Projekt z Szeregów Czasowych

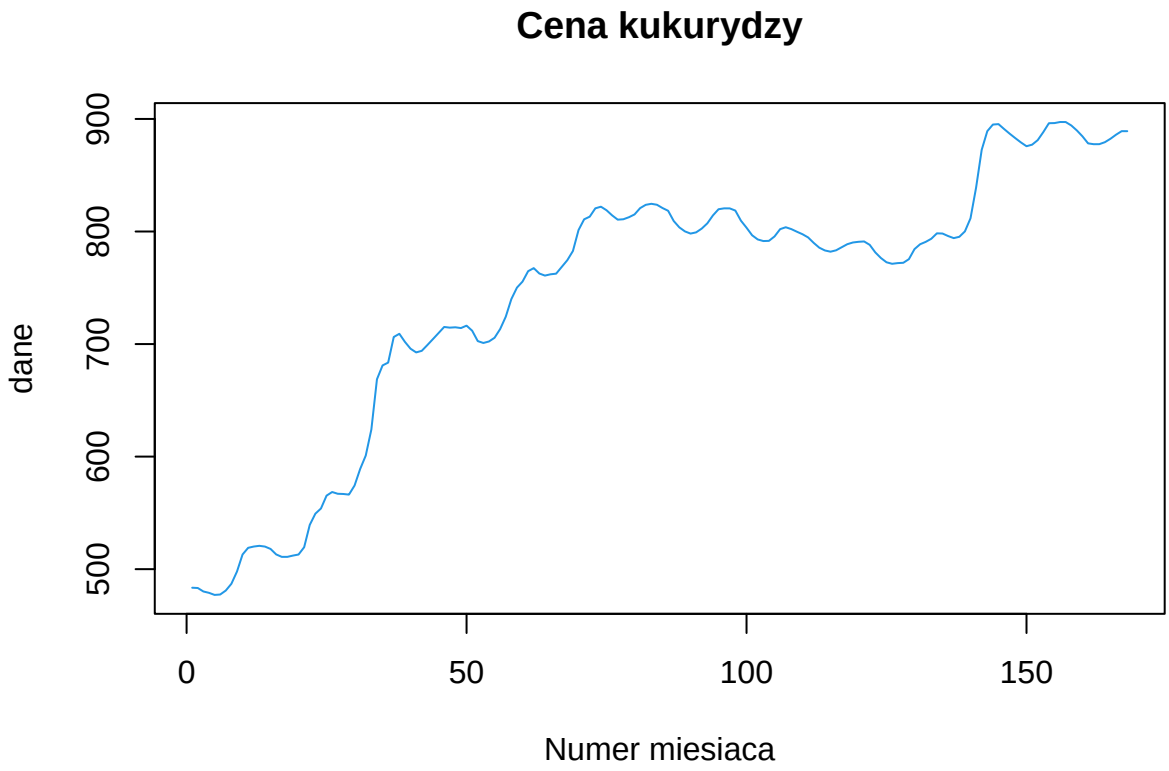
Jan Moskal i Szymon Makulec

2025-01-18

Wstęp

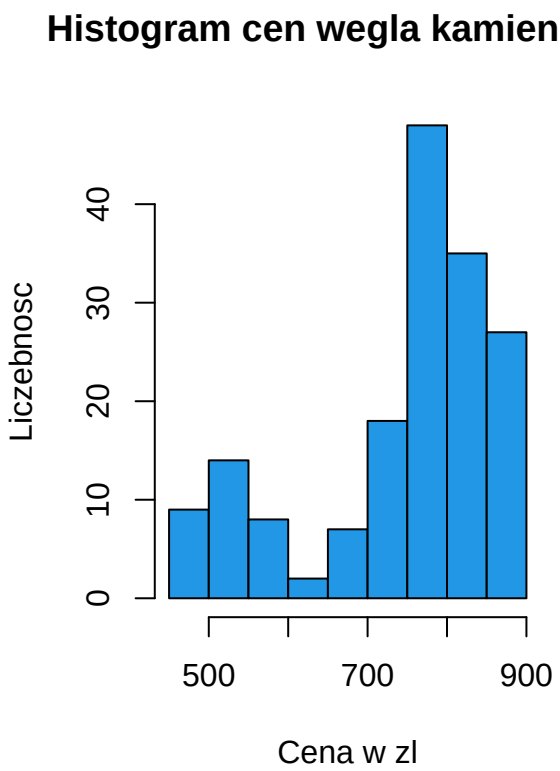
Dane, które będziemy analizować, pochodzą ze strony Głównego Urzędu Statystycznego (https://bdl.stat.gov.pl/bdl/dane/podgrup/temat) znajdują się w grupie "Przeciętne ceny detaliczne towarów i usług konsumpcyjnych", w podgrupie "Ceny detaliczne wybranych towarów i usług konsumpcyjnych (dane miesięczne)" i dotyczą cen węgla kamiennego za toną. Dane o przeciętnych cenach obejmują notowania co miesiąc dla całej Polski. Projekt ma na celu analizę tego szeregu czasowego, aby zrozumieć zmiany cen węgla kamiennego w Polsce w latach 2006-2019.

