MA 5124 Financial Time Series Analysis and Forecasting

Chapter 1: Introduction to time series and forecasting
Lesson 1

Dr. Priyanga Talagala

23-06-2024

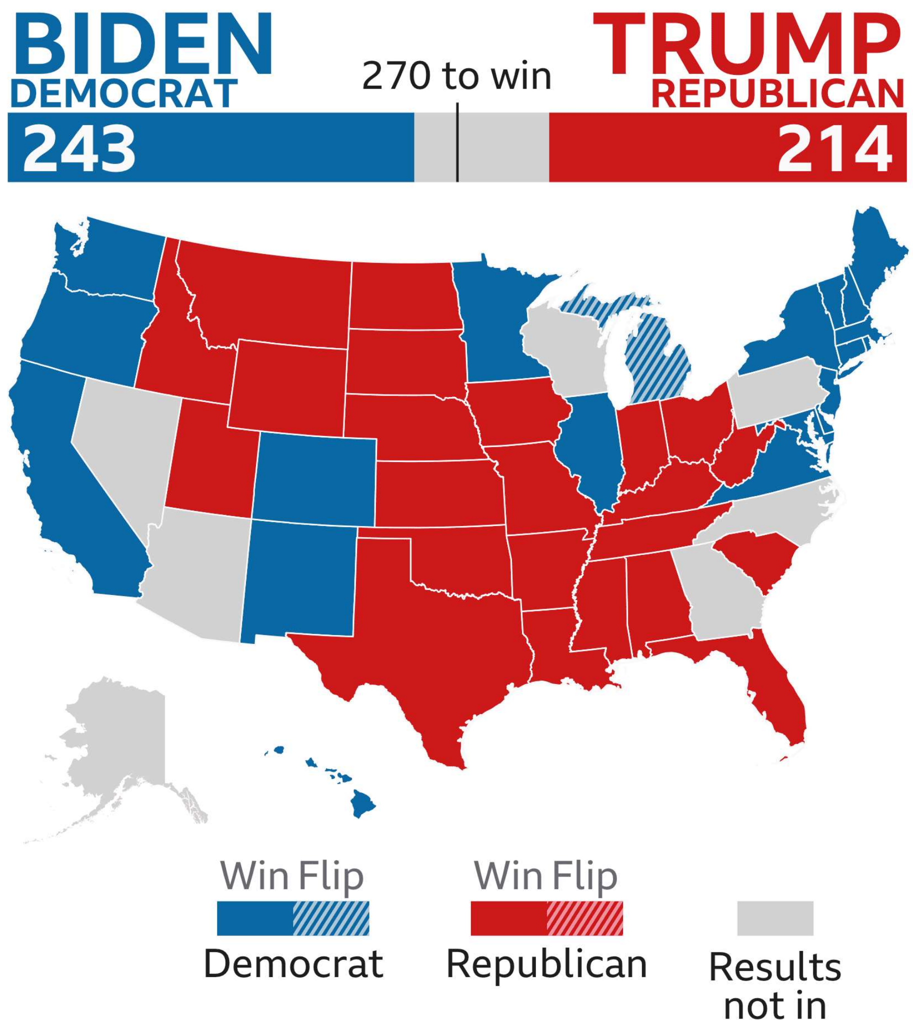
Types of Data

- Cross-sectional data
- Time series data
- Pooled data
- Panel data

1. Cross-sectional data

- A cross-sectional data set consists of a sample of individuals, households, countries or any other type of unit at **a specific point in time.**
- Sometimes, data across all units do not correspond to exactly the same time point.
- Example: A survey that collects data from questionnaire surveys of different units within a month.
- In this case, we can ignore the minor time differences in collection.

ID	Monthly Income (in LKR)
1	83000
2	150000
3	40000
4	65000



Source: NEP/Edison via Reuters, 07:47 GMT (02:47 EST)

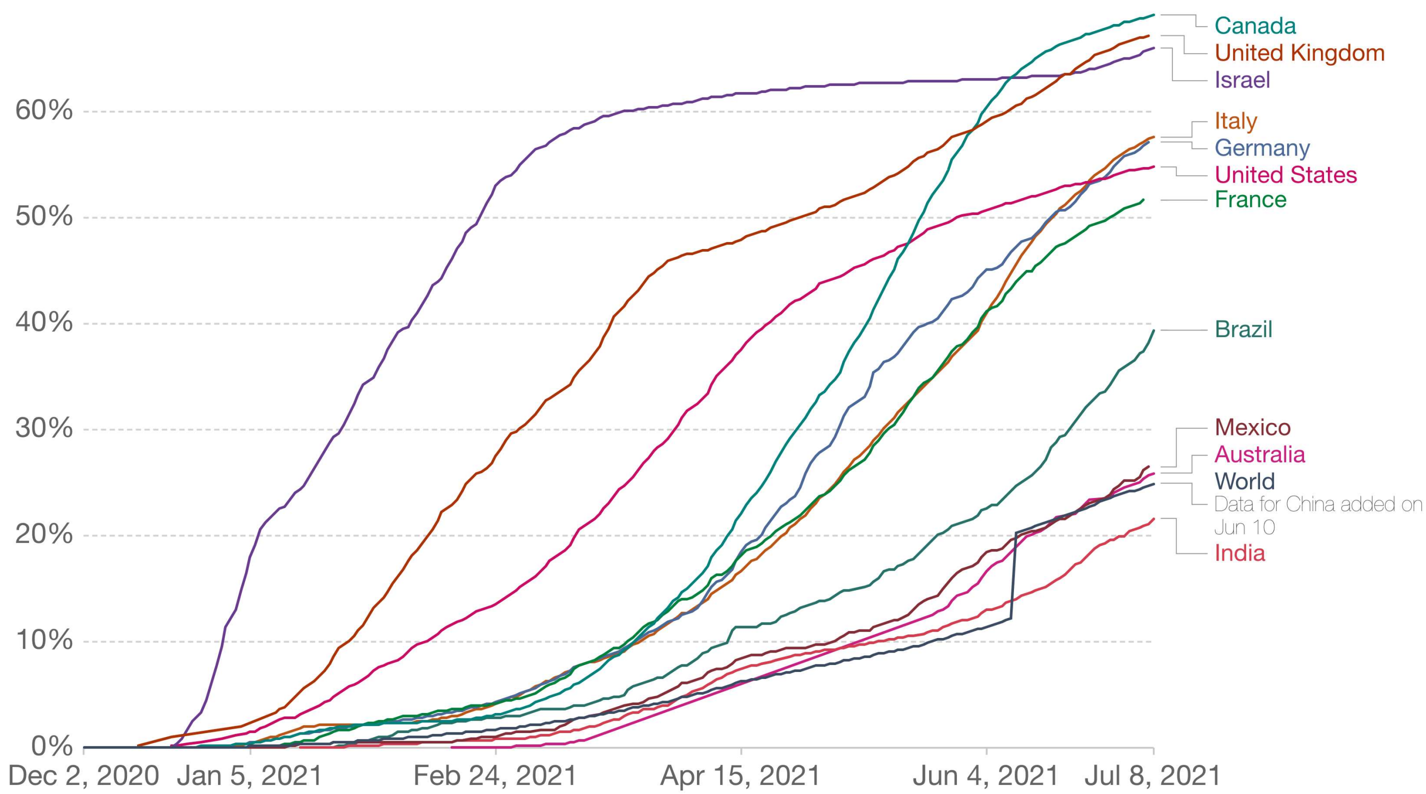
BBC

2. Time series data

- A time series is a sequence of observations taken **sequentially in time**.
- Time series data are arranged in chronological order and can have different time frequencies (eg: biannual, annual, quarterly, monthly, weekly, daily, hourly, etc.)
- Examples of time series data
  - Annual Google profits
  - Monthly rainfall
  - Weekly retail sales
  - Daily confirmed COVID-19 cases and deaths
  - Hourly electricity demand

