nawet w najzimniejszym scenariuszu nie bije rekordu zimna lat 2000-2024.

Interpretacja wyników

Aby ocenić skuteczność prognozy, wybrano dwa kryteria – MAE i RMSE. Użyto ich zarówno do oceny prognozy na bazie SARIMY, jak i prognozy na bazie ETS.

```r
kryteria <- c("MAE", "RMSE")
accuracy(prognoza_sarima)[,kryteria]
```

```
##      MAE     RMSE
## 1.305697 1.716736
```

Średnia wartość bezwzględna błędu predykcji wynosi ok. 1,3°C. Z kolei wynik RMSE świadczy o tym, że prognoza średnio myli się o ok. +/- 1,7°C.

Porównanie wyników dla metod SARIMA i ETS

Model ETS również jest odpowiedni do badania średnich miesięcznych temperatur. W związku z tym wybrano tę metodę celem porównania jej wyników z wynikami modelu SARIMA.

```
fit_ets <- ets(ts_dane, opt.crit="mse", ic="bic")
summary(fit_ets)
```

```
## ETS(A,N,A)
##
## Call:
##  ets(y = ts_dane, opt.crit = "mse", ic = "bic")
##
##   Smoothing parameters:
##     alpha = 0.0242
##     gamma = 0.0133
##
##   Initial states:
##     l = 13.4212
##     s = -9.2794 -4.0648 1.2572 7.3669 11.4673 11.9175
##            8.8894 4.1897 -1.1028 -7.0049 -11.0351 -12.601
##
##   sigma:  1.7838
##
##      AIC     AICc      BIC
## 1979.952 1981.716 2034.896
##
## Training set error measures:
##                       ME     RMSE      MAE MPE MAPE      MASE      ACF1
## Training set 0.1245952 1.739943 1.343735 Inf  Inf 0.7064205 0.2011485
```

Wydajność modelu jest stosunkowo dobra. ME na poziomie 0,12 wskazuje na to, że prognozy modelu są delikatnie zaniżone. RMSE na poziomie 1,74 oznacza, że model myli się średnio o +/- 1,74°C. MASE na poziomie 0,71 sugeruje, że model jest lepszy od prostego modelu naiwnego. ACF1 na poziomie 0,2 świadczy o niskiej dodatniej korelacji między błędami w czasie.

```
cat("SARIMA BIC:", BIC(sarima), "\n")
```

```
## SARIMA BIC: 1149.191
```

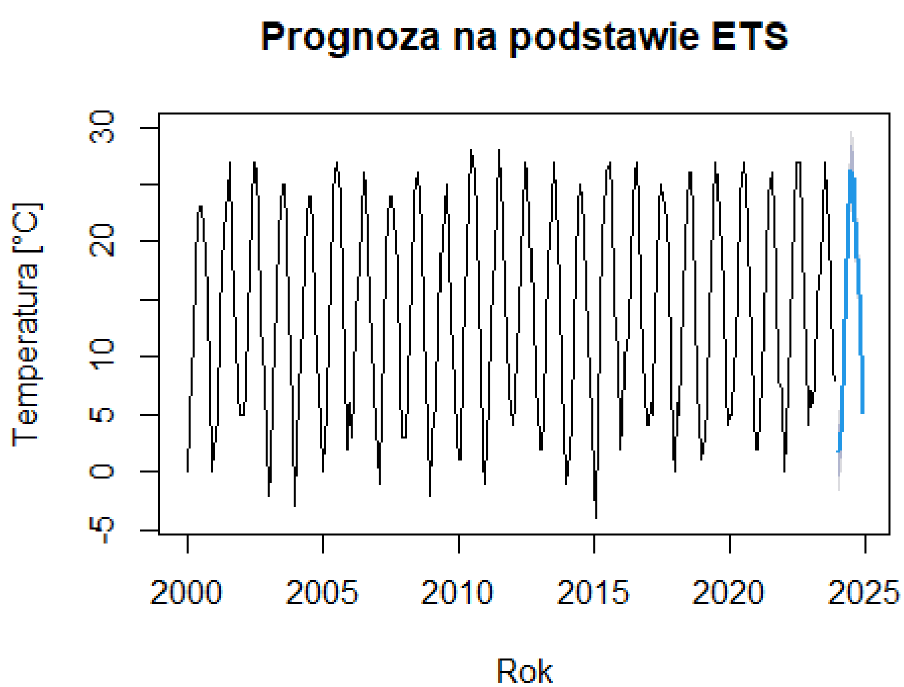
```
cat("ETS BIC:", BIC(fit_ets), "\n")
```

```
## ETS BIC: 2034.896
```

Model SARIMA ma znacznie niższy współczynnik BIC w porównaniu z modelem ETS. Sugeruje to, że model SARIMA zapewnia lepsze dopasowanie do danych w porównaniu z modelem ETS.

