This compulsory assignment consists of 4 exercises. Exercises must be solved in groups (assigned in Canvas) and uploaded to Canvas by the submission deadline. Please provide a single `.html` or `.pdf`-file, entitled `dat320_comp1_groupX.html` (or `.pdf`), where `X` should be replaced by your group number (1-16). The file should be structured into sections (one section per exercise) and subsections (one subsection per task). The usage of R markdown (R package `knitr`) is strongly recommended - you find a sample file in the folder "R markdown example". Further, check out this short introduction video: `https://rmarkdown.rstudio.com/lesson-1.html`.

Exercise 1 (Data processing with R)

The *ozone* dataset contains ground measurements of wind speed, temperature, and other parameters related to the ozone level between 1998 and 2004. The dimension is 2536 x 73, where all the columns whose column names start with "T" represent temperatures averaged hourly ("T.0": average temperature 00:00-00:59, "T.1": average temperature 01:00-01:59, etc.), and those, whose column names start with "WSR" indicate the wind speed (same naming scheme). Other columns can be ignored for this exercise.

Material provided:

- `ozone.csv`: dataset

Tasks:

(a) *Load the ozone dataset from the `.csv`-file as a data frame in R. Clean the dataset by removing all columns except for "Date", "WSR.0"-"WSR.23" (hourly wind speed measurements) and "T.0"-"T.23" (hourly temperature measurements). Transform the column "Date" into a Date format. Plot variables "T.0" and "WSR.0" over all dates and compare them. Is there a trend or seasonality?*

(b) *Investigate the dataset for missing values. Note that some dates are missing entirely — in such cases, add new rows with the correct dates and `NA` values for the measurements (no hard-coding!). Finally, make sure that the data frame is sorted correctly (w.r.t. dates). Further, write an R-function to remove measurements on 29.02. in leap years and apply it to the data frame. You can check your code by dividing the number of rows by 365 (the result should be exactly 7).*

(c) *Reshape the dataset from wide to long (use `reshape` from the `stats` package - refer to the documentation of the `reshape` method). In the end, the data frame should have one column for dates, one for the time of day, one for wind speed, and one for temperature (see Tab. 1). The dimension of the resulting data frame should be 61320 x 4. Finally, create a new column named "DateTime" in POSIXct format, which contains the starting points of the observation intervals (e.g., for the very first entry, "1998-01-01 00:00:00").*

(d) *Compute and plot the time series of (a) yearly average temperatures (± standard deviation), and (b) yearly median wind speeds (along with minimum and maximum*

wind speed) (use `dplyr::group_by` *and* `dplyr::summarize`*). NA-values should be ignored.*

date	time	wind	temperature
...	...	...	...

Table 1: Data frame after reshaping.

Exercise 2 (Missing values)

For this exercise, we use daily average temperature data recorded in Ås between 01.01.2012 and 13.12.2021. The dataset contains three columns where the first represents the date ("Date"), the second contains the temperature ("T.full"), and the third is the temperature with missing values ("T.missing"). You will work with the column "T.missing" and investigate options for missing value imputation. The "T.full" time series represents the ground truth and will be used to evaluate imputations.

Material provided:

- `temperature.csv`: dataset

Tasks:

(a) *Load the dataset from the* `.csv`*-file as a data frame in R. Describe properties of the time series denoted as "T.missing" (exploratory analysis), focusing on missing values in particular. Characterize missing values as either (a) single missing points or (b) missing intervals. How long is the longest sequence of missing values? (use the* `imputeTS` *package.) Remove days of February 29 (you may use your coded function from Exercise 1).*

(b) *Replace the missing values with each of the following metrics:*

(i) global mean,

(ii) last observation carried forward (LOCF)

Plot all options of the imputed time series and describe the differences. Compute the MSE between your fit and the ground truth, which is given in column "T.full". Which imputation performs best? Why?