Share of people who received at least one dose of COVID-19 vaccine

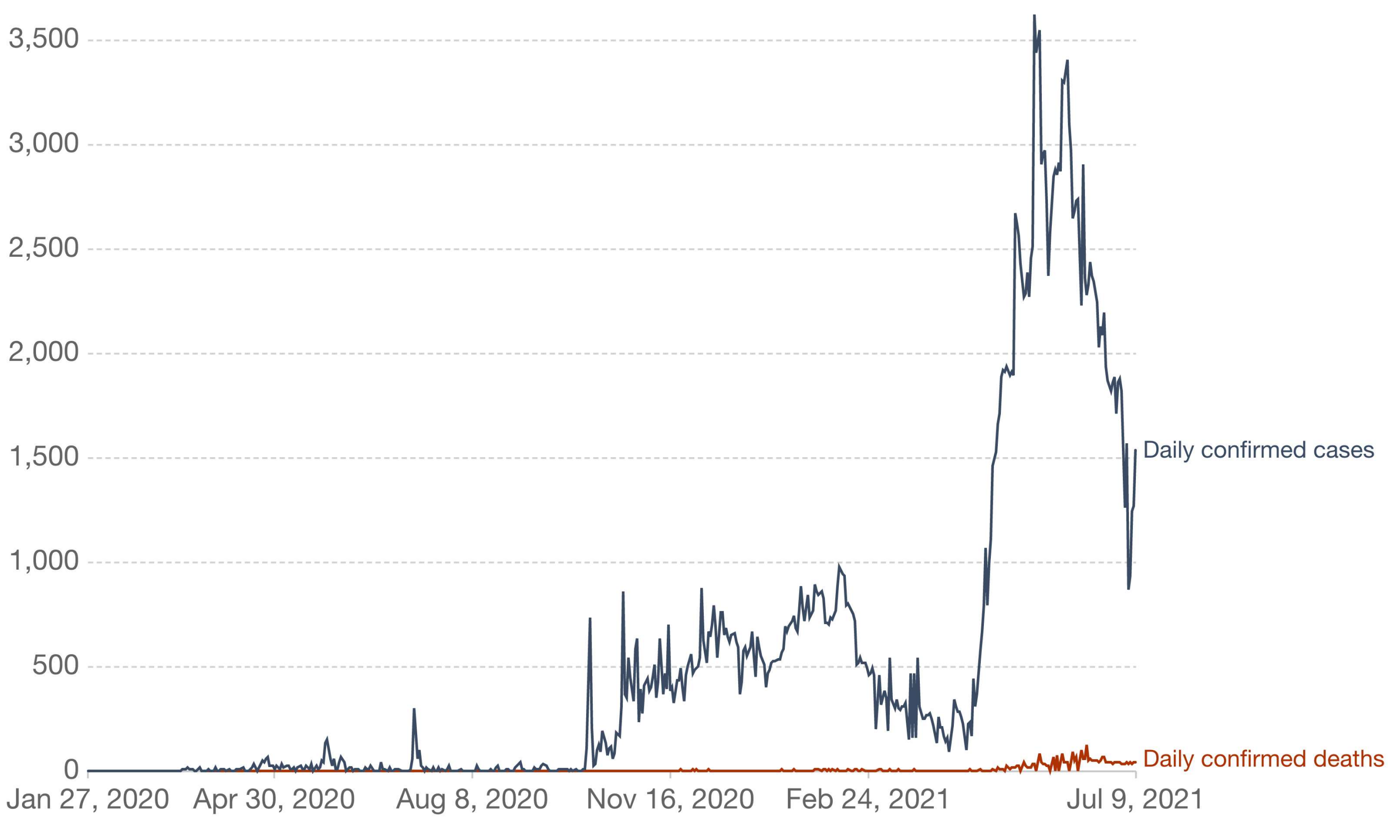
Share of the total population that received at least one vaccine dose. This may not equal the share that are fully vaccinated if the vaccine requires two doses. This data is only available for countries which report the breakdown of doses administered by first and second doses.



Source: Official data collated by Our World in Data

Daily confirmed COVID-19 cases and deaths, Sri Lanka

The confirmed counts shown here are lower than the total counts. The main reason for this is limited testing and challenges in the attribution of the cause of death.



Source: Johns Hopkins University CSSE COVID-19 Data – Last updated 10 July, 09:02 (London time) OurWorldInData.org/coronavirus • CC BY

3. Pooled data

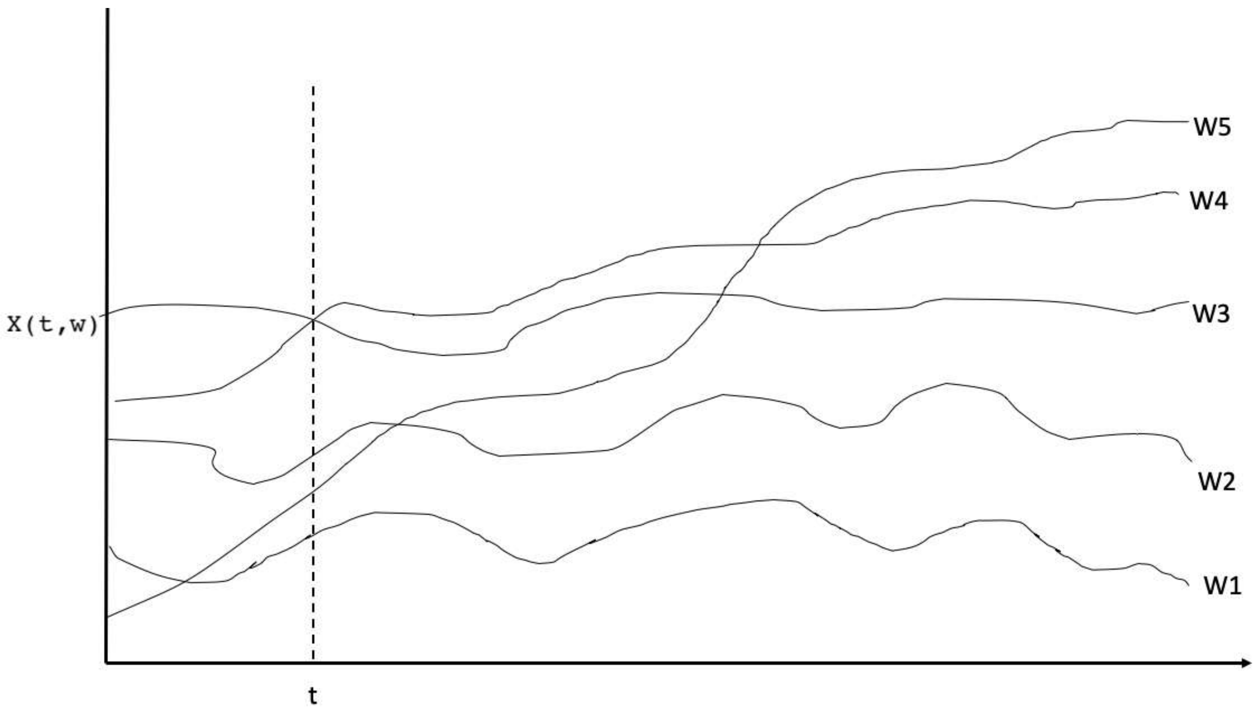
- Pooled data occur when we have a “time series of cross sections,” but the observations in each cross section do not necessarily refer to the same unit.

4. Panel data

- This is a **special type** of pooled data in which the samples of the same cross-sectional units observed over time.

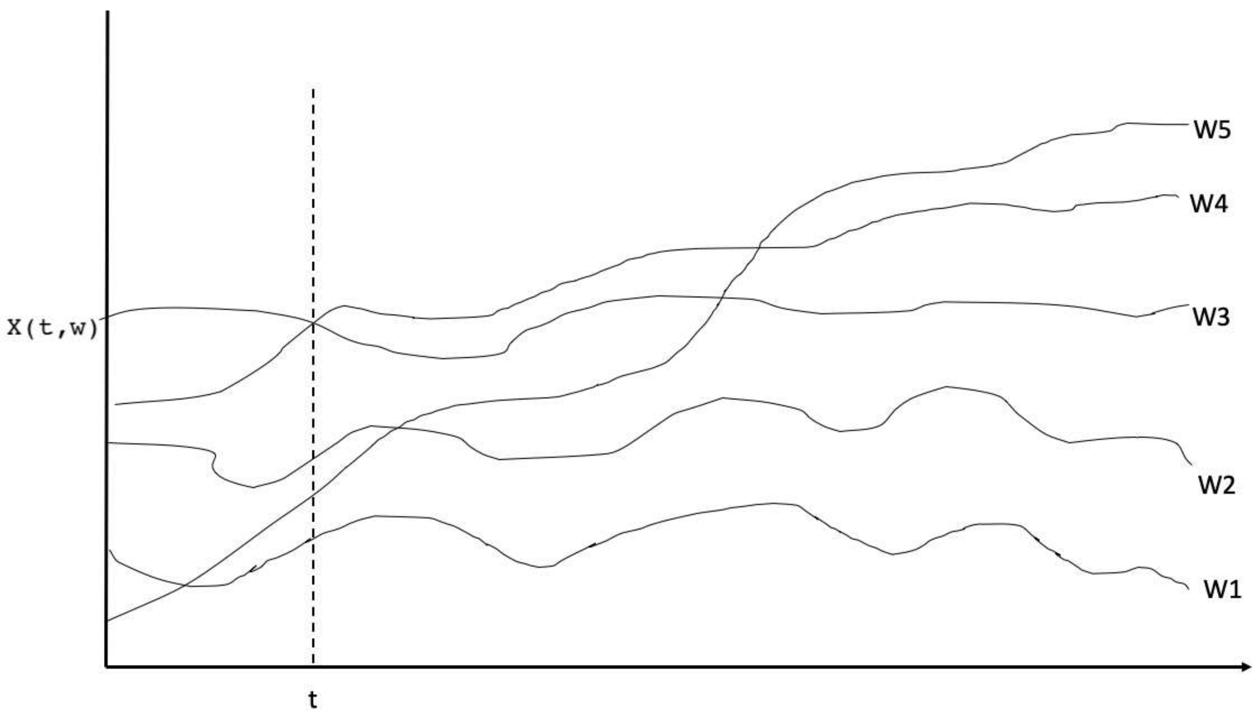
Stochastic Processes

A stochastic process is a family of indexed random variables $\{X(t,\omega); t \in T; \omega \in \Omega\}$ defined on a probability space $(\Omega, \beta, \mathbf{P})$ where T is an arbitrary set.



