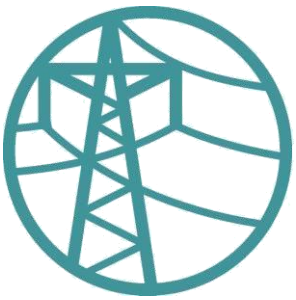


WYDZIAŁ
ELEKTROTECHNIKI
I INFORMATYKI
POLITECHNIKI RZESZOWSKIEJ

Zakład Systemów Złożonych

Flejszar Grzegorz & Jędrzejczyk Jakub

Szereg czasowy

Projekt zaliczeniowy

Opiekun:
dr inż. Marek Bolanowski

Rzeszów, 2023

1. Temat projektu

Tematem projektu była próba przeanalizowania ramki danych zawierającej dzienne minimalne temperatury w Melbourne w Australii (link do bazy danych: https://www.kaggle.com/datasets/paulbrabban/daily-minimum-temperatures-in-melbourne) przy użyciu szeregu czasowego.

2. Opis tworzenia projektu

2.1. Wczytanie i obróbka danych

Pierwszym krokiem było zaimportowanie bibliotek potrzebnych do wykonania analizy. Użyte biblioteki to:

- `pandas`
- numpy

- `matplotlib`
- `seaborn`
- `statsmodels`
- `prophet`
- `prophet`
- `sklearn.metrics`

2.2. Badanie stacjonarności i sezonalności szeregu

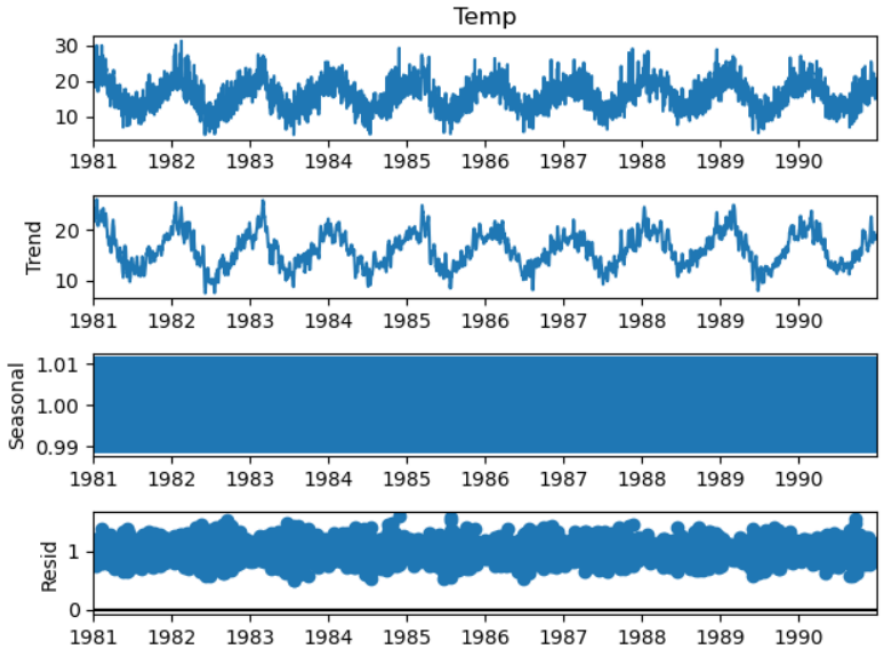
Stacjonarność szeregu oznacza, że jego statystyczne cechy, takie jak średnia są stałe w czasie i nie wykazuje tendencji do sezonowości czy innych niestabilnych wzorców, co ułatwia ich dalszą analizę i modelowanie. Aby ocenić, czy szereg czasowy jest stacjonarny, została użyta metoda testu Augmented Dickey-Fullera (ADF) wywołana funkcją „adfuller".

```
(-4.440933664385148,
 0.0002510472415292601,
 20,
 3631,
 {'1%': -3.4321522387754775,
  '5%': -2.862336328589075,
  '10%': -2.567193897993964},
 16651.240027625234)
```

Na podstawie informacji obliczonych przez funkcję można stwierdzić, że jest duże prawdopodobieństwo, że szereg jest stacjonarny. Świadczą o tym wartości -4.440933664385148 i 0.0002510472415292601, które są mniejsze od poziomów istotności 1%, 5% i 10%. Obie te wartości są też stosunkowo małe co zwiększa przekonanie o stacjonarności szeregu.