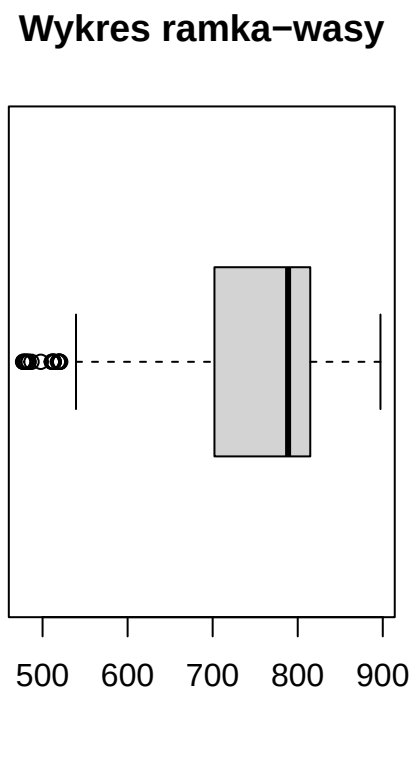
Wstępna analiza szeregu



Jak widzimy z wykresu, cena dość szybko wzrosła do cen powyżej 700 zł. Widzimy również, że ogólny trend jest rosnący.

Robimy wykresy typu boxplot oraz histogram, żeby zobaczyć rozkład danych.