(c) *Use further pre-implemented options for missing value imputation from the* `imputTS` *package and try to minimize the MSE between your imputation and the ground truth. Try at least 3 other methods and tune the parameters (if any). What are the advantages and disadvantages of the suggested imputation strategies?*

(d) *Decompose the imputed time series (best imputation option from steps b and c), as well as the ground truth time series "T.full". For this purpose, first transform the time series into an* `ts` *object and specify the frequency. Try a short and a long seasonality*

window (e.g., `s.window = 5` and `s.window = "periodic"`). Plot and interpret the result of the STL. How do different imputations affect the decomposition?

(e) (optional) *As another imputation method, replace missing values by the mean over the same day across all years, e.g., if 02.01. of year 1 is missing, compute an average over the values measured on 02.01. in all other years to replace. Implement the imputation method yourself as an R function and compare the results to the methods used in previous tasks.*

Exercise 3 (Transformations)

The *covid* dataset contains daily numbers of new cases with COVID-19 per 1 million inhabitants for multiple countries in the world, according to official statistics. The dataset was collected between 01.01.2020 and 21.02.2022.

Material provided:

- `covid.csv`: dataset

Tasks:

(a) *Load the dataset from the `.csv`-file as a data frame in R, restrict it to the column "new_cases_per_million", and extract the countries with iso-codes ("SWE", "DNK", "NOR", "GBR", "ITA" and "IND") in the time frame from 16.03.2020 to 01.01.2022. Perform an exploratory data analysis and plot the time series. What do you see? Are the data complete?*

(b) *To get an even more detailed overview of the countries, compute and plot*

- *the pairwise correlations,*
- *the pairwise cross-correlation functions, and*
- *the autocorrelation functions*

for all countries. The correlation matrix can be visualized using `matrixplot` or as a heatmap.

(c) *Perform a principal component analysis (PCA) and plot the scores and loadings. What can be concluded from the PCA results regarding similarities and differences between countries?*

(d) *For example in Sweden, the COVID-19 infections were not registered on a daily basis. Hence, multiple values are set to 0. Smooth the time series of each country to resolve this issue. Optimize the type of smoothing and the kernel size parameter k, and repeat the PCA analysis. Compare with c).*

Exercise 4 (Autocorrelation and cross-correlation)

Auto- and cross-correlation are two major concepts in time series analysis. They are

particularly helpful for exploratory analysis and provide information about the dynamics and characteristics of a time series.

Tasks:

(a) *Let $(x_t)_{t\in T}$ be a time series with regular time index set $T = \{1, \ldots, t_{\max}\}$. Further, assume that an inverted version of the autocorrelation function, $\widetilde{acf} : \mathbb{N} \to [0, 1]$ is given as*

$$\widetilde{acf}(k) = \left(x_t, B^{-k}(x_t)\right),\ t \leq (t_{\max} - k).$$

Show that this inverted autocorrelation function is equal to the autocorrelation (as defined in the lecture), i.e.,

$$acf(k) = \widetilde{acf}(k),\ \forall k \in \mathbb{N}.$$

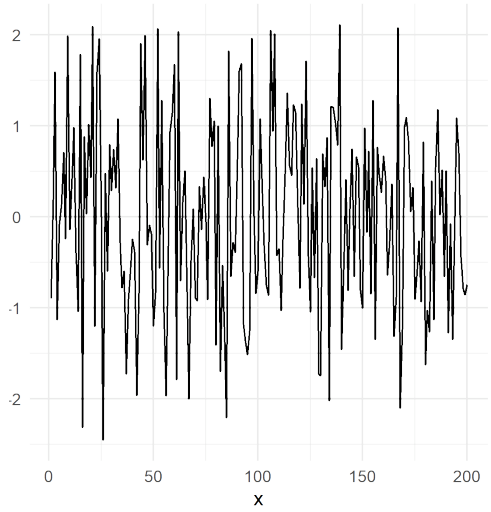
(b) *Does the same statement as in (a) hold when comparing the cross-correlation function $ccf(.)$ to an inverted cross-correlation function, given by $\widetilde{ccf} : \mathbb{Z} \to [0, 1]$,*

