

Analiza szeregu czasowego

Liczba udzielonych noclegów w hotelach w Polsce 2019-2024 r.



***Patrycja Mazur Weronika Poniedziałek***

***Nr. albumu: 100991 Nr. albumu: 101002***

## WSTĘP

Celem projektu jest analiza szeregu czasowego ilustrującego miesięczną liczbę udzielonych noclegów ogółem w hotelach w Polsce w latach 2019–2024. Dane zostały pobrane z serwisu GUS (Bank Danych Lokalnych):

https://bdl.stat.gov.pl/bdl/metadane/podgrupy/478

Analiza szeregu czasowego obejmuje identyfikację kluczowych cech, takich jak stacjonarność, sezonowość oraz występowanie zależności autokorelacyjnych. Wykorzystane metody pozwolą na charakterystykę dynamiki zmian w liczbie noclegów oraz na określenie potencjalnych wzorców i zależności w danych.

## Obraz zawierający tekst, diagram, Wykres, liniaWIZUALIZACJA DANYCH

Wykres pokazuje ogólną zmienność danych w czasie, z wyraźnymi sezonowymi wzorcami.

Bardzo wyraźnie widać duży spadek na początku roku 2020 trwający do połowy 2021 r. z wyraźnym jednym wzrostem w połowie 2020 r. Wiedząc, że w tych latach panował COVID-19, możemy przypuszczać, że to właśnie przez to liczba wynajmowanych noclegów gwałtownie spadła, gdyż był to okres utrudnionych wyjazdów.

Możemy zauważyć, że w połowie roku liczba wynajmowanych pokoi jest znacząco wyższa niż na końcu bądź początku roku. Możemy wyciągnąć wnioski, że ludzie o wiele więcej wyjeżdżają w wakacje niż na święta. Wskazuje to na preferencje podróżnicze związane z porami roku.

Obraz zawierający diagram, tekst, Wykres, linia

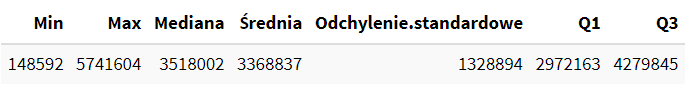
Opis wygenerowany automatycznie

Obraz zawierający tekst, zrzut ekranu, diagram, linia

Opis wygenerowany automatycznie

Wykresy wskazują, że w niektórych miesiącach liczba noclegów jest wyraźnie wyższa (np. w miesiącach letnich), a w innych niższa (np. w miesiącach zimowych). Możemy zauważyć dużą zmienność danych w zależności od miesiąca. Są to istotne wahania sezonowe.

