Seminar 3

Time series plotting, stationarity, differencing and decomposition

Obs. This seminar contains exercises related to the topics discussed in Lecture 4 and 5: time series plotting, stationarity, differencing and part of decomposition. While the questions sometimes mention R and/or specific functions from R, you can try to solve the requirements (or parts of it) in other programming languages as well.

1. Read the content of "Earthquake.csv", which contains data about the total number of earthquakes in the world for a year, which had a magnitude over 7. There is data for 99 years, but we do not have the exact years, just the sequence of values. (File was taken from https://online.stat.psu.edu/stat510/book/export/html/661)

a) Transform the data into a tsibble. (The file contains just a sequence of numbers, to transform it into a tsibble, you need an index column. First you need to transform it into a tibble, then you can artificially create a sequence of years, with *seq(1901, 1999)* and add this as a column – use *mutate* – and finally transform it into a tsibble)
b) Autoplot the data.
c) According to what you see on the plot, do you think the data has a trend or does it seem to be stationary?
d) Plot the autocorrelation of the data for at least 16 lags. What do the autocorrelations tell you about the data? Is it stationary?
e) Use the KPSS test to check if the data is stationary.
f) Use the ADF test to check if the data is stationary. Do the two tests agree? If not, which one do you tend to agree with?
a) Using the *unitroot_ndiffs* and *unitroot_nsdiff* functions, check how many times the time series should be differenced to be stationary. If needed, difference the data and check again the autocorrelation.
g) Use STL decomposition to decompose the data. What do you observe?
h) A decomposition is good if the remainder is white noise. Get the remainder from the decomposition and plot its autocorrelations. Is it white noise?

3. Read the content of the "Temperature.csv", which contains hourly data about the temperature for more than 9 months (data was taken from: https://www.kaggle.com/datasets/vitthalmadane/ts-temp-1/data)

a) Transform it into a tsibble.
b) Autoplot it.
c) Create seasonal plots with different periods ("day", "week", "month"). Do you think there is a seasonal component in the data?
d) Take only the data for the first week (168 observations) and create a seasonal plot for them with daily period. Does it seem like you have daily seasonality in data?

e) Take only the data for the first 10 weeks (1680 observations) and create a seasonal plot for them with weekly period. Does it seem like you have weekly seasonality in data?
f) Plot the autocorrelation of the data for at least 16 lags. What do the autocorrelations tell you about the data? Does it have seasons? Does it have a trend?
g) Use the KPSS test to check if the data is stationary.
h) Use the ADF test to check if the data is stationary. Do the two tests agree? If not, which one do you tend to agree with?
i) Using the *unitroot_ndiffs* and *unitroot_nsdiff* functions, check how many times the time series should be differenced to be stationary. If needed, difference the data and check again the autocorrelation.
j) Use STL decomposition to decompose the original data. What do you observe?
k) Plot the autocorrelation of the remainder from the decomposition. Does it show that this is a good decomposition?

4. Consider the time series about monthly Vitamin prescription data from the PBD data. You can get this time series in R with the following code:

```
PBS |>
  filter(ATC2_desc == "VITAMINS") |>
  select(Month, Concession, Type, Cost) |>
  summarize(TotalCost = sum(Cost)) |>
  mutate(Cost = TotalCost / 1000000) -> vitamins
```

a) Autoplot it.
b) Do you think that it has trend or seasons?
c) Create seasonal plots. Do you think there is a seasonal component in the data?
d) Create a seasonal subseries plot.
e) Plot the autocorrelation in data (consider around 40 lags). Does it confirm your conclusions about season and trend from the previous questions?
f) Use the KPSS test to check if the data is stationary.
g) Use the ADF test to check if the data is stationary. Do the two tests agree? If not, which one do you tend to agree with?
h) Using the *unitroot_ndiffs* and *unitroot_nsdiff* functions, check how many times the time series should be differenced to be stationary.
i) Difference the data once. Plot the differenced data and its autocorrelation.
j) Do a seasonal difference. Plot the differenced data and its autocorrelation.
k) Why do the two differenced time series have a different number of observations?
l) Compare the two differenced data. Which seems more "stationary"?
m) Do a seasonal differencing followed by a regular one. Plot the differenced data and its autocorrelation. Does it seem better than the autocorrelation of the once differenced data?
n) Do an STL decomposition on the original data. Look at the seasonal component, how it changes over time.
o) Plot the autocorrelation of the remainder. Does it seem to be white noise?
p) Do an STL decomposition of the twice differenced data. What do you see? Why is it a bad idea to decompose differenced data?