There are many ways of visualizing a stochastic process.

(i) For each choice of $t \in T$, $X(t,\omega)$ is a random variable.
(ii) For each choice of $\omega \in \Omega$, $X(t,\omega)$ is a function of t.
(iii) For each choice of ω and t, $X(t,\omega)$ is a number.
(iv) In general it is an ensemble (family) of functions $X(t,\omega)$ where t and w can take different possible values.



Time series data

- **The observed time series or time series to be analyzed is a particular realization of a stochastic process.**
- Anything that is observed sequentially over time is a time series.
- In this course, we will only consider time series that are observed at **regular intervals of time.**
- **Irregularly spaced time series** can also possible, but are beyond the scope of this course

Forecasting

- Forecasting is about **predicting the future** as accurately as possible, given all of the information available, including historical data and knowledge of any future events that might impact the forecasts.
- Forecasting is estimating how the sequence of observations will continue into the future.

Factors affecting forecastability

Something is easier to forecast if:

- we have a good understanding of the factors that contribute to it
- there is lots of data available;
- the forecasts cannot affect the thing we are trying to forecast.
- there is relatively low natural/unexplainable random variation.
- the future is somewhat similar to the past

Types of Methods

- **Qualitative** forecasts
  - Judgmental forecasting is the only option if no historical data (for new product, new market conditions), or if the data available are not relevant to the forecasts.
  - See fpp3 Chapter 6: https://otexts.com/fpp3/judgmental.html.
- **Quantitative** forecasts: can be applied
  - if numerical information about the past is available
  - if it is reasonable to assume that some aspects of the past patterns will continue into the future

Quantitative forecasts

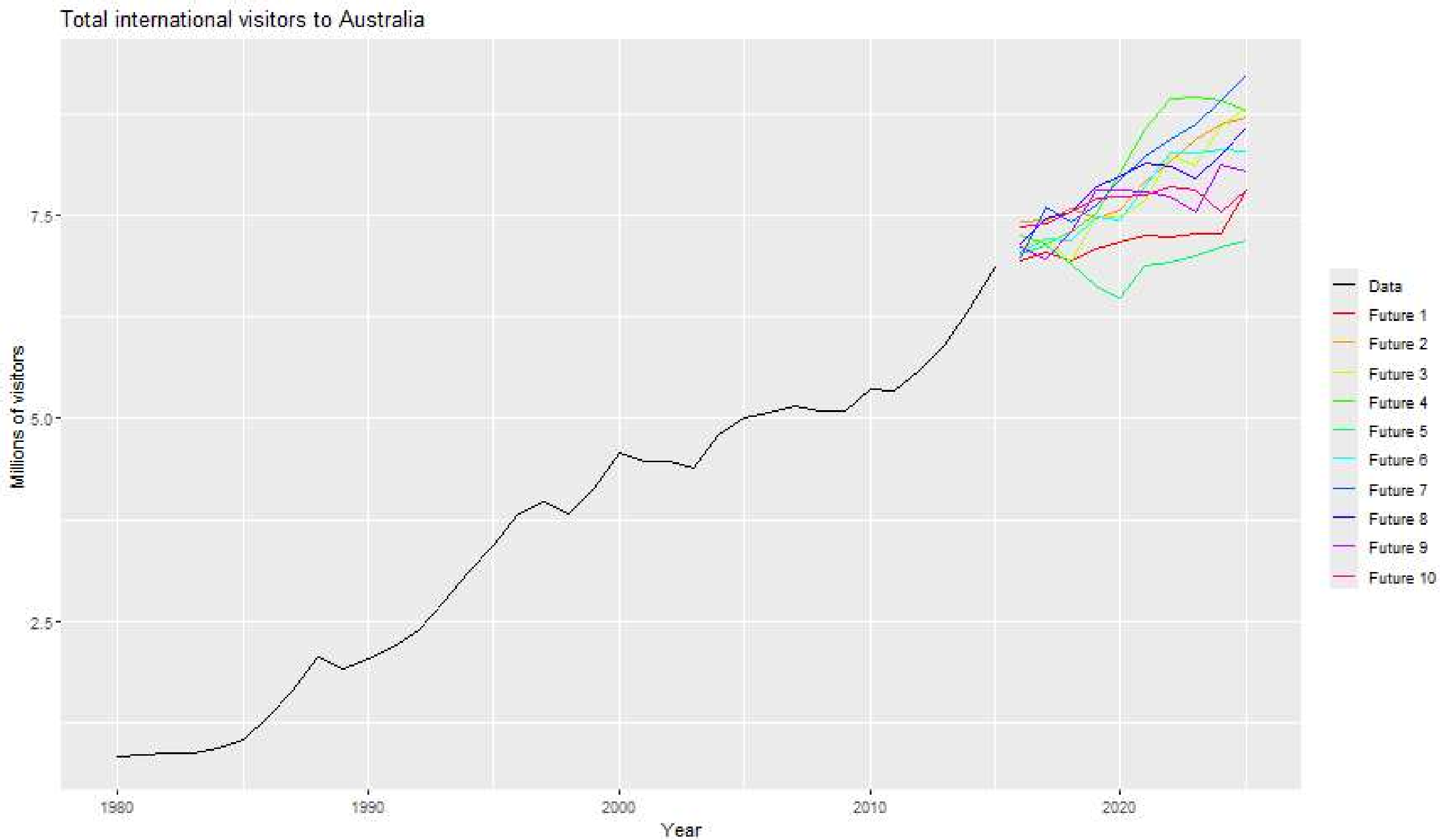
- Most quantitative forecasting problems use either
  - Time series data (collected at regular intervals over time).
  - Cross-sectional data (collected at a single point in time).

Basic steps in a forecasting task

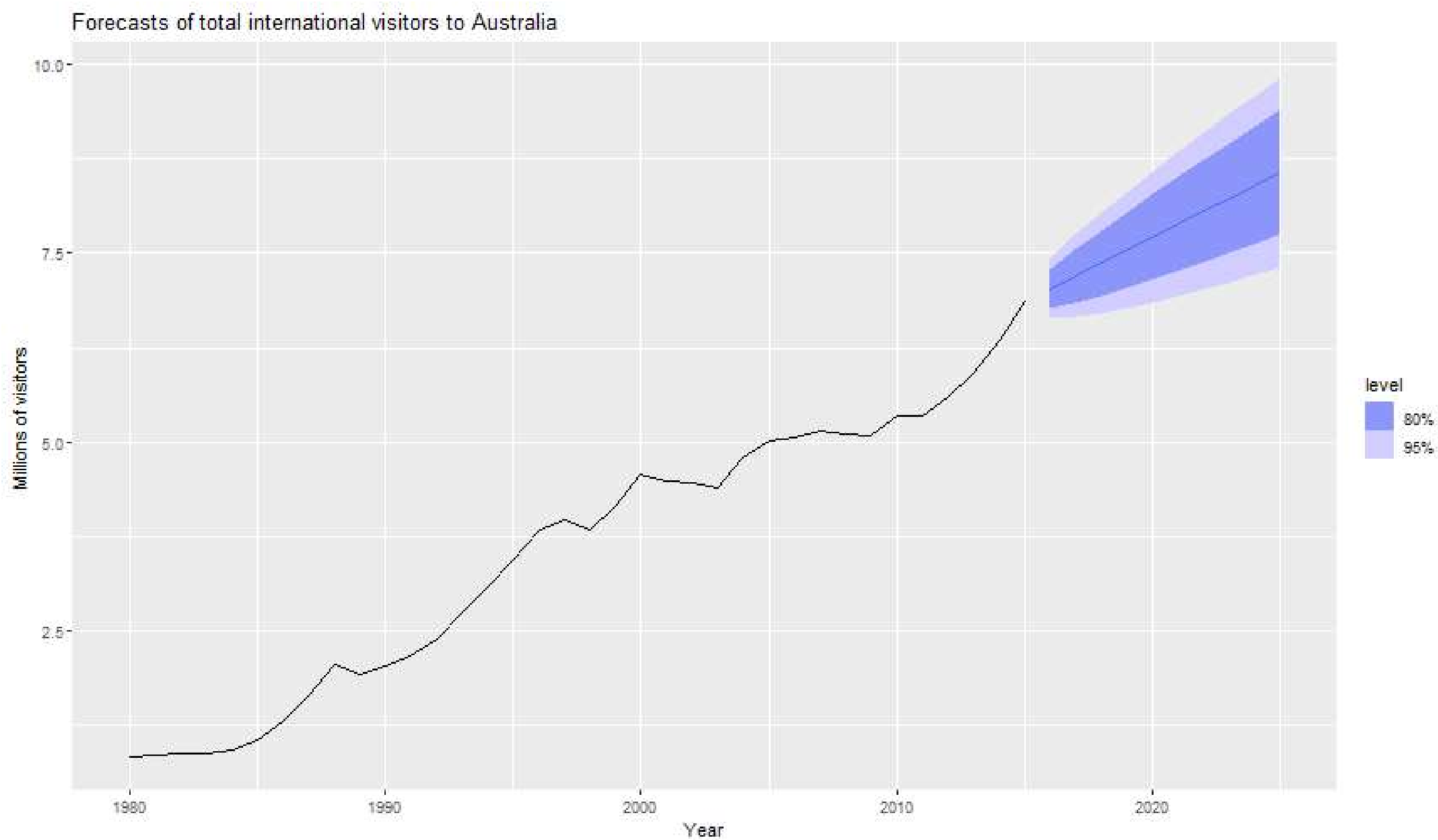
- Problem definition
- Collect data
- Preliminary (exploratory) analysis (data Visualization)
- Modelling
- Evaluate the fitted model

