1. **Introduction Time Series**
   1. Szereg czasowy sklada się z 2 zmiennych: monotonicznego klucza(czas) oraz wartości wyrażonej przez ten klucz
   2. dane zrodlowe Rok – Kraj – kraj – kraj (w wierszach wartości)
   3. przekształcenie na szeregi czasowe (wczytać csv 🡪 przeksztalcic: R: funkcja ts)
   4. odczytac zakres i czestotliwosc (wartości ts . start / frequency / end) + ts.plot w wykresie liniowym / wyświetlenie brakujących danych funkcja „window”
   5. **CEL analizy czestościowej:** wyodrębnienie składowych (trend +szum+ ewentualnie skladowa cykliczna) oraz ocena wplwu czynnikow sezonowych
   6. Chcemy przeprowadzić prognozę na podstawie trendu a nie szumu czy cykliczności? Nie Arima ale: normalizacja seri danych, wygładzenie serii danych, wyodrębnienie składowych a potem usuniecie z serii np. składowej cyklicznej
   7. **Dekompozycja szeregu czasowego (model addytywny albo multiplikatywny)**  - funkcja, można zrobić scatterplot matrix
2. **Preprocessing** 
   1. Dla szerego niemonotonicznego można zastosować logarytmizacje
   2. **Wyodrębnienie składowych szeregu czasowego** Ewentualnie można usrednic dane, tak żeby niesystematyczne składniki poszczególnych obserwacji znosily się nawzajem –
      1. uzycie sredniej ruchomej (zastąpienie każdego elementu szeregu przez zwykla lub wazona srednia n sąsiadujących wartości, gdzie n jest szerokoscia okna wygładzenia) 🡪 R moving average filter pozwala okreslic szerokość okna wygładzenia i uzyta metode <https://machinelearningmastery.com/moving-average-smoothing-for-time-series-forecasting-python/> .
      2. W R szeregi czasowe można wygładzić m.in. za pomocą SMA
      3. **Runmed (R**
      4. **Decompose R**
      5. w dekompozycji pomocne jest tez obliczenie wspolczynnika autokorelacji ACF: jeśli kolejne wartości ACF zblizaja się do zera a następnie oscylują wokół niego to szereg czasowy zawiera skladowa stala / jesli zblizaja się do zera a następnie przyjmują wartości ujemne – to szereg ma trend / jeśli wartości ACF oscylują wokół zera i co pewien czas sa wyraźnie większe od zera to szereg czasowy zawiera wahania sezonowe
   3. Zamiast średnich można uzyc median, powinna być stosowana gdy w szerego wystepuja wartość odstające (mediana, w przeciwieństwie do sredniej nie zależy od wartości nietypowych). Zaleta: wyniki sa gładsze i przynajmniej bardziej reprezentatywne przy tej samej szerokości okna. Wada – nie można uzyc wag. R: median filter funkcje smooth i runmed
   4. **ANALIZA SEZONOWOSCI: trzeba zdefiniować sezonować:** Sezonowość (druga po trendzie skladowa szeregow czasowych) – zaleznosc korelacyjna rzedu k miedzy i-tym elementem szeregu a (i-k)-tym elementem. Miara sezonowości jest autokorelacja. Można ja usunąć poprzez przekształcenie każdego i-tego elementu szeregu na jego roznice z (i-k)-tym elementem. Robi się to po to, żeby:
3. Przygotować dane treningowe na potrzeby modeli ARIMA aby uzyskac szregi stacjonarne (szeregi o stalych w czasie sredniej, wariancji i autokorelacji).
4. Analiza aby zanalizować ukryta nature szregu

Metody sprawdzenia stacjonarności szeregu czasowego:

Box.test

Adf.test

Dickey-Fuller

Kpss.test

**Sprawdzić** czy szereg jest stacjonarny i czy potrzebne jest jego roznicowanie

1. **Wydzielenie danych testowych**
   1. nie można stosować losowego podzialu obserwacji na testowe i treningowe bo prowadzi do uzyskania niekompletnych szeregow.
   2. Należy ograniczyć dlugosc danych treningowych czyli usuwając z nich ostatnie x obserwacji, które posluza jako dane testowe
2. **Prognozowanie**
   1. **Autoregresja liniowa** najprostszy model
   2. **Naiwne metody prognozowania**
   3. **Modele sredniej wazonej** prognozy oblicza się na podstawie wazonej sredniej wcześniejszych wartości, jeśli wagi maleja geometrycznie, jest to model Holt-Wintersa. Podstawowa wersja tego modelu nie uwzglednia ani trendu, ani sezonowości, a wiec nie u mnie
   4. **Modele ARIMA – (**autoregressive integrated moving averages) zakladaja niezerowa autokorelacje skladowe przypadkowej, co często prowadzi do dokładniejszych prognoz. Jest polaczeniem modelu autoregresyjnego(AR) z modelem sredniej kroczącej (MA). Model Arima ma szanse być dokładny jeżeli w szeregu czasowym wystepuja silne autokorelacje lub czastkowe autokorelacje. Sa zdefiniowane dla stacjonarnych szeregow czasowych (czyli szeregow o stalych w czasie: sredniej, wariancji, autokorelacji) Jeżeli dane zrodlowe nie spelniaja tego warunku, konieczne jest zroznicowaie szeregu.
      1. **Różnicowanie 1, 2, 3, rzedu**
      2. **Parametry modelu** (ARIMA(p,d,q): d=odpowiada za rząd roznicowaia ; p=parametr autoregresji / q = parametr sredniej kroczącej – tu pomocne sa wspolczynniki autokorelacji i czastkowej autokorelacji; **parametr p** – liczbba wartości zmiennej uwzgledniona przy prognozowaniu jej przyszłych wartości (np. arima (2, x, y) – wartość sprzedaży w marcu na podstawie wynikow z lutego i marca), **parametr q** wpływ coraz starszych zaburzen (składowej przypadkowej poprzednich wartości w serii) na biezaca wartość zmiennej.
      3. **Model MA(2) =ARIMA (0,0,2), AR(2) = arima(2,0,0)**
      4. **Sprawdz wartosci ACF i PACF zróżnicowanego szeregu czasowego i porównaj – wybierz model**
      5. Sprawdz auto.arima (dane uzupełnione o prognozy modelu ARIMA)
   5. **Modele nieliniowe –** zalea maja wieksza stabilność i dokladnosc dlugoterminowa, ale prognozy krótkoterminowe sa mniej dokładne. Model NNAR(p,k) jeśli model nie miał składowej okresowej, a jeśli miał to NNAR(p,P,k)[m] = p=parametr autoregresji/P uzyta w formule autoregresja składowa okreswa, k – liczba ukrytych wezlow sieci, m – okresowość szeregu czasowego (paramtr size) // przy uzyciu lambda można przeksztalcic metoda Boxa-**Coxa dane treningowe**
3. **Ocena modelu:** 
   1. **Skrocic szereg czasowy i wziąć go jako test**
   2. Miary bledu: można obliczyć wektor bledu
   3. Można zastosować miary: najprostsza to sredni blad bezwzględny MAE (mean absolute error)

**Slownik:**

**Trend** – monotoniczny skladnik serii danych – rosnacy albo malejący, szereg czasowy może mieć tylko jeden rend. // dlugookresowa tendencja do jednokierunkowych zmian (wzrostu lub spadku) wartości, lub staly poziom który wystepuje gdy w szeregu nie ma trendu a wartości oscylują według stałego poziomu

**OKRES –** w ramach którego można wyroznic:

Wahania sezonowe (wahania wokół trendu lub stałego poziomu, powtarzające się w określonym czasie nieprzekraczającym jednego roku)

Wahania cykliczne (długookresowe wahania wokół trenud lub stałego poziomu)

Skladowa cykliczna – wiele szeregow czasowych wykazuje wahania okresowe wyraźnie. Warto zauwazyc ze parametr, który jest trendem w pewnym przedziale w innym przedziale czasu może okazac się zmienna okresowa // Sezonowość (druga po trendzie skladowa szeregow czasowych) – zaleznosc korelacyjna rzedu k miedzy i-tym elementem szeregu a (i-k)-tym elementem. Miara sezonowości jest autokorelacja. Można ja usunąć poprzez przekształcenie każdego i-tego elementu szeregu na jego roznice z (i-k)-tym elementem. Robi się to po to, żeby:

1. Przygotować dane treningowe na potrzeby modeli ARIMA aby uzyskac szregi stacjonarne (szeregi o stalych w czasie sredniej, wariancji i autokorelacji).
2. Analiza aby zanalizować ukryta nature szregu

**Szum** – pozostaje po odjęciu trendu i cykliczności. Zaswsze jest nieregularny, ale ma charakterystykę i niektóre algorytmy np. ARIMA obliczają na jego podstawie przyszle wartości

**Model** addytywny – sezonowe wahania sa stale w czasie

Model multiplikatywny – sezonowe wahania sa coraz większe

Autokorelacje = korelacja miedzy kolejnymi opóźnionymi wartościami tej samej zmiennej (ilustracja wykres funkcji lag2.plot). Mierzy się ja za pomocą miar autokorelacji ACF (uwzglendia wszystkie obserwacje badanego okresu, czyli umozliwia zbieranie np. wpływu dochodow narodowych brutto z ostatnich 5 lat na dochod tegoroczny) - biblioteka r funkcja Acf i Pacf z bliblioteki forecasti czastkowej autokorelacji PACF

Korelacja w czasie miedzy zmiennymi = polaczenie korelacji i autokorelacji

Sredia kroczaca - SMA, od ang. *simple moving average -* zwykła [średnia arytmetyczna](https://pl.wikipedia.org/wiki/%C5%9Arednia_arytmetyczna) wartości z ostatnich n okresów

Wazona srednia ruchoma - WMA, od ang. *weighted moving average.* przypisuje różne wagi danym z poszczególnych okresów

Wykładnicza średnia ruchoma (EMA, od ang. *exponential moving average*) jest odmianą średniej ważonej, w której znaczenie coraz bardziej odległych w czasie okresów maleje w sposób wykładniczy

Szerokość okna wygładzenia – horizon?