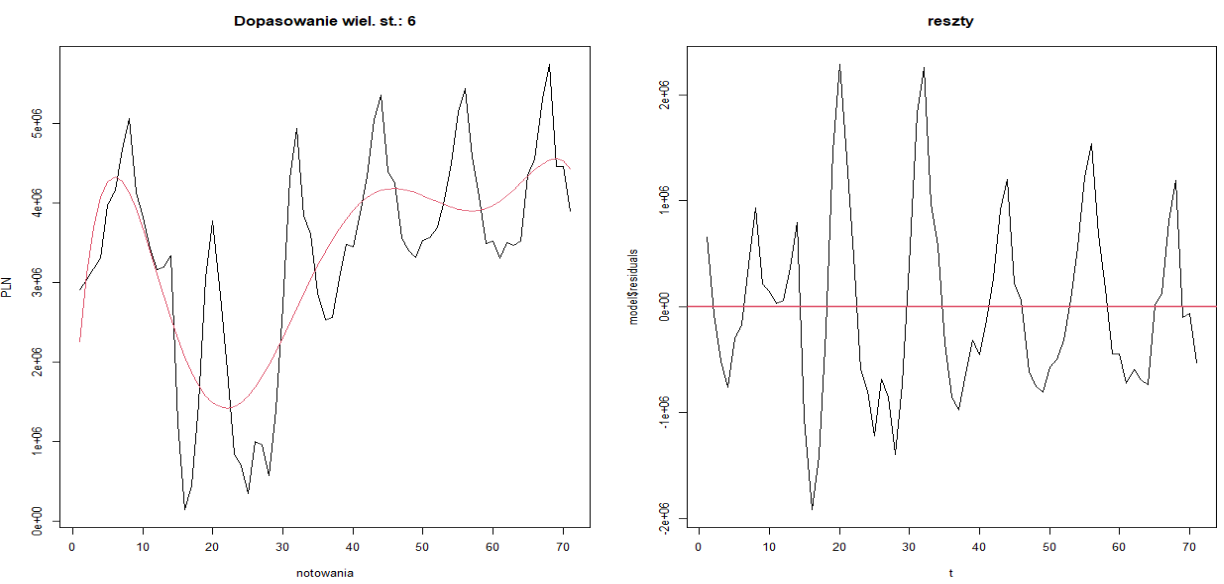
## **STATYSTYKI OPISOWE**



## **DOPASOWANIE MODELU WIELOMIANOWEGO**

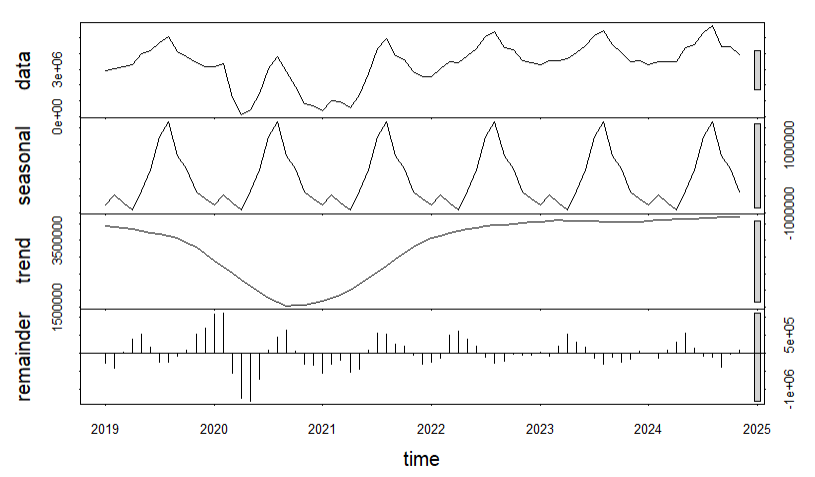
Obraz zawierający tekst, zrzut ekranu, linia, diagram

Opis wygenerowany automatycznieW wyniku analizy szeregu czasowego, model wielomianowy o stopniu 6 najlepiej opisuje dane, minimalizując kryterium AIC. To optymalne dopasowanie zapewnia równowagę między jakością modelu a jego złożonością, skutecznie przewidując przyszłe wartości w szeregu czasowym.



## **DEKOMPOZYCJA ZA POMOCĄ FUNKCJI STL**

Wykres dekompozycji STL przedstawia trzy główne składniki danych:

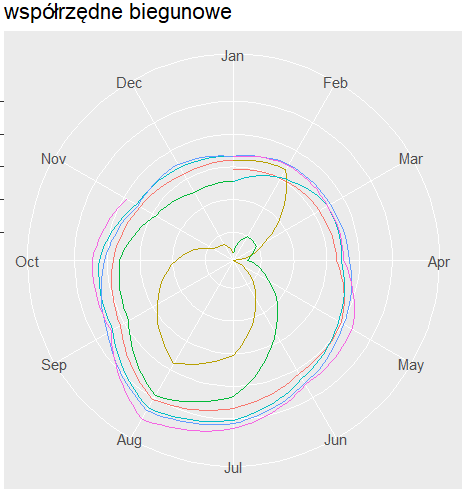
**- Trend**: W 2020 roku zauważalny był duży spadek, a od 2022 roku obserwuje się delikatny wzrost w porównaniu do 2019 roku.

**- Sezonowość**: Powtarzające się wzorce, szczególnie widoczne w środku roku, kiedy obserwuje się wzrost wynajmowanych hoteli.