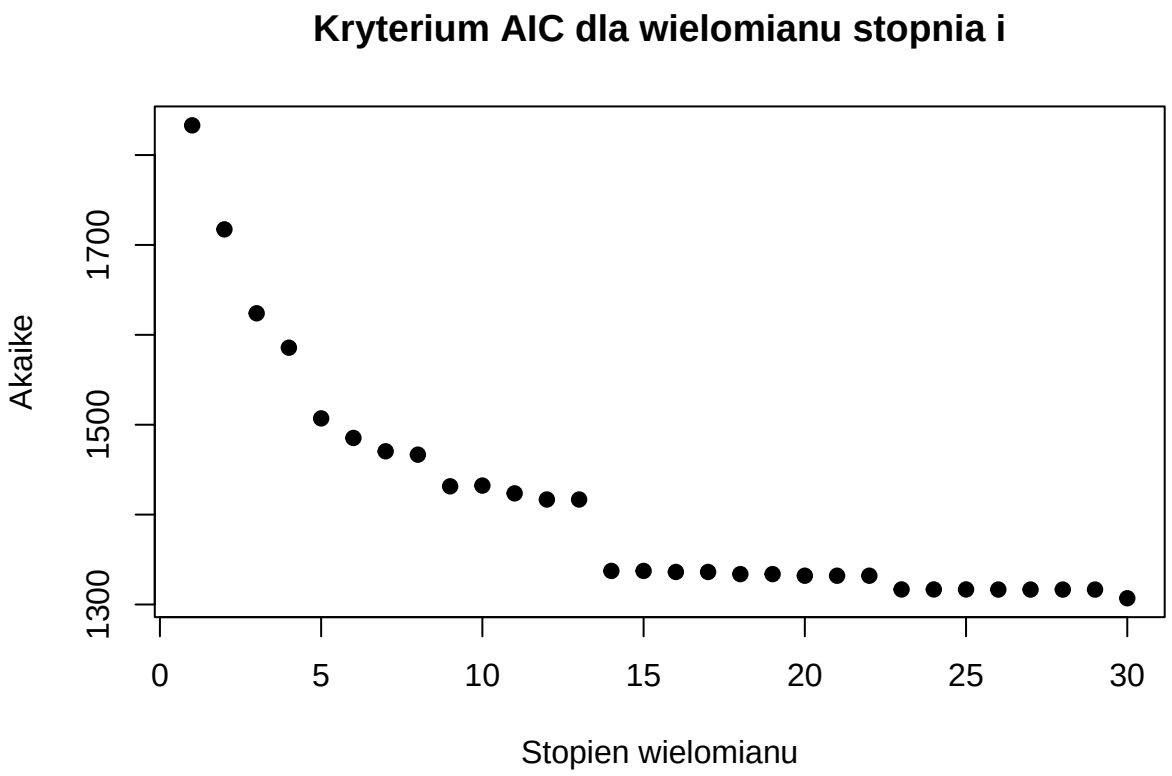


Możemy zauważyć, że rozkład jest lewostronnie asymetryczny, bierzę się to z tego co już zauważyliśmy z wykresu liniowego czyli, że ceny od 700 zł za tonę zaczęły się już po 2 latach od pierwszej obserwacji z szeregu a pozostałe 12 lat oscylowało co do wartości od 700 do 900 zł za tonę. Z wykresu pudełkowego możemu zauważyć nawet dokładniej, że kwartyl pierwszy wynosi około 700 a kwartyl trzeci około 810 co w przełożeniu na nasz problem oznacza, że połowa obserwacji, czyli z 7 lat znajduje się na tym małym przedziale.