The statistical forecasting perspective

Sample futures



Forecast intervals



Statistical forecasting

- Thing to be forecast: a random variable, y_t.
- Forecast distribution: If $\mathcal{I}$ is all observations, then $y_t|\mathcal{I}$ means "the random variable y_t given what we know in $\mathcal{I}$".
- The **point forecast** is the mean (or median) of $y_t|\mathcal{I}$
- The **forecast variance** is $\text{var}[y_t|\mathcal{I}]$
- A prediction interval or **interval forecast** is a range of values of y_t with high probability.
- With time series, $y_{t|t-1} = y_t|\{y_1, y_2, \dots, y_{t-1}\}$.
- $\hat{y}_{T+h|T} = \text{E}[y_{T+h}|y_1, \dots, y_T]$ (an h-step forecast taking account of all observations up to time T).

Frequency of a time series: Seasonal periods

- **Frequency**: number of observation per natural time interval of measurement

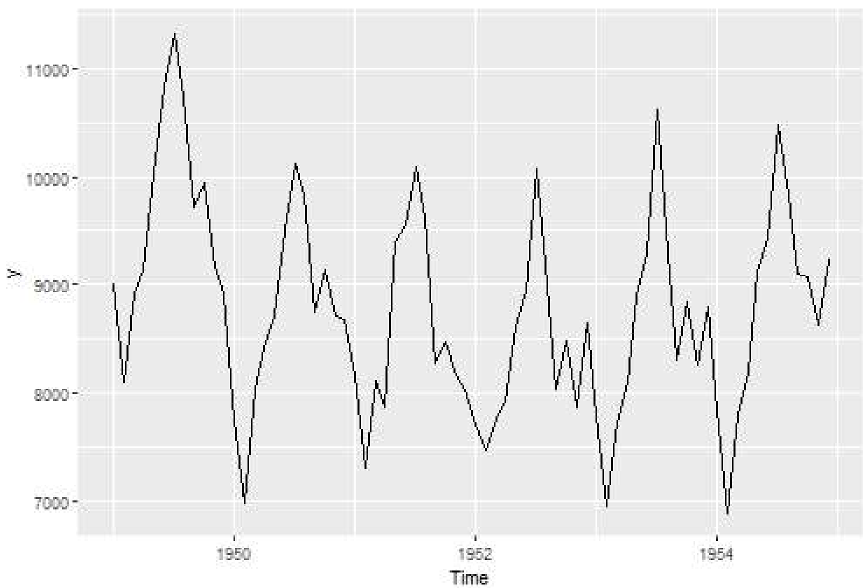
Data	Frequency
Annual	1
Quarterly	4
Monthly	12
Weekly	52 or 52.18

Frequency of a time series: Seasonal periods

- Multiple frequency setting

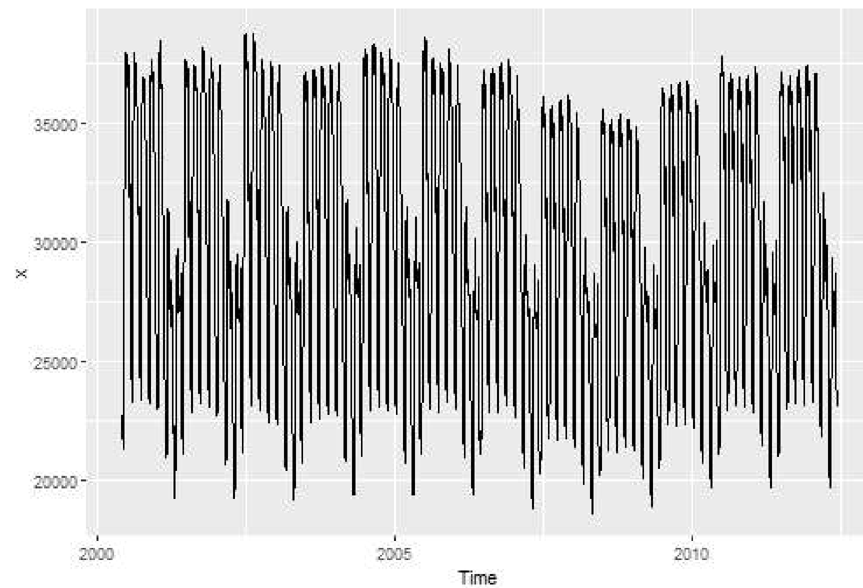
Data	Minute	Hour	Day	Week	Year
Daily				7	365.25
Hourly			24	168	8766
Half-Hourly			48	336	17532
Minutes		60	1440	10080	525960
Seconds	60	3600	86400	604800	31557600

Monthly time series



- Length of the series: 72
- Monthly seasonality

Half-hourly Time Series



- Length of the series: 4032
- Daily seasonality and weekly seasonality

Time series patterns

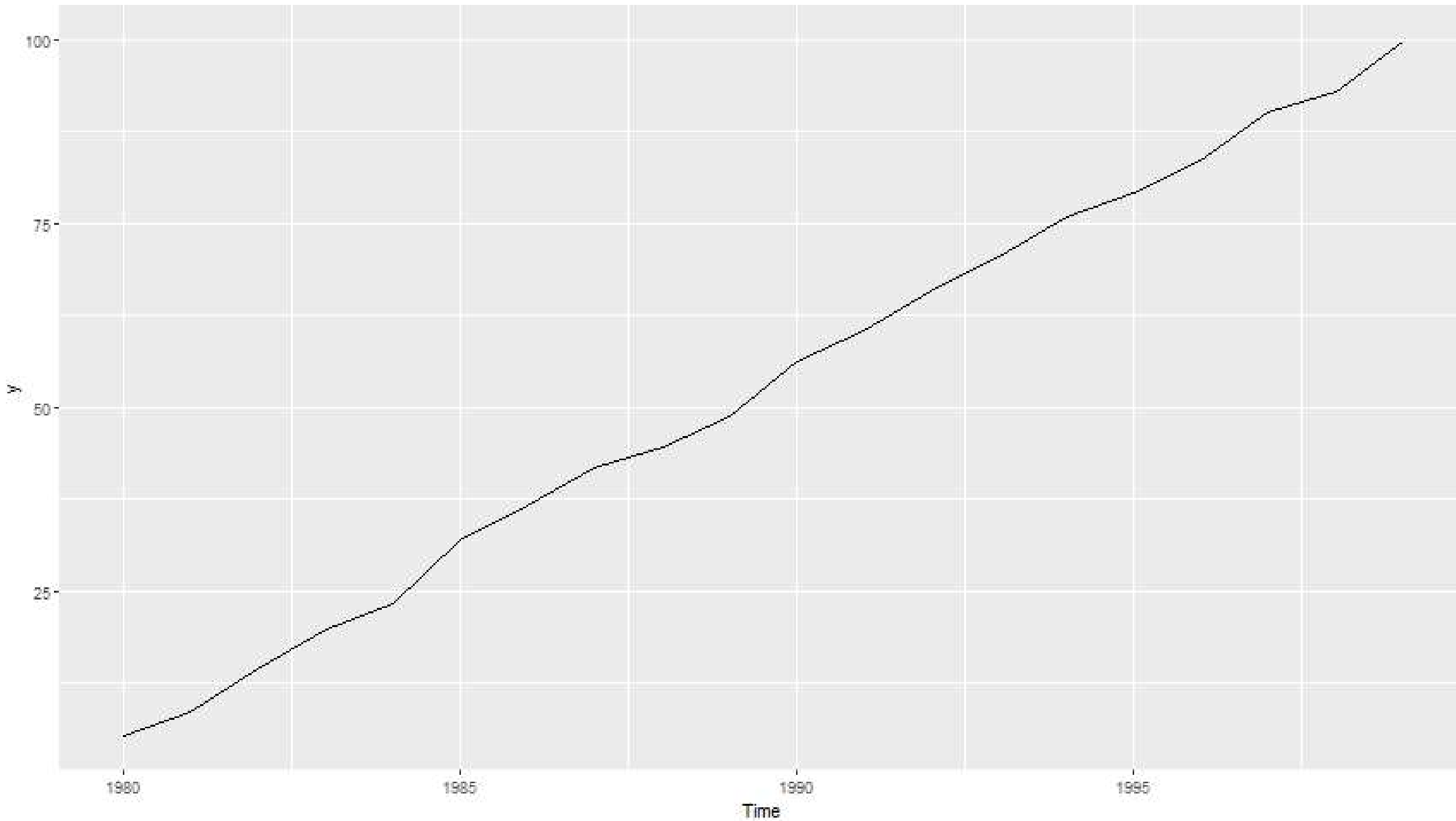
- **Trend** pattern exists when there is a long-term increase or decrease in the data.
- **Seasonal** pattern exists when a series is influenced by seasonal factors (e.g., the quarter of the year, the month, or day of the week).
- **Cyclic** pattern exists when data exhibit rises and falls that are not of fixed frequency (duration usually of at least 2 years).