Sezonalność oznacza z kolei, że występują regularne i powtarzające się wzorce w danych w określonym czasie (np. co rok, co miesiąc lub co tydzień). Te wzorce mogą wynikać z czynników sezonowych

(np. pory roku lub święta). Sezonalność została sprawdzona dwoma metodami – metodą addytywną przy użyciu funkcji „additive” oraz metodą multiplikatywna „multiplicative”. W przypadku drugiej metody konieczne było dodanie 5 do wszystkich wartości temperatury aby wyeliminować przypadek, że wartość przyjmowana przez funkcję jest ujemna. Taki zabieg nie zmienia linii trendu a sprawił, że można było użyć metody multiplikatywnej.

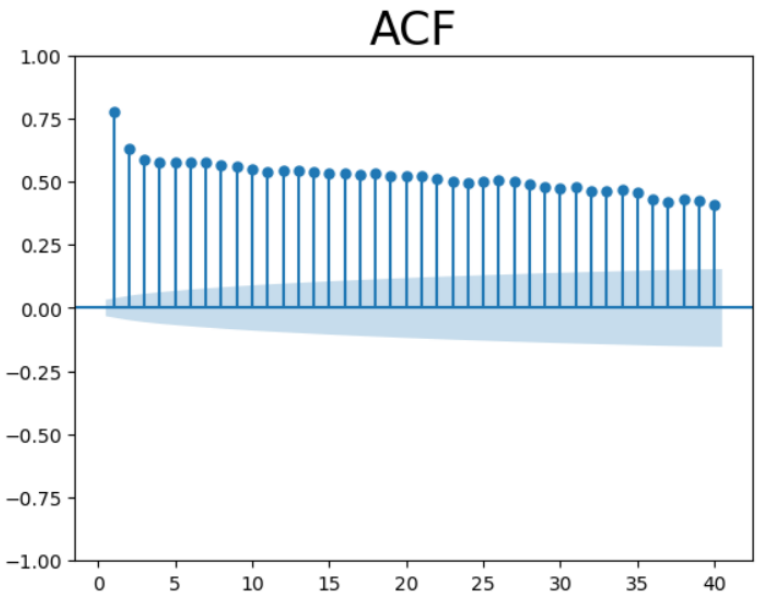


Z uwagi na to, że oryginalny wykres jak i wykres trendu są do siebie podobne można stwierdzić, że w analizowanym zbiorze danych nie ma sezonalności.

2.3. Badanie autokorelacji

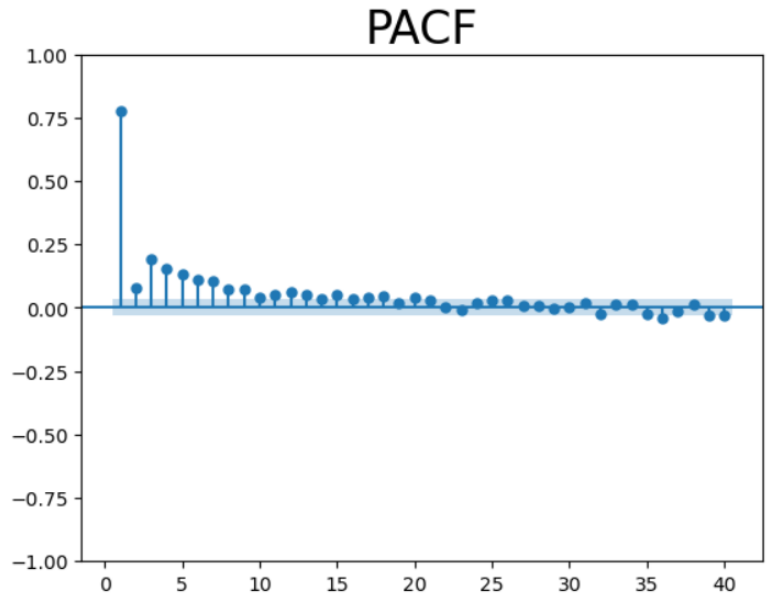
Autokorelacja w szeregach czasowych oznacza występowanie zależności pomiędzy wartościami w szeregu a ich opóźnieniami (lagami) i opisuje podobieństwo między wartościami obserwowanymi w różnych punktach czasu. Autokorelacja może być dodatnia lub ujemna. Dodatnia świadczy o tym, że zmienne są zależne od siebie w sposób można by rzecz „wprost proporcjonalny” co oznacza, że wzrost jednej wartości niesie za sobą wzrost drugiej wartości. Autokorelacja ujemna oznacza powiązaniu „odwrotnie proporcjonalnym” – wzrost wartości jednej ze zmiennych oznacza spadek wartości drugiej zmiennej.

Podobnie jak sezonalność autokorelacja została zbadana dwoma metodami. Pierwszą z nich była metoda ACF (Autocorrelation Function), wywołana funkcją „plot_acf”. Funkcja ta dla każdego opóźnienia (lagu) oblicz wartość korelacji między wartościami szeregu czasowego a wartościami przesuniętymi o dane opóźnienie. Następnie tworzy wykres wyliczonej korelacji, na osi X podane są opóźnienia a na osi Y wartość korelacji, której przedział mieści się w zakresie [-1,1].