Ponownie przeprowadzono prognozę, tym razem dla modelu ETS. Prognozowany okres jest taki sam, jak w przypadku SARIMA.

```
prognoza_ets <- forecast(fit_ets, h=12)
prognoza_ets

##          Point Forecast      Lo 80     Hi 80      Lo 95     Hi 95
## Jan 2024       1.832556 -0.4535279  4.118639 -1.6637074  5.328819
## Feb 2024       3.305516  1.0187652  5.592266 -0.1917675  6.802799
## Mar 2024       7.329781  5.0423633  9.617198  3.8314777 10.828084
## Apr 2024      13.209497 10.9214131 15.497581  9.7101746 16.708820
## May 2024      18.481669 16.1929184 20.770419 14.9813270 21.982011
## Jun 2024      23.195350 20.9059330 25.484766 19.6939890 26.696710
## Jul 2024      26.260399 23.9703163 28.550482 22.7580197 29.762778
## Aug 2024      25.771892 23.4811431 28.062640 22.2684941 29.275289
## Sep 2024      21.748535 19.4571205 24.039949 18.2441190 25.252950
## Oct 2024      15.570614 13.2785345 17.862694 12.0651808 19.076048
## Nov 2024      10.166892  7.8741467 12.459637  6.6604409 13.673342
## Dec 2024       5.062726  2.7683799  7.357073  1.5538263  8.571627

plot(prognoza_ets, main = "Prognoza na podstawie ETS", ylab = "Temperatura [°
C]", xlab = "Rok")
```



Prognoza oparta na modelu ETS wskazuje, że, podobnie jak w przypadku prognozy modelu SARIMA, przewidywane temperatury w nadchodzącym roku będą zbliżone do poziomów obserwowanych w poprzednich latach. Jednakże, przy wyższych przedziałach ufności, modele biorą pod uwagę możliwość wystąpienia rekordowo ciepłego lata w 2024 roku. Z drugiej strony, nawet w najzimniejszym scenariuszu, zima 2024 roku nie powinna być chłodniejsza niż najzimniejsze zimy w okresie od 2000 do 2023 roku.

```
accuracy(prognoza_ets)[,kryteria]

##      MAE     RMSE
## 1.343735 1.739943
```

Średnia wartość bezwzględna błędu predykcji wynosi ok. 1,3°C. Z kolei wynik RMSE świadczy o tym, że prognoza średnio myli się o ok. +/- 1,7°C. Kryteria te mają podobne wartości dla prognoz opartych na każdym z modeli, jednakże prognozy na bazie modelu ETS obarczone są nieco większym błędem niż prognozy z modelu SARIMA.

Podsumowanie

Analiza miała na celu prognozowanie przyszłych średnich miesięcznych temperatur w nowojorskim Central Parku na podstawie danych historycznych z okresu 2000-2023, pochodzących z National Centers for Environmental Information (NCEI).

Do analizy wykorzystano dwa modele prognozowania: SARIMA oraz ETS. Dane zostały wstępnie przetworzone poprzez konwersję wartości temperatury ze stopni Fahrenheita na stopnie Celsjusza i zaokrąglenie wyników. Wykresy autokorelacji i częściowej autokorelacji ujawniły roczne i półroczne wzorce sezonowe, co uzasadniało zastosowanie modeli sezonowych.

Wybrano model ARIMA(0,0,1)(2,1,2)[12]. Wyniki testów diagnostycznych, w tym testu Shapiro-Wilka i testu Boxa-Ljunga, wykazały, że reszty modelu nie mają idealnego rozkładu normalnego, ale nie wykazują znaczącej autokorelacji. BIC modelu wyniósł 1149,19, co wskazuje na dobre dopasowanie do danych.

Prognozy modelu ETS wykazały większe odchylenia w porównaniu do modelu SARIMA. BIC modelu wyniósł 2034,896, co sugeruje, że jest mniej efektywny niż model SARIMA.

Model SARIMA okazał się bardziej adekwatny do prognozowania średnich miesięcznych temperatur w Central Parku w porównaniu z modelem ETS, co potwierdziły niższe wartości BIC oraz brak znaczącej autokorelacji w resztach modelu. Pomimo pewnych niedoskonałości, model SARIMA jest w stanie dobrze odwzorować sezonowe wzorce temperatur w badanym okresie.

Dalsze badania mogłyby skupić się na uwzględnieniu dodatkowych zmiennych, takich jak opady deszczu czy poziomy zanieczyszczeń, aby jeszcze bardziej poprawić dokładność prognoz.