$$\widetilde{ccf}(k) = \left(x_t, B^{-k}(y_t)\right),\ t \leq (t_{\max} - k)?$$

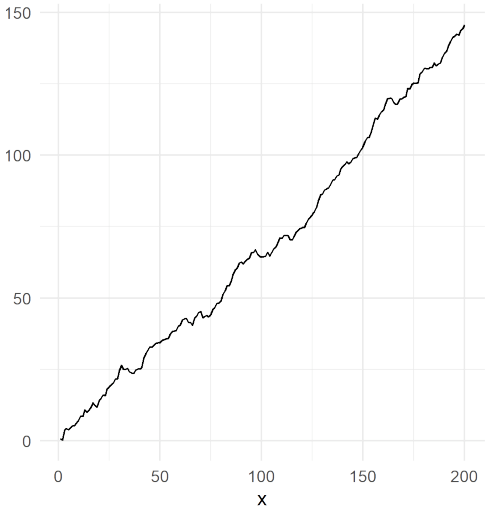
Justify your answer.

(c) *Look at the line plots of time series in Fig. 1, and the autocorrelation functions in Fig. 2. Assign the time series (a)-(d) to their correct autocorrelation functions (a)-(d)! Describe properties of the time series that can be identified in the autocorrelation functions.*

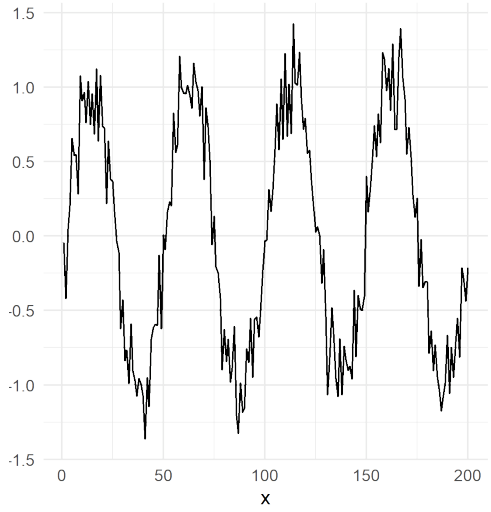
(d) *Fig. 3 shows two time series subject to a phase shift (shift along the time-axis). What can we say about the lag of the phase shift? What do their autocorrelation functions and their cross-correlation function look like? (Provide a description of function characteristics we can derive from the line plots, or a sketch of acf and ccf)*