**- Reszty**: W 2020 roku występują istotne odchylenia w resztach, związane z anomaliami, takimi jak pandemia COVID-19.

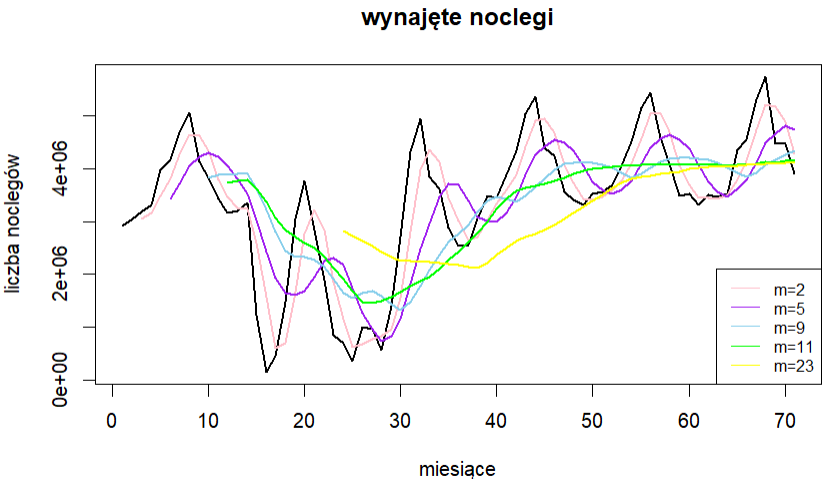
## **SEZONOWOŚĆ**

Obraz zawierający diagram, Wykres, linia, tekst

Opis wygenerowany automatycznie

Wykresy sezonowe potwierdzają, że dane charakteryzują się cyklicznością, czyli w określonych miesiącach liczba udzielonych noclegów wzrasta (miesiące letnie) lub spada (miesiące zimowe). Sezonowość jest stabilna w czasie, co oznacza, że wzorce w danych są powtarzalne co roku.

## **METODA ŚREDNIEJ RUCHOMEJ**

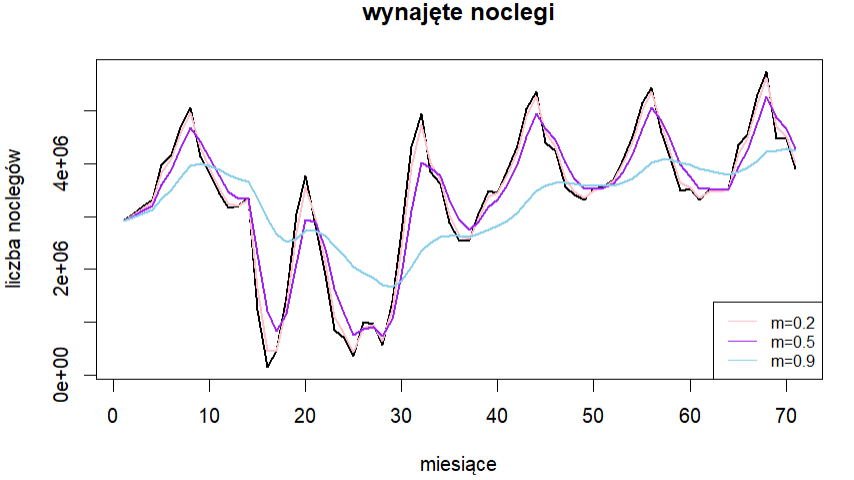


Szereg został wygładzony, eliminując krótkoterminowe wahania i podkreślając długoterminowe trendy. Średnia ruchoma polega na obliczeniu średniej wartości danych na określonym okresie (m), przy czym w każdej iteracji uwzględnia się kolejne wartości, przesuwając "okno" analizowanej serii danych. Dzięki temu identyfikujemy ogólne wzorce i zmiany w szeregu czasowym.

## **METODA WYKŁADNICZYCH WAG RUCHOMEJ ŚREDNIEJ**

Metoda ta w przeciwieństwie do prostej średniej ruchomej daje bardziej dynamiczny obraz trendów w szeregach czasowych oraz przypisuje większy wpływ najnowszym danym.

Pozwala na szybsze uwzględnianie zmian w danych.



Wraz ze wzrostem współczynnika „m”, wykładnicza średnia ruchoma staje się bardziej wygładzona, ponieważ wpływ najnowszych danych ma większe znaczenie, a wahania starszych wartości są stopniowo minimalizowane.

## **ANALIZA LOSOWOŚCI DANYCH**

runs.test() – Test losowości (test przejść), jest używany, aby sprawdzić, czy reszty są rozmieszczone losowo, tzn. czy zmiany w danych nie są zależne od poprzednich wartości.

**Hipoteza zerowa (H₀):** Reszty są rozmieszczone losowo — zmiany w danych nie są powiązane.

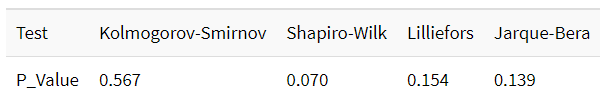
**Hipoteza alternatywna (H₁):** Reszty są nierównomiernie rozmieszczone — istnieje jakiś ukryty czynnik wpływający na dane.

Wynik testu biegów (p-value = 5.65e-08) jest bardzo mały, co oznacza, że odrzucamy hipotezę zerową o braku odchyleń od losowości.