Seasonal or cyclic?

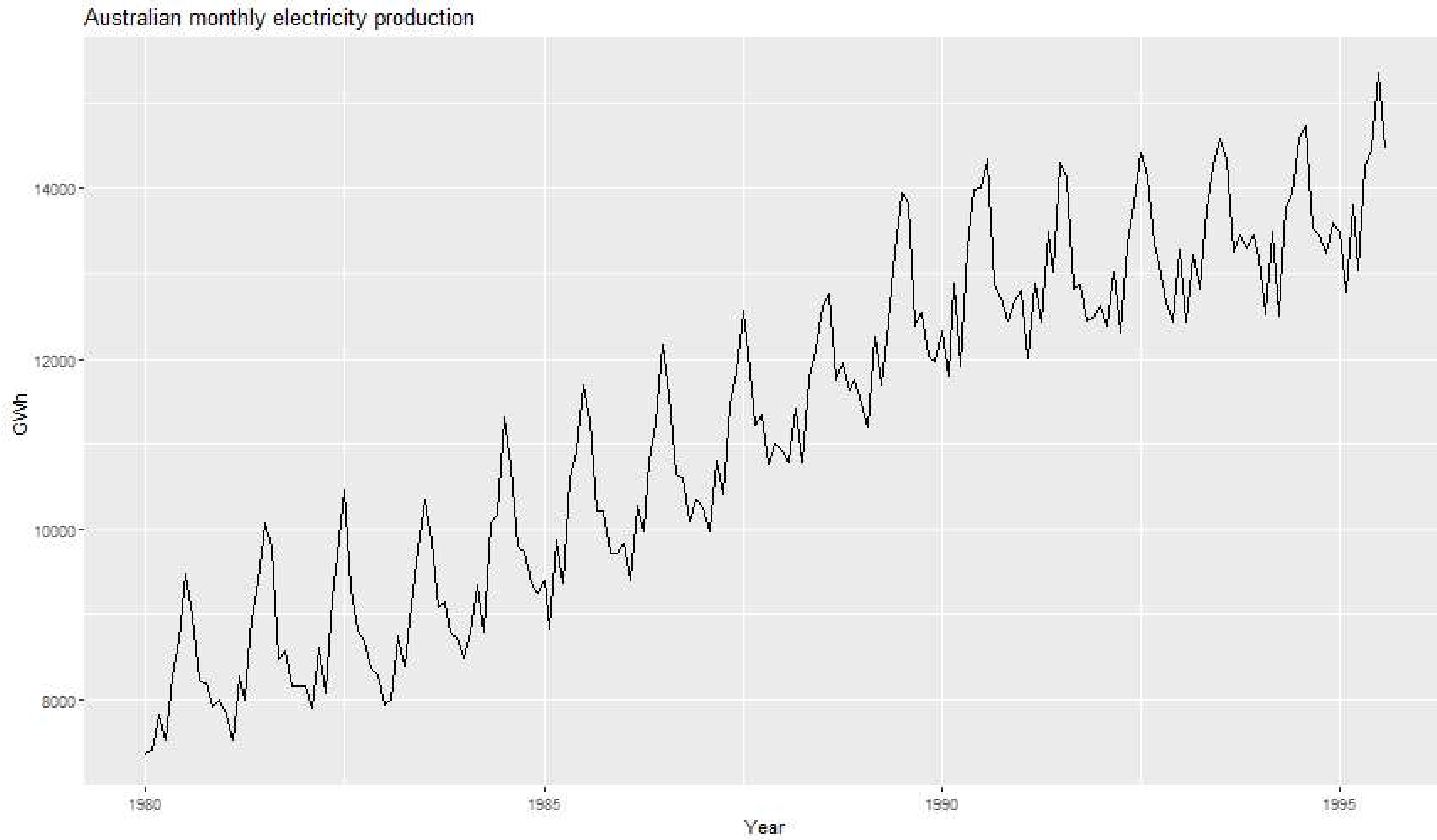
Differences between seasonal and cyclic patterns:

- seasonal pattern constant length; cyclic pattern variable length
- average length of cycle longer than length of seasonal pattern
- magnitude of cycle more variable than magnitude of seasonal pattern
- **The timing of peaks and troughs is predictable with seasonal data, but unpredictable in the long term with cyclic data.**