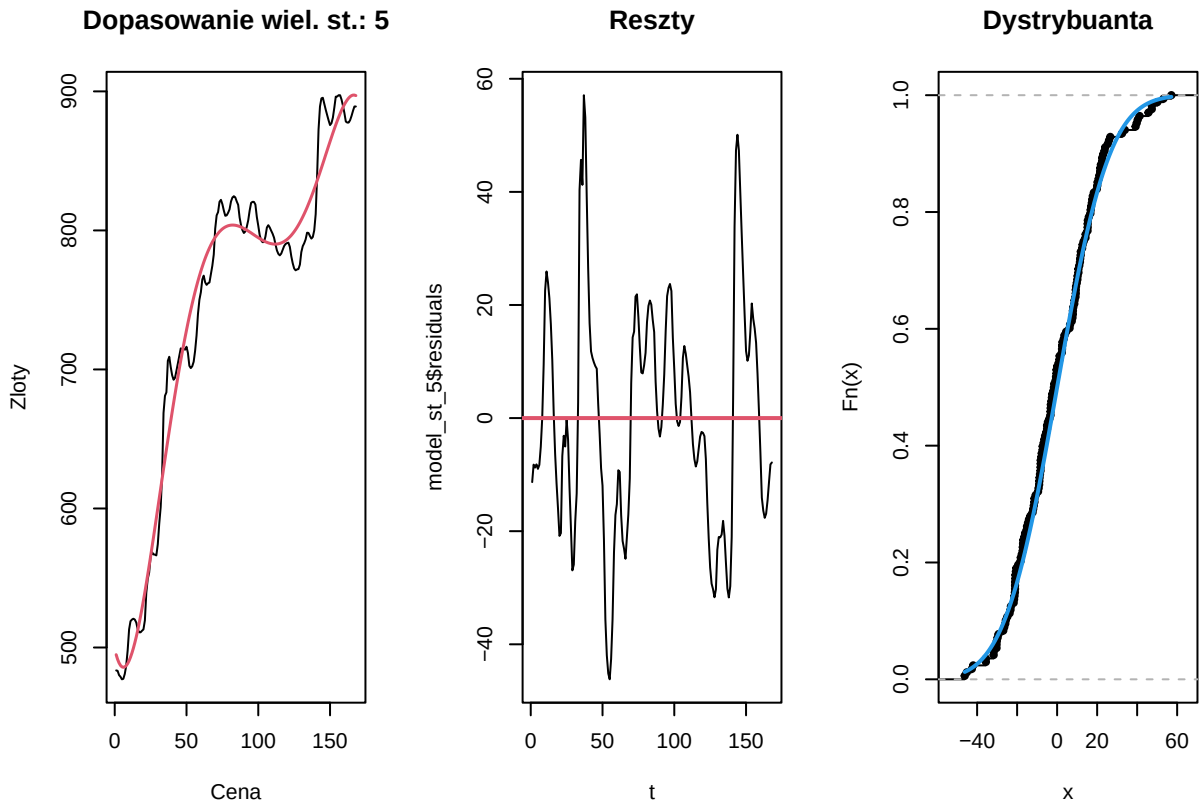
Podstawowe statystyki

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   477.1   702.2   788.7   744.1   814.5   897.3
```

Szukamy najlepszego wielomianu opisującego nasz szereg



Z kryterium wyboru stopnia wielomianu wybieramy wielomian stopnia 5.



Badanie reszt

Zbadamy reszty z modelu stopnia 5.

```
##
##  Runs Test
##
## data:  reszty
## statistic = -11.143, runs = 13, n1 = 84, n2 = 84, n = 168, p-value <
## 2.2e-16
## alternative hypothesis: nonrandomness
```

```
##
##  Asymptotic one-sample Kolmogorov-Smirnov test
##
## data:  reszty
## D = 0.48095, p-value < 2.2e-16
## alternative hypothesis: two-sided
```

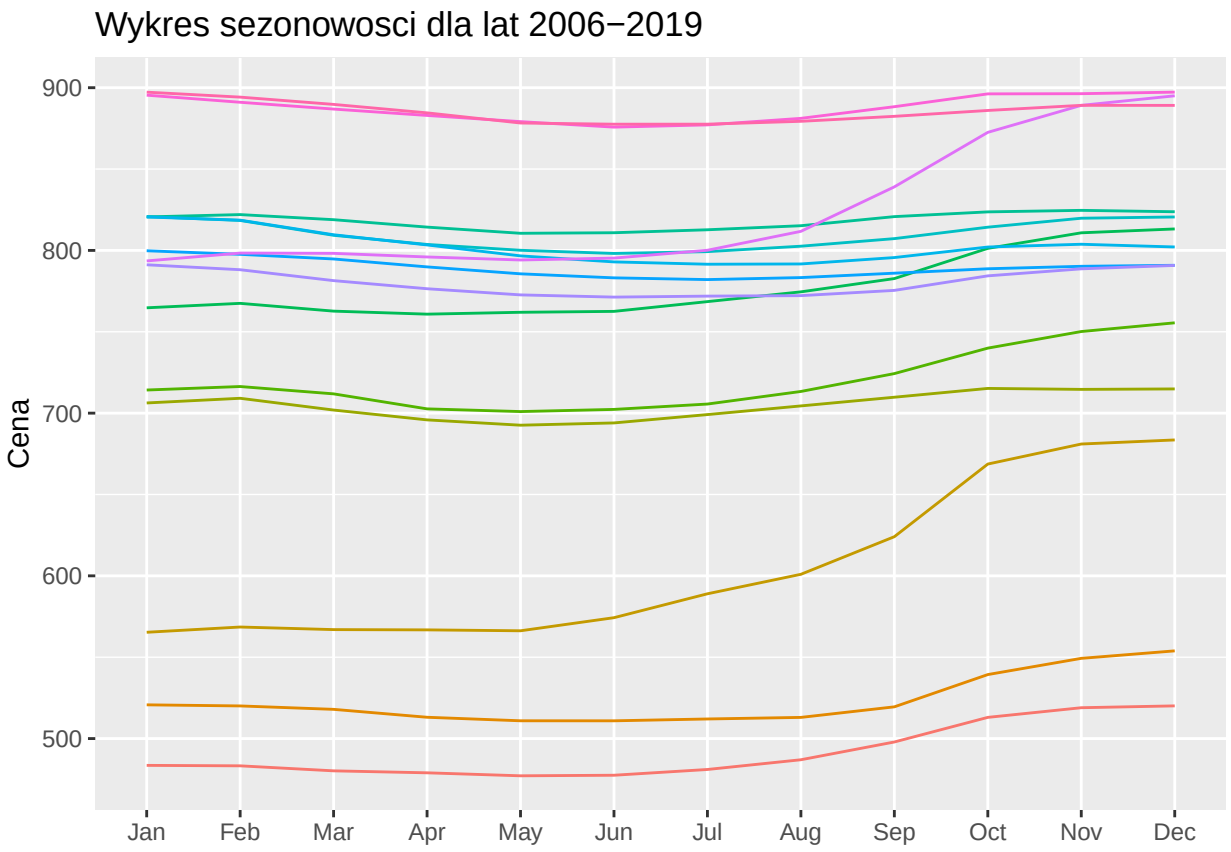
```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  reszty
## D = 0.035656, p-value = 0.8673
```