Obraz zawierający tekst, diagram, Wykres, linia

Opis wygenerowany automatycznie

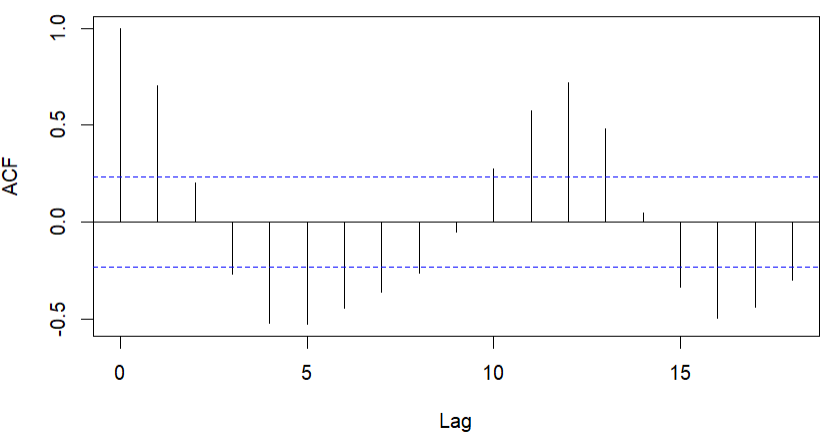
## **NORMALNOŚĆ**



W każdym z przeprowadzonych testów, p-wartość>0.05, więc nie mamy podstaw do odrzucenia hipotezy zerowej. Zatem przyjmujemy, że reszty mają rozkład normalny.

## **KORELACJE**

Autokorelacja reszt (ACF - Autocorrelation Function) mierzy, jak mocno dana wartość reszt jest związana z wcześniejszymi wartościami.

Obraz zawierający diagram, linia, Rysunek techniczny, Równolegle

Opis wygenerowany automatycznieWyświetlany jest wykres autokorelacji dla opóźnień do 50.

## **TESTY DOTYCZĄCE KORELACJI**

### Test istotności współczynnika korelacji dla przesunięcia 𝜏

*𝜏*= 2

**Hipoteza zerowa (H₀):** Brak istotnej korelacji między zmiennymi dla przesunięcia 𝜏.

**Hipoteza alternatywna (H₁):** Istnieje istotna korelacja między zmiennymi dla przesunięcia 𝜏.

p-value = 0.09147 > 0.05

Nie mamy podstaw do odrzucenia H₀.

### Test Ljung-Box – test na obecność autokorelacji dla wielu opóźnień

𝜏= 15

**Hipoteza zerowa (H₀):** Brak istotnej autokorelacji – reszty są niezależne dla wszystkich opóźnień (1, 2, …, 15).

**Hipoteza alternatywna (H₁):** Istnieje istotna autokorelacja – przynajmniej jedna autokorelacja reszt różni się od zera dla przynajmniej jednego opóźnienia.

p-value < 2.2e-16

Odrzucamy hipotezę zerową, zatem możemy wnioskować, że dla co najmniej jednego opóźnienia istnieje istotna autokorelacja.

## **TESTY STACJONARNOŚCI**

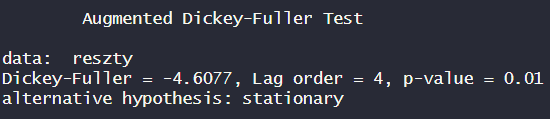
### Test Dickeya-Fullera (ADF)

Obraz zawierający design

Opis wygenerowany automatycznie przy niskim poziomie pewności**Hipoteza zerowa (H₀):** Szereg jest niestacjonarny.

P – wartość < 0.05, więc odrzucamy H0

Obraz zawierający symbol, clipart, kreatywność

Opis wygenerowany automatycznie

### Obraz zawierający design Opis wygenerowany automatycznie przy niskim poziomie pewnościTest KPSS

**Hipoteza zerowa (H₀):** Szereg jest stacjonarny.

P – wartość > 0.05, więc nie mamy podstaw do odrzucenia H0

Obraz zawierający symbol, clipart, kreatywność

Opis wygenerowany automatycznieObraz zawierający tekst, Czcionka, zrzut ekranu

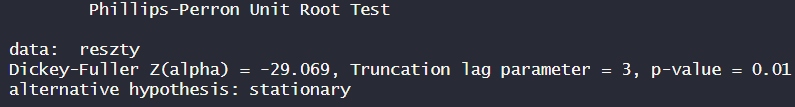
Opis wygenerowany automatycznie

### Obraz zawierający design Opis wygenerowany automatycznie przy niskim poziomie pewnościTest Phillipsa-Perrona (PP)

P – wartość < 0.05, więc odrzucamy H0

**Hipoteza zerowa (H₀):** Szereg jest niestacjonarny.

Obraz zawierający symbol, clipart, kreatywność

Opis wygenerowany automatycznie

Wyniki testów pomagają ocenić, czy wygenerowane szeregi są stacjonarne.