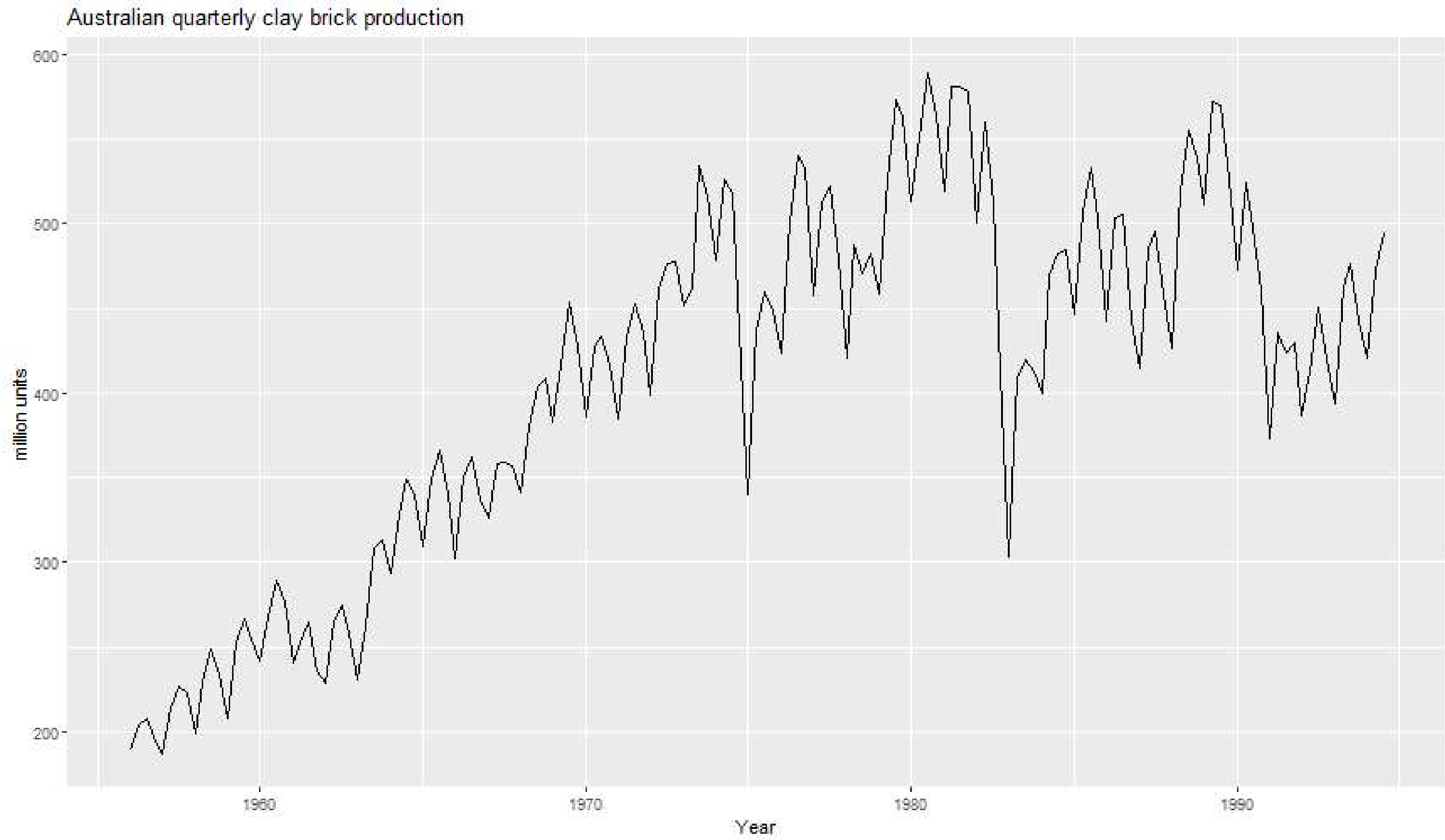
Time series patterns



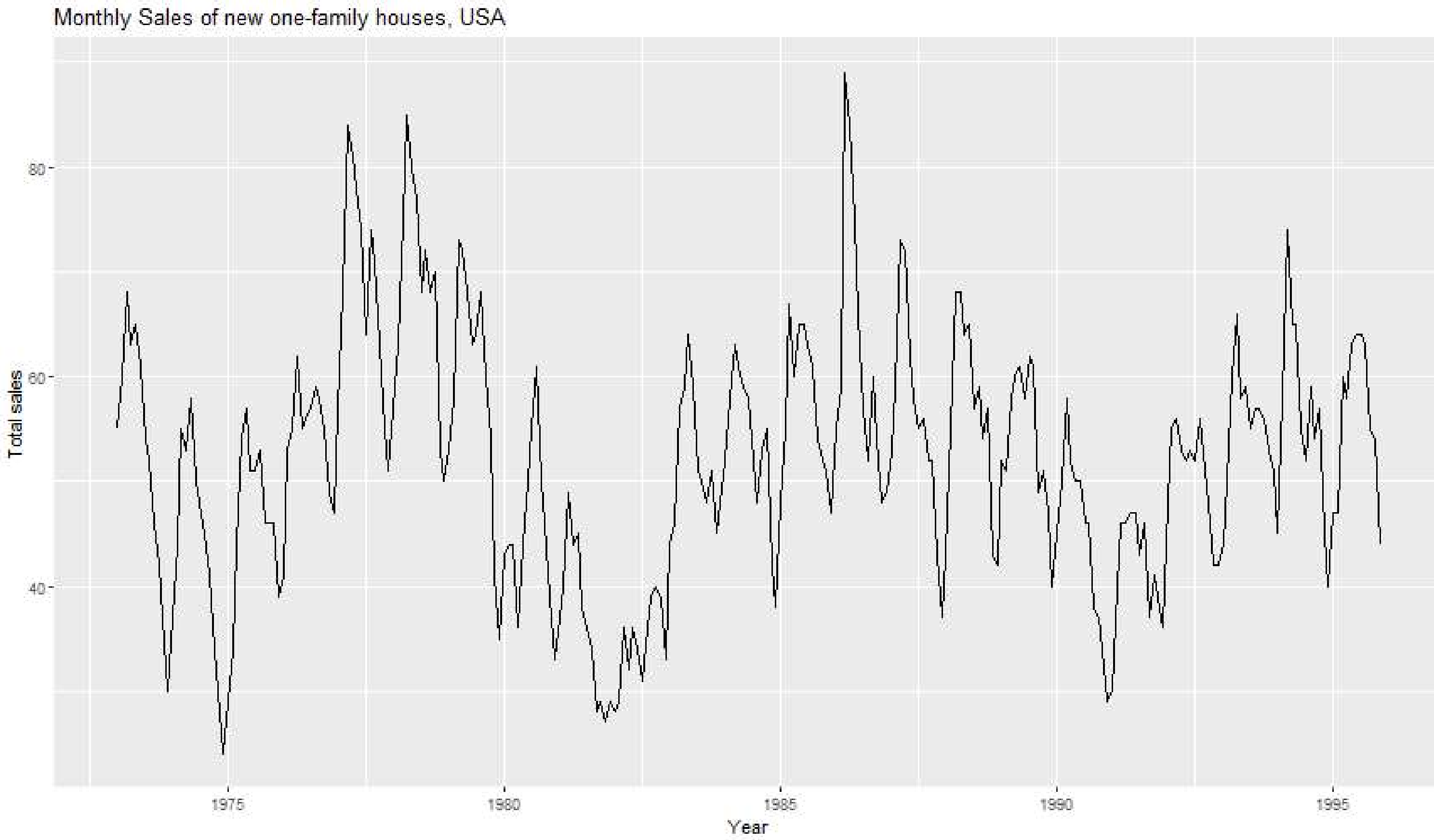
Time series patterns



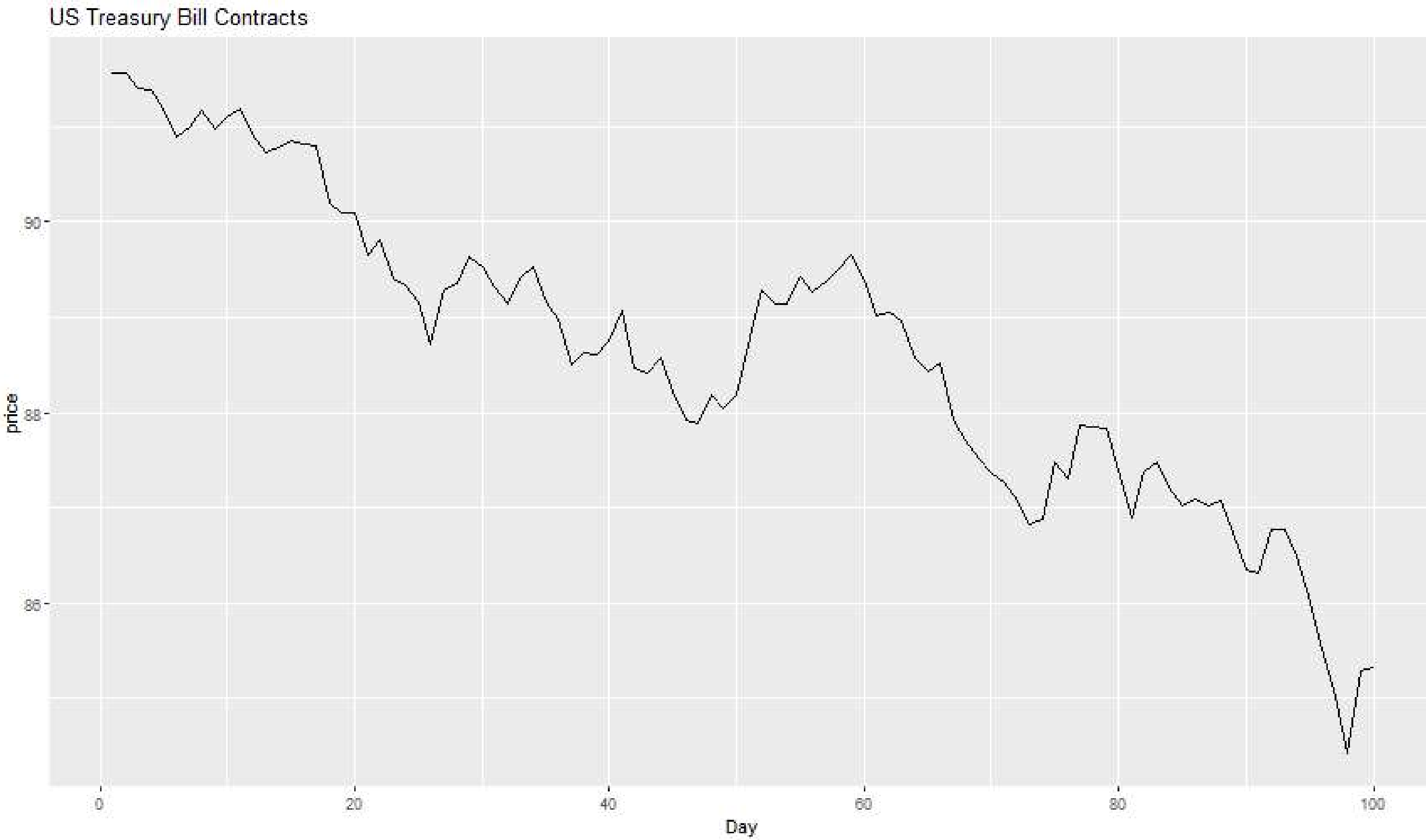
Time series patterns



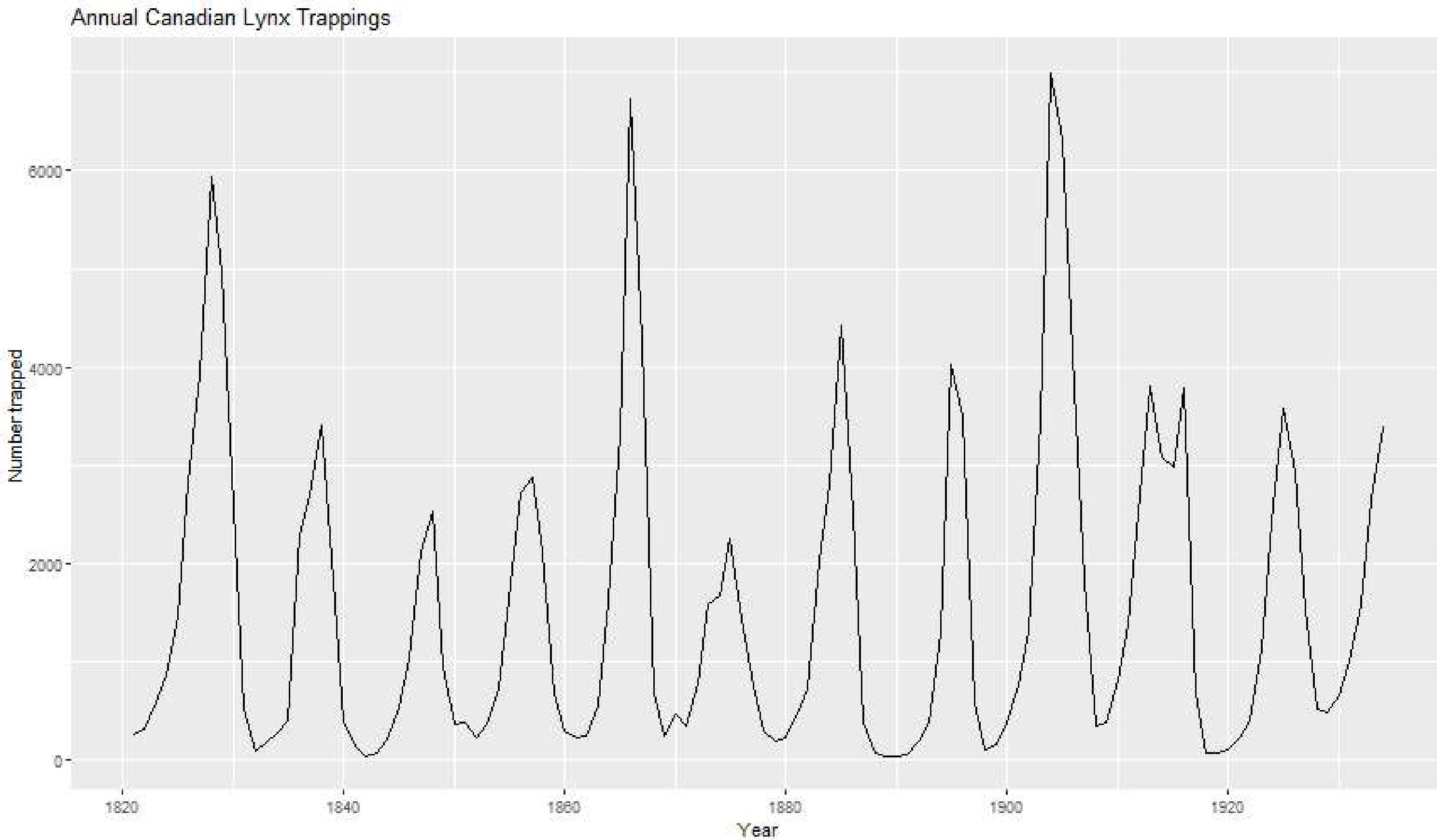
Time series patterns



Time series patterns



Time series patterns

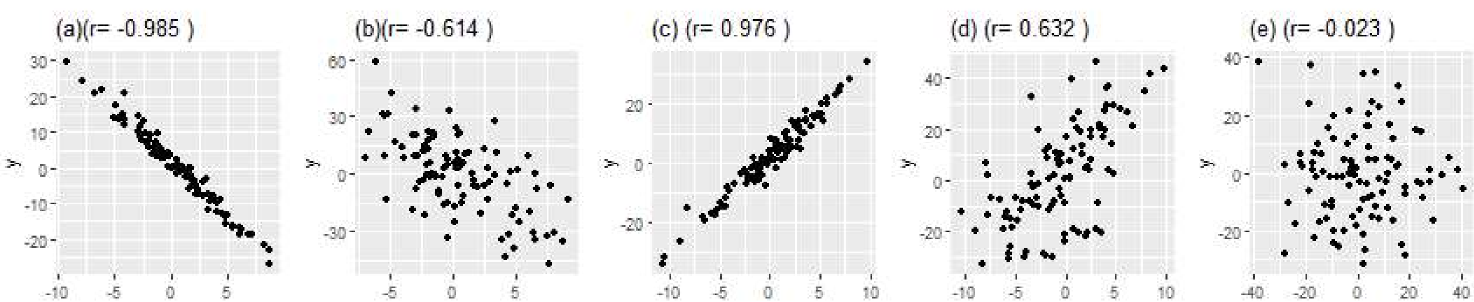


Numerical data summaries

- **Covariance** and **correlation**: measure extent of linear relationship between two variables (x and y).