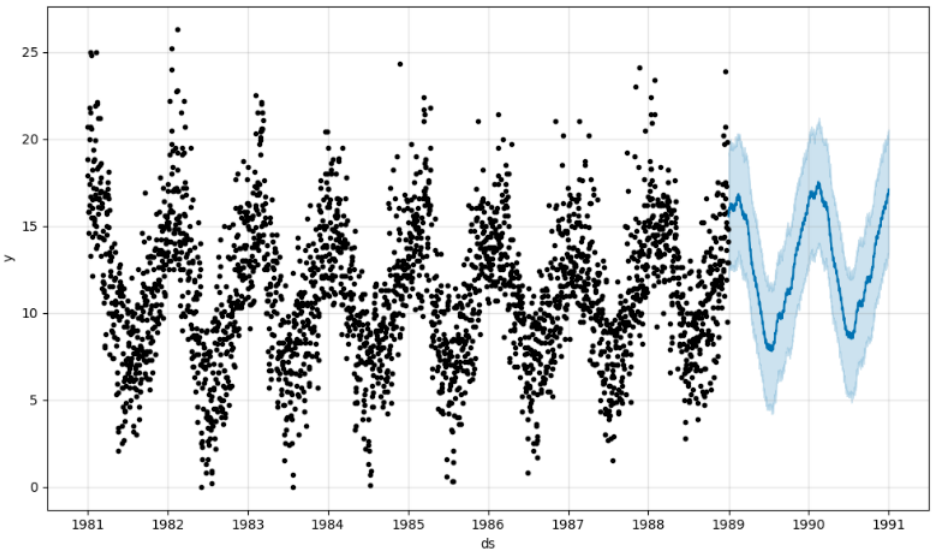
Wartości autokorelacji znajdują się dużo powyżej niebieskiego obszaru oraz dla opóźnień od 1 do około 20 są większe od 0.5 co świadczy o dużym stopniu korelacji między poszczególnymi wartościami temperatur, właśnie dla tych opóźnień. Innymi słowy wartości z poprzedniego okresu są dobrymi indykatorami przyszłych wartości. Natomiast wraz ze wzrostem opóźnień wartości autokorelacji maleją co wskazuje na to, że starsze wartości temperatur mają słabszy wpływ na przyszłe temperatury.

Druga metoda, czyli PACF (Partial Autocorrelation Function) do w przeciwieństwie do ACF nie oblicza autokorelacji biorąc pod uwagę cały zakres opóźnień a jedynie wartość początkową i końcową. W niektórych przypadkach ma to korzystny efekt, ale w przypadku naszej analizy tak nie jest. Wynika to z faktu, że znacznie trudniej jest przewidzieć temperaturę, gdy autokorelacja niemal na całym wykresie jest bliska zeru.