## **HOMOSKEDASTYCZNOŚĆ**

Na podstawie testów Breuscha-Pagana i Goldfelda–Quandta nie mamy podstaw do odrzucenia hipotezy zerowej o jednorodności wariancji reszt.

## **MODEL ARIMA DLA RESZT**

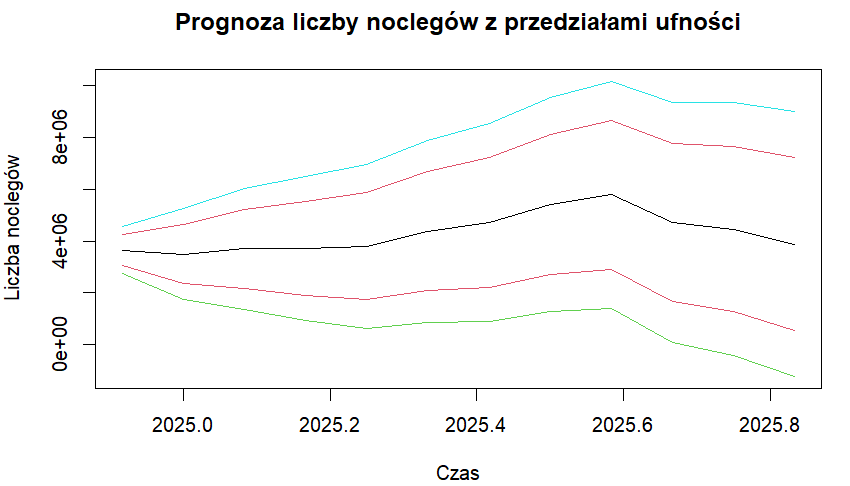
auto.arima() – wybiera odpowiedni model Arima (0,1,1)

Obraz zawierający tekst, diagram, Wykres, linia

Opis wygenerowany automatycznie

Model ARIMA został automatycznie dopasowany do danych, uwzględniając zarówno trend, jak i sezonowość w danych. Dopasowany model można uznać za dobrze opisujący dane, co widać po wynikach diagnostycznych i zgodności z obserwowanymi danymi historycznymi.

## **DALSZE WIZUALIZACJE PROGNOZY Z PRZEDZIAŁAMI UFNOŚCI**



Wykres pokazuje nie tylko przewidywaną wartość, ale także zakres niepewności tych przewidywań.

- Czarna środkowa linia to prognoza liczby noclegów.

- Zielona linia to dolna granica przedziału ufności.

**Trend**: Prognoza liczby noclegów rośnie od początku roku do połowy 2025 roku, osiągając maksimum w okolicach lata (lipiec/sierpień), co może być związane z sezonem wakacyjnym.

**Zakres niepewności**:

* Na początku i końcu roku przedziały ufności (odległość między zieloną a niebieską linią) są szersze, co oznacza większą niepewność prognozy.
* W środku roku, kiedy liczba noclegów jest najwyższa, przedziały są węższe, co sugeruje większą pewność w przewidywaniu.

## **PODSUMOWANIE**

Projekt analizuje miesięczną liczbę noclegów w polskich hotelach w latach 2019–2024, skupiając się na trendach, sezonowości i wpływie czynników zewnętrznych, takich jak pandemia COVID-19.

Dane wykazały wyraźne wzorce sezonowe, z najwyższymi wartościami latem i spadkami zimą, a także znaczący spadek liczby noclegów na początku pandemii, po którym nastąpiło gwałtowne odbicie.

Wykorzystano modele statystyczne, w tym analizę STL i ARIMA, aby uchwycić kluczowe zależności oraz przygotować prognozy na 2025 rok, które przewidują wzrost liczby noclegów latem i spadek pod koniec roku (czyli tak jak co roku).

Wyniki, które otrzymałyśmy w dużej mierze mogą wspierać branże turystyczną.

Obraz zawierający symbol, clipart, kreatywność

Opis wygenerowany automatycznieObraz zawierający design

Opis wygenerowany automatycznie przy niskim poziomie pewności