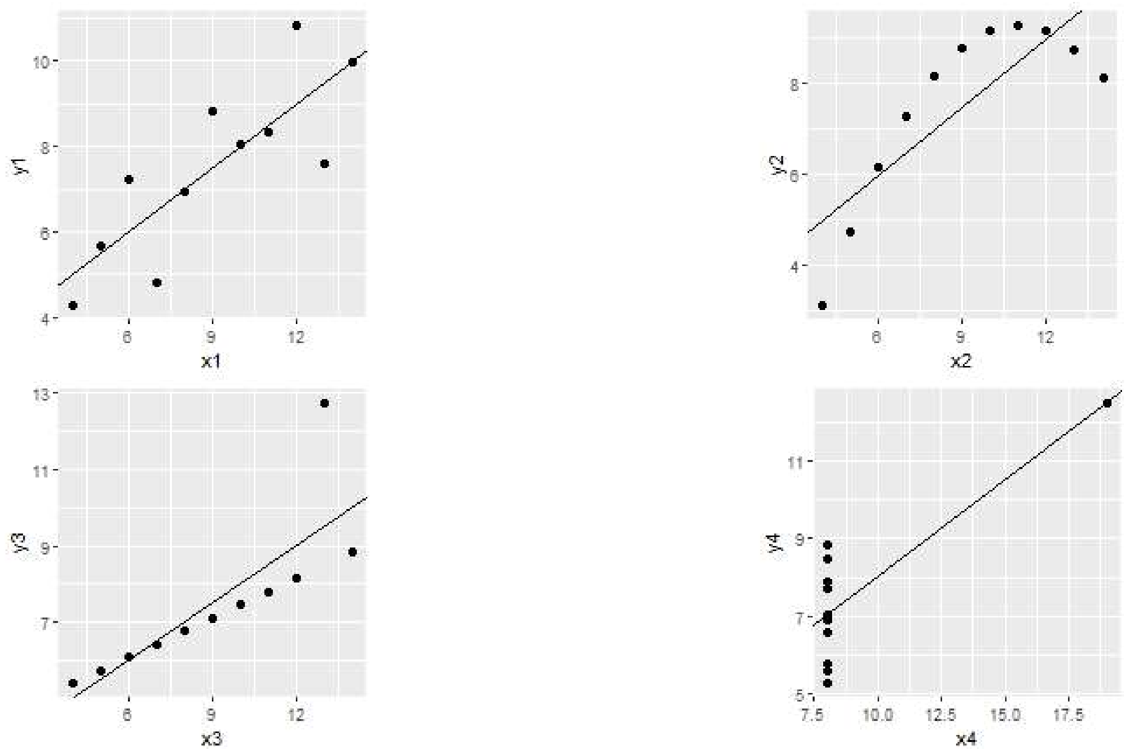
$$r = \frac{\sum(x-\bar{x})(y-\bar{y})}{\sqrt{\sum(x-\bar{x})^2\sum(y-\bar{y})^2}}$$

- Lies between -1 and +1



Correlation coefficient

Which one has the highest correlation?



- All these have $r = 0.82$.

Autocorrelation

Autocovariance (c_k) and **auto**correlation (r_k): measure linear relationship between lagged values of a time series y.

