Seminar 2

Working with data in R and Time series plotting

Obs. This seminar contains exercises related to working with data in R (including visualizations) and to plotting different time series and identifying whether they contain trends, cycles or seasons. 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 the "Sunspots.csv" file and transform it into a tsibble (use the *as_tsibble)* function. File was taken from https://www.kaggle.com/datasets/robervalt/sunspots/.

a) Rename the "Monthly Mean Total Sunspot Number" column into "Sunspots". (**Hint**: when you need to use column names containing several words, you need to put them between ` `)
b) Drop the column containing the line numbers.
c) Check what happens if instead of using ` ` as suggested in the hint at point a), you use " " for changing the name of a column.
d) What is the frequency of your tsibble data? Does this seem correct to you?
e) Transform your Date column, so that it has a monthly frequency (**Hint:** use the *yearmonth* function)
f) Get the *summary* of the *Sunspots* attribute.
g) Autoplot your data
h) Look at the autoplot of the entire data. Does it have trends?
i) Look at the autoplot of the entire data. Does it have seasons? If yes, determine their exact length.
j) Look at the autoplot of the entire data. Does it have cycles? If yes, try to determine an approximate average length for them.
k) Create a seasonal plot. Does it confirm your answer to the previous 3 points (about the existence of trend, cycle and season)?
l) Play around with the value of the *period* parameter for the seasonal plot. Can you find a value for which the plot shows some pattern? Values need to be multiples of 12 (need to be full years).
m) Save in a tsibble called *sunspotsSmall* those instances when the number of sunspots is between the minimum and the first quartile (you can hardcode the interval endpoint values returned by the *summary* function).
n) Autoplot this reduced tsibble. What do you observe on the plot?
o) How many observations are in *sunspotsSmall*?
p) Since there is implicit missing data in *sunspotsSmall* we cannot use it for seasonal plot. Replace these missing values using the *fill_gaps* function, with parameter *.full* set to FALSE.
q) Autoplot the resulting time series. What do you observe?

2. Read the data from the DailyDelhiClimate.csv file and transform it into a tsibble. File was taken from https://www.kaggle.com/datasets/sumanthvrao/daily-climate-time-series-data.

a) What is the frequency of the data?
b) What are the dimensions (number of observations and number of attributes) of the data?
c) Is the climate data in a wide or in a long format? Transform it into the other format (and save it in a different variable name), using the *pivot_longer* or *pivot_wider* function.
d) Autoplot the long format. What do you notice?
e) Let's plot the long format, but on 4 different subplots. For this you need to use *ggplot()*, *geom_line* and *facet_wrap* (facets are the feature which creates subplots based on the value of an attribute). What do you observe on the plot? (Compared to the results of point d).
   - For ggplot or geom_line you need to pass as parameter the data, and the aesthetic mapping (in our case what to put on axis x and what to put on axis y)
   - For facet_wrap you need to pass as parameter the following: *vars(NAME),* where NAME is the name of the attribute based on which you want to split into subplots
f) Add another parameter to *facet_wrap:* scales = "free_y". What changes?
g) Let's go back to our original data. Let's autoplot the *humidity* time series. Does it have trend, season and/or cycle? If it has season, try to determine the exact length; if it has cycles, try to determine an average length.
h) Create a seasonal plot of *humidity*, using a year as a period. Does this confirm your answer from the previous point about the existence of trend and/or season?
i) What if you use a month as a period? Do you have trend and/or season?
j) Repeat the previous 3 points for the *meantemp* and *wind_speed* time series.
k) Autoplot the *meanpressure* time series. What do you observe?
l) Autoplot only the first 1000 observations from *meanpressure* (you can use function *head).* Does it look different from the plot of the entire time series?
m) Get the summary of the *meanpressure* time series. What is the minimum and the maximum value?
n) Consider the following paragraph (taken from https://barometricpressure.today/cities/new-delhi-in). How does that relate to what you see in your data?
   In New Delhi, India, the barometric pressure varies throughout the year due to the changing seasons. During the summer months, from April to June, the pressure tends to be lower, ranging between 990 to 1008 millibars. In contrast, the pressure increases during the winter, from November to February, reaching values of around 1012 to 1026 millibars.
o) Let's filter only those observations which contain *meanpressure* between 990 and 1026 millibars (save them in a new tsibble). How many observations do we have?
p) Autoplot the *meanpressure* from this new tsibble. Does it have a trend or season?
q) Try to create a seasonal plot for the *meanpressure* of this new tsibble. What is the problem?
r) Convert implicit missing values into explicit ones using the *fill_gaps* function.
s) Create a seasonal plot. Do you have a trend and/or season?
t) Let's look at the outliers in the *meanpressure* data (observations with value less than 990 or greater than 1026). Save them in a separate tsibble. Are there values that are close to the regular ones, which might not actually be outliers?