```
##
##  Shapiro-Wilk normality test
##
## data:  reszty
## W = 0.98779, p-value = 0.1533
```

```
##
##  Box-Ljung test
##
## data:  reszty
## X-squared = 390.21, df = 12, p-value < 2.2e-16
```

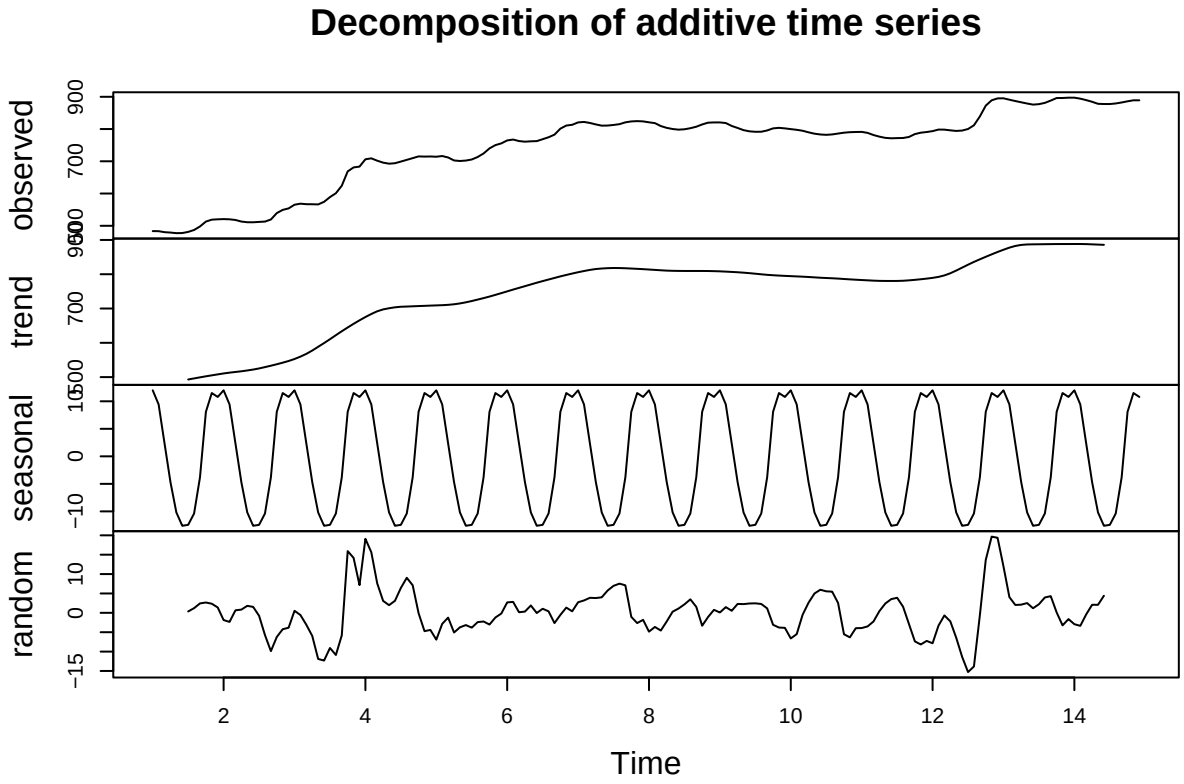
Wyniki sugerują, że wybrany model nie jest w stanie odpowiednio uchwycić struktury danych, ponieważ reszty są skorelowane, mają nienormalny rozkład i odrzucają hipotezy o losowości. Może to oznaczać, że stopień wielomianu jest nieodpowiedni lub że struktura danych wymaga bardziej złożonego modelu.

Analiza trendów fazowych



Z wykresu sezonowości widzimy powtarzający się trend wzrostu cen węgla kamiennego w okresie od sierpnia do listopada. W okresie od stycznia do maja zauważalny jest nieznaczny trend spadkowy cen.

Dekompozycja



Po wykonaniu dekompozycji widzimy, że trend jest rosnący w dziedzinie. Widzimy również, że występuje sezonowość w częstotliwości 12 miesięcznej. W kwestii reszt wydają się one oscylować wokół zera, jednak przez ich nieregularność będzie się trzeba im lepiej przyjrzeć.

Stacjonarność szeregu