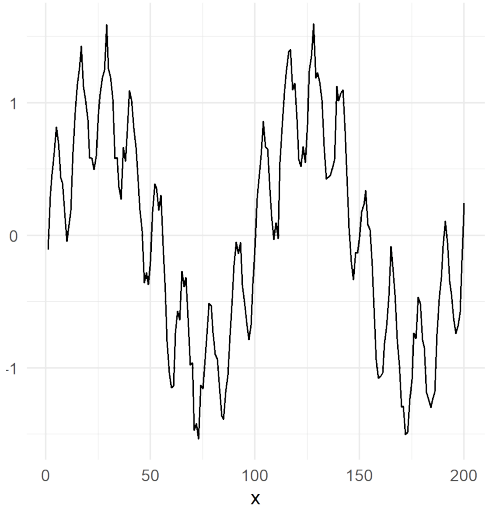
(a) Time series 1



(b) Time series 2

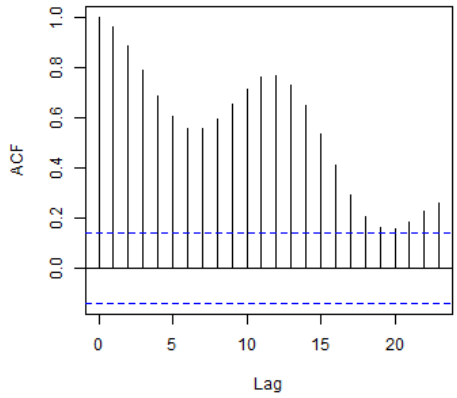


(c) Time series 3

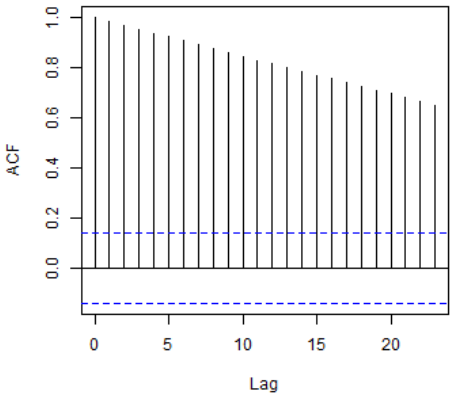


(d) Time series 4

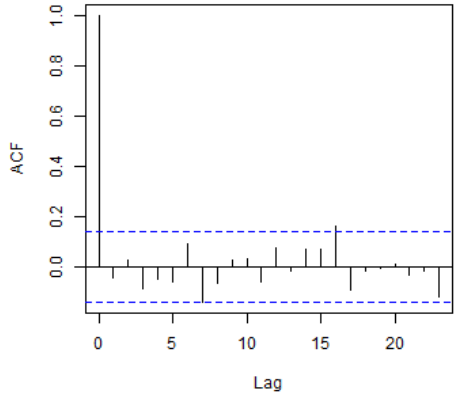
Figure 1: Time series.



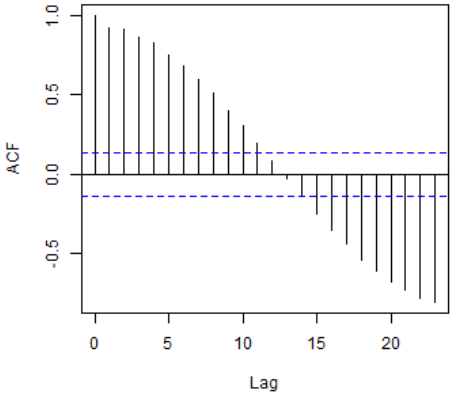
(a) ACF 1



(b) ACF 2

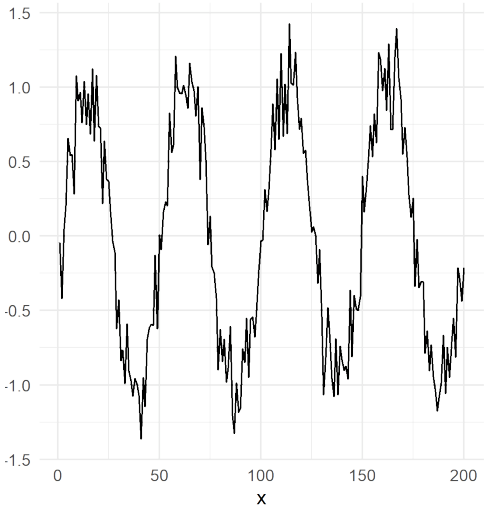


(c) ACF 3

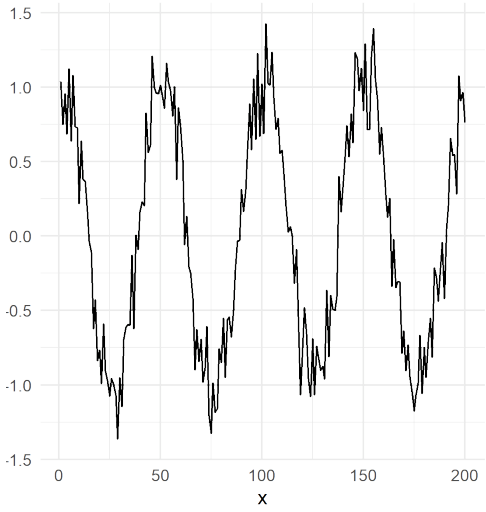


(d) ACF 4

Figure 2: Autocorrelation functions.



(a) Time series 3



(b) Time series 5

Figure 3: Time series with phase shift.