Czy szereg czasowy jest stacjonarny? Sprawdza się za pomocą autokorelacji i autokorelacji częściowej box.test / tseries adf.test / kpss.test

Wartość p (po testach stacjonarności) to **graniczny poziom istotności** (prawdopodobieństwo uzyskania takiej jak zaobserwowana lub bardziej oddalonej od zera wartości pewnej statystyki (np. roznicy średnich) przy zalozeniu ze hipoteza zerowa jest spelniona. Im wartość p jest mniejsza tym bardziej prawdopodobna jest hipoteza alternatywna, w tym wypadku hipiteza, kora zakłada ze jest szereg stacjonarny, a wiec jego różnicowanie jest niepotrzebne.

**Forcasting ARIMA**

https://d.adroll.com/cm/aol/out?advertisable=J2JBYR2WMFF6RO2QUVOG4Chttps://d.adroll.com/cm/n/out?advertisable=J2JBYR2WMFF6RO2QUVOG4C

R packages forecast, tseries , ggplot2

https://d.adroll.com/cm/n/out?advertisable=J2JBYR2WMFF6RO2QUVOG4CARIMA models are a popular and flexible class of forecasting model that utilize historical information to make predictions. This type of model is a basic forecasting technique that can be used as a foundation for more complex models. In this tutorial, we walk through an example of examining time series for demand at a bike-sharing service, fitting an ARIMA model, and creating a basic forecast. We also provide a checklist for basic ARIMA modeling to be used as a loose guide

CHECKLIST

•Explaining seasonal patterns in sales

•Predicting the expected number of incoming or churning customers

•Estimating the effect of a newly launched product on number of sold units

•Detecting unusual events and estimating the magnitude of their effect

**ARIMA CHEATSHEET**

1.Examine your data

•Plot the data and examine its patterns and irregularities

•Clean up any outliers or missing values if needed

•tsclean() is a convenient method for outlier removal and inputting missing values

•Take a logarithm of a series to help stabilize a strong growth trend

2.Decompose your data

•Does the series appear to have trends or seasonality?

•Use decompose() or stl() to examine and possibly remove components of the series

3.Stationarity

•Is the series stationary?

•Use adf.test(), ACF, PACF plots to determine order of differencing needed

4.Autocorrelations and choosing model order •Choose order of the ARIMA by examining ACF and PACF plots

5.Fit an ARIMA model

6.Evaluate and iterate

•Check residuals, which should haven no patterns and be normally distributed

•If there are visible patterns or bias, plot ACF/PACF. Are any additional order parameters needed?

•Refit model if needed. Compare model errors and fit criteria such as AIC or BIC.

•Calculate forecast using the chosen model

**TRESC ZADANIA**

|  |
| --- |
| In R or Python prefered |
| **T0: Merge** table 1 and 2 in a meaningful manner |
| **T1:** Create a forecast for each of the brands with a horizon of 3 and 12 using two predictive models.  https://d.adroll.com/cm/n/out?advertisable=J2JBYR2WMFF6RO2QUVOG4C  Horizon h = 3 short term forecasting horizon  Horizon h = 12 long-term horizon  The **forecast horizon** is the length of time into the future for which [forecasts](https://ec.europa.eu/eurostat/statistics-explained/index.php?title=Glossary:Forecasting) are to be prepared. These generally vary from short-term forecasting horizons (less than three months) to long-term horizons (more than two years).  Forecasting: Producing a prediction about the value of a random variable y at  horizon h conditional on the information set in period t    Sometimes one want to predict the distribution so that the object  of interest is    The first model should be an auto-arima model and the second one can be chosen freely. |
| **T2:** Out of the two forecasts build an ensemble by your liking. Essentially both forecasts need to be used to come to **one single forecast.**  Why does it matter what you will do with a forecast?  There may be biased forecasts with small MSE and unbiased forecasts with large MSE. Two different users may disagree which forecast is best for them. There is a large decision theoretic literature linking utility functions and optimality of forecast |
| **T3:** Go to quandl and identify a fitting external factor and present a meaningful presentation of the relationship and its effect on sales (e.g: https://www.quandl.com/search?filters=%5B%22Metals%22%5D&query=GDP). Hint: Think of **T1** |