```
##
##  Augmented Dickey-Fuller Test
##
## data:  dane
## Dickey-Fuller = -1.7305, Lag order = 5, p-value = 0.6888
## alternative hypothesis: stationary
```

```
## Warning in kpss.test(dane): p-value smaller than printed p-value
```

```
##
##  KPSS Test for Level Stationarity
##
## data:  dane
## KPSS Level = 2.7326, Truncation lag parameter = 4, p-value = 0.01
```

```
##
##  Phillips-Perron Unit Root Test
##
## data:  dane
## Dickey-Fuller Z(alpha) = -4.5142, Truncation lag parameter = 4, p-value
## = 0.856
## alternative hypothesis: stationary
```

Z wszystkich testów wynika, że szereg jest niestacjonarny.

```
## Series: dane
## ARIMA(4,1,0) with drift
##
## Coefficients:
##          ar1      ar2     ar3      ar4   drift
##       0.9059  -0.4425  0.5447  -0.4685  2.4353
## s.e.  0.0676   0.0881  0.0872   0.0670  0.7722
##
## sigma^2 = 21.63:  log likelihood = -491.99
## AIC=995.97   AICc=996.5   BIC=1014.68
```

Model ARIMA(4,1,0) wskazuje, że szereg czasowy wymagał różnicowania pierwszego rzędu, aby stać się stacjonarnym. Współczynniki autoregresyjne sugerują zależność od czterech poprzednich wartości, a obecność dryfu oznacza trend wzrostowy. Niemamy składnika średniej ruchomej. Teraz przejdziemy do zbadania reszt.

```
##
##  Runs Test
##
## data:  reszty
## statistic = -0.31874, runs = 82, n1 = 75, n2 = 93, n = 168, p-value =
## 0.7499
## alternative hypothesis: nonrandomness
```

Nie odrzucamy hipotezy o losowości reszt.