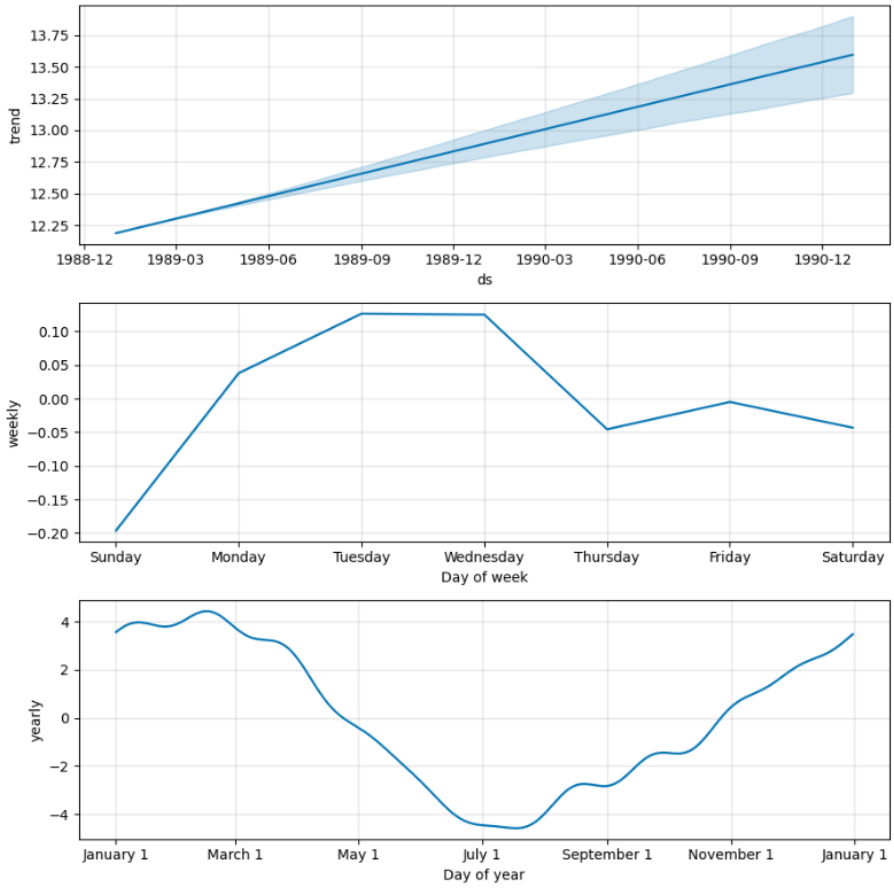


2.4. Modelowanie

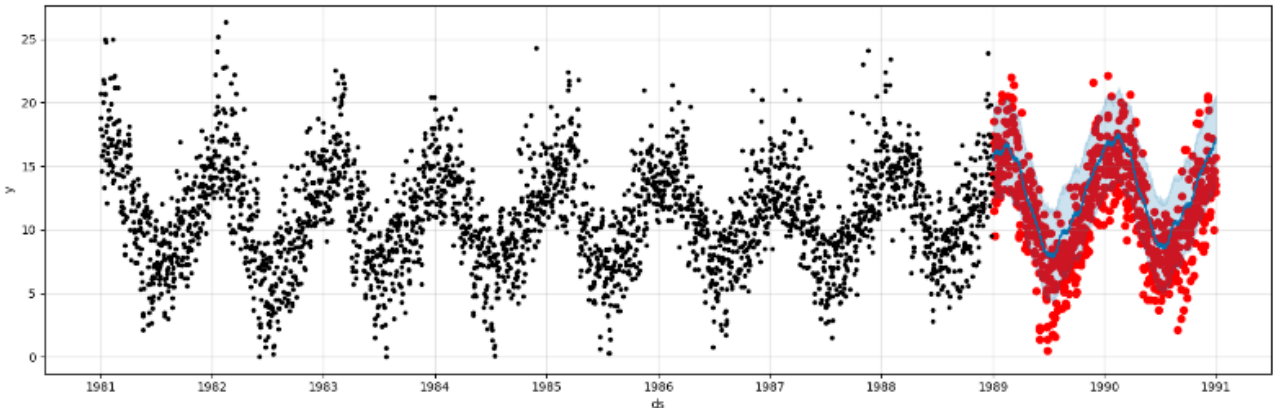
Do modelowania została użyta biblioteka Prophet, która została opracowana przez Facebook'a w celu prognozowania szeregów czasowych . Wywoływana jest przy użyciu funkcji o tej samej nazwie co nazwa biblioteki. Funkcja ta przyjmuje jednak tylko ściśle określoną ramkę danych – kolumny muszą się nazywać odpowiednio „ds" i „y", dlatego też przed wywołaniem funkcji „Prophet" konieczna była zmiana nazw kolumn ramki danych. Wynikiem działania modelu był następujący wykres:



Niebieska linia wskazuje najbardziej prawdopodobne temperatury jakie mogą wystąpić, a jasno-niebieskie widmo oznacza możliwy zakres w jakim wartości temperatur się znajdą.



Z kolejnych wykresów można wywnioskować, że predykcja wraz z czasem zmniejsza swoją skuteczność, ponieważ rozstrzał możliwych wartości przedstawionych na wykresie o nazwie „trend” wraz z upływem czasu jest coraz większy. Dodatkowo okazuje się, że statystycznie wtorki i środy w Melbourne były najcieplejszymi dniami tygodnia, a lipiec był najchłodniejszym miesiącem roku.



Ostatni wykres obrazuje jakie były rzeczywiste temperatury w stosunku do tego jak zostały przewidziane przez algorytm. Można zauważyć, że prawdziwe wartości temperatur były nieco niższe od tych wyliczonych przy użyciu szeregu czasowego. Potwierdza to średni błąd bezwzględny, wyliczony przy pomocy funkcji „mean_absolute_error”, który wyniósł 2.38 °C.

3. Problemy przy tworzeniu projektu

Początkowo w planie było aby wykorzystać do analizy dane na temat meteorytów które spadły na świecie. Nie było to jednak możliwe ponieważ wiele meteorytów spadało tego samego dnia, po czym następowała przerwa. Wykorzystanie algorytmów szeregów czasowych w takim przypadku nie miało by sensu ponieważ między datami byłyby duże przerwy i predykcje byłyby z tego powodu bardzo mało dokładne.