```
##
##  One Sample t-test
##
## data:  reszty
## t = -0.016825, df = 167, p-value = 0.9866
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  -0.7036653  0.6917731
## sample estimates:
##   mean of x
## -0.00594608
```

Nie odrzucamy hipotezy o średniej równej 0.

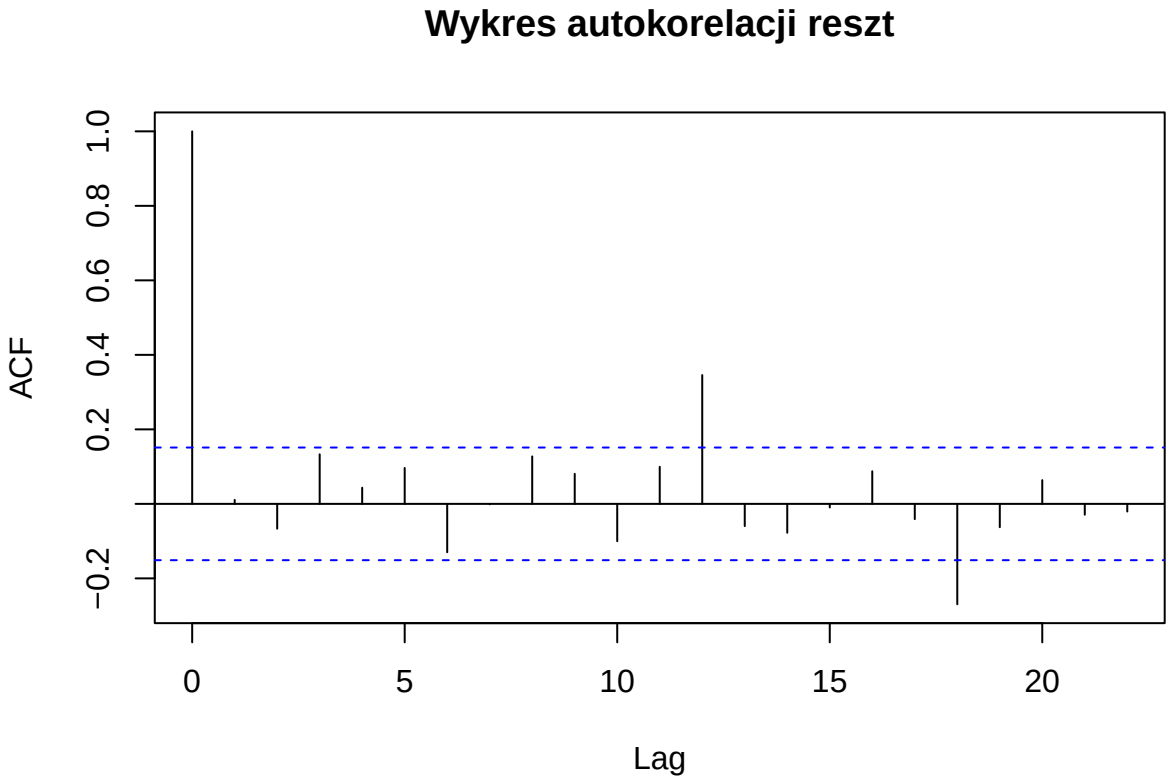
```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  reszty
## D = 0.10715, p-value = 6.983e-05
```

```
##
##  Shapiro-Wilk normality test
##
## data:  reszty
## W = 0.89614, p-value = 1.765e-09
```

Odrzucamy hipotezę o normalności reszt.

```
##
##  Box-Ljung test
##
## data:  reszty
## X-squared = 38.365, df = 12, p-value = 0.0001338
```

Odrzucamy hipotezę o braku korelacji w resztach.




```
##
##  studentized Breusch-Pagan test
##
## data:  reszty ~ t
## BP = 2.8114, df = 1, p-value = 0.0936
```

Nie odrzucamy hipotezy o jednorodności wariancji.

Model ARIMA(4, 1, 0), mimo że dobrze odwzorowuje poziom szeregu czasowego, nie wychwytuje wszystkich zależności w danych, co widać po obecności autokorelacji w resztach. Wskazuje to na potencjalną potrzebę dalszej optymalizacji modelu, aby lepiej uchwycić zależności czasowe. Brak heteroskedastyczności sugeruje, że wariancja reszt jest stabilna w czasie, więc model GARCH nie jest konieczny.

Podsumowanie projektu

W ramach analizy szeregu czasowego cen węgla kamiennego w Polsce w latach 2006-2019, przeprowadziliśmy szereg kroków w celu zrozumienia struktury danych. Dopasowanie wielomianu stopnia 5 nie poprawiło jakości modelu, ponieważ reszty z tego modelu były skorelowane i miały nienormalny rozkład. Dekompozycja szeregu ujawniła rosnący trend i sezonowość, z wyraźnym wzrostem cen w okresie od sierpnia do listopada. Testy stacjonarności wskazały na niestacjonarność szeregu, co skutkowało zastosowaniem modelu ARIMA(4,1,0), który po różnicowaniu pierwszego rzędu stabilizował szereg. Mimo że model ARIMA dobrze odwzorowywał poziom szeregu, reszty wykazywały autokorelację, co sugeruje konieczność dalszej optymalizacji modelu w celu pełniejszego uchwycenia zależności czasowych. Możnabyłoby kontunuować pracę aby otrzymać lepiej radzący sobie model jednak my w tym miejscu zakończymy.