- We measure the relationship between:

y_t and y_{t-1}

y_t and y_{t-2}

y_t and y_{t-3}

$\dots$

y_t and y_{t-k}

- We denote the sample autocovariance at lag k by c_k and the sample autocorrelation at lag k by r_k.

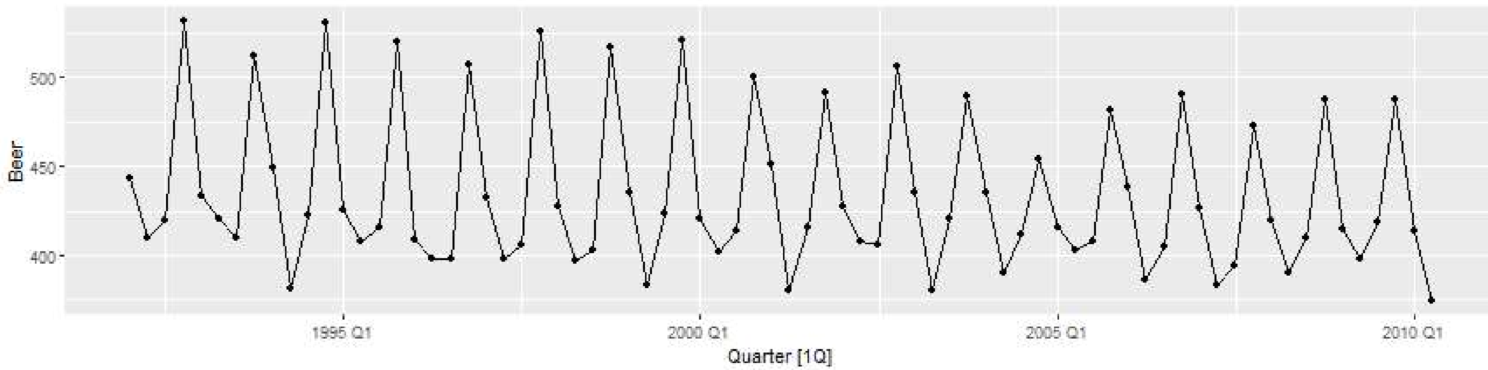
Then define

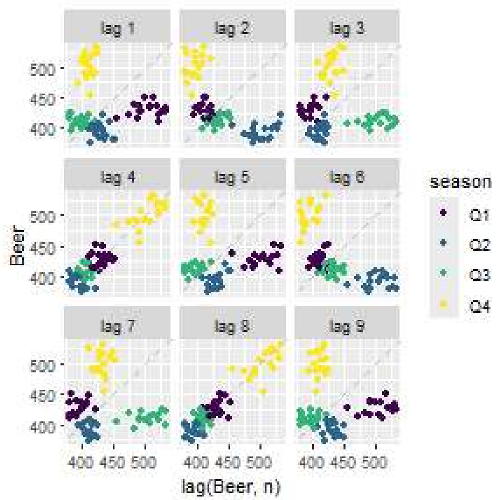
$$r_k = \frac{c_k}{c_0} = \frac{\sum_{t=k+1}^{T}(y_t - \bar{y})(y_{t-k} - \bar{y})}{\sum_{t=1}^{T}(y_t - \bar{y})^2}$$

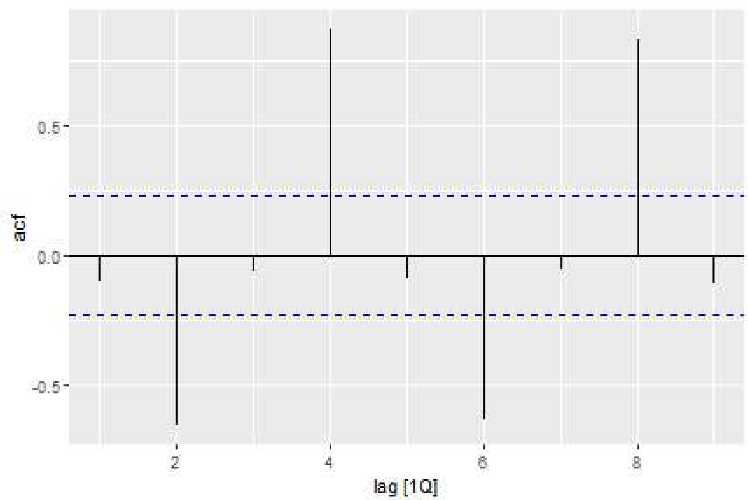
Autocorrelation

- r_1 indicates how successive values of y relate to each other
- r_2 indicates how y values two periods apart relate to each other
- r_k is almost the same as the sample correlation between y_t and y_{t-k}.

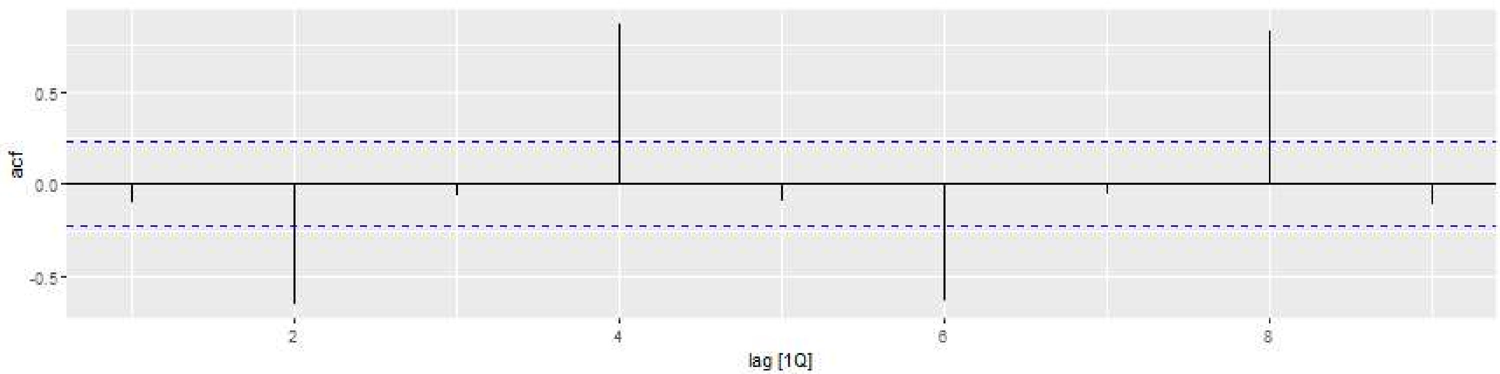
Autocorrelation: Results for first 9 lags for beer data:





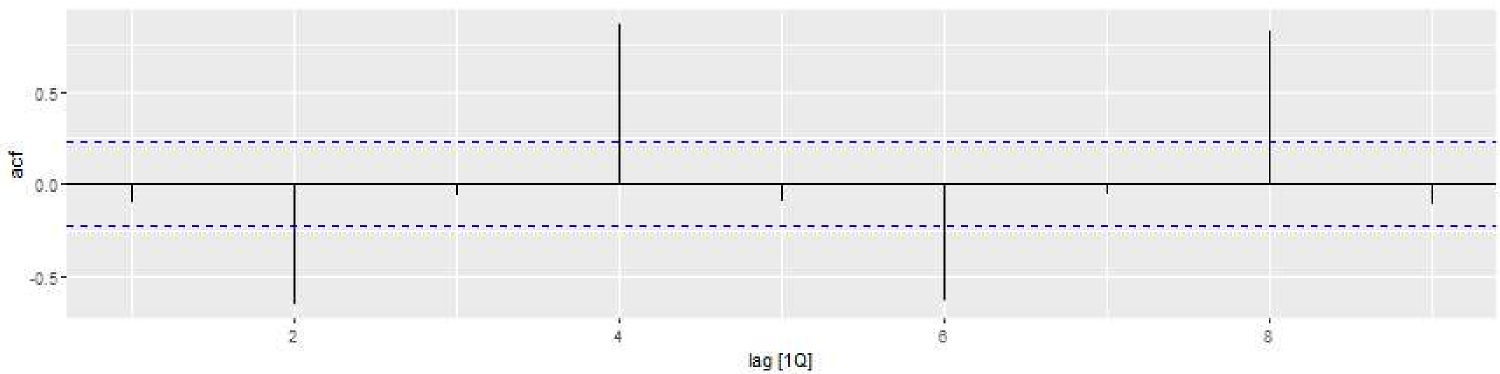


Autocorrelation: Results for first 9 lags for beer data:



- r_4 is **positive and higher** than for the other lags. This is due to the **seasonal pattern** in the data.
  - the peaks (troughs) tend to be 4 quarters apart.
  - the spikes every 4 lags after this $(r_8, r_{12}, \ldots)$ decrease in size as the lag number increases.

Autocorrelation: Results for first 9 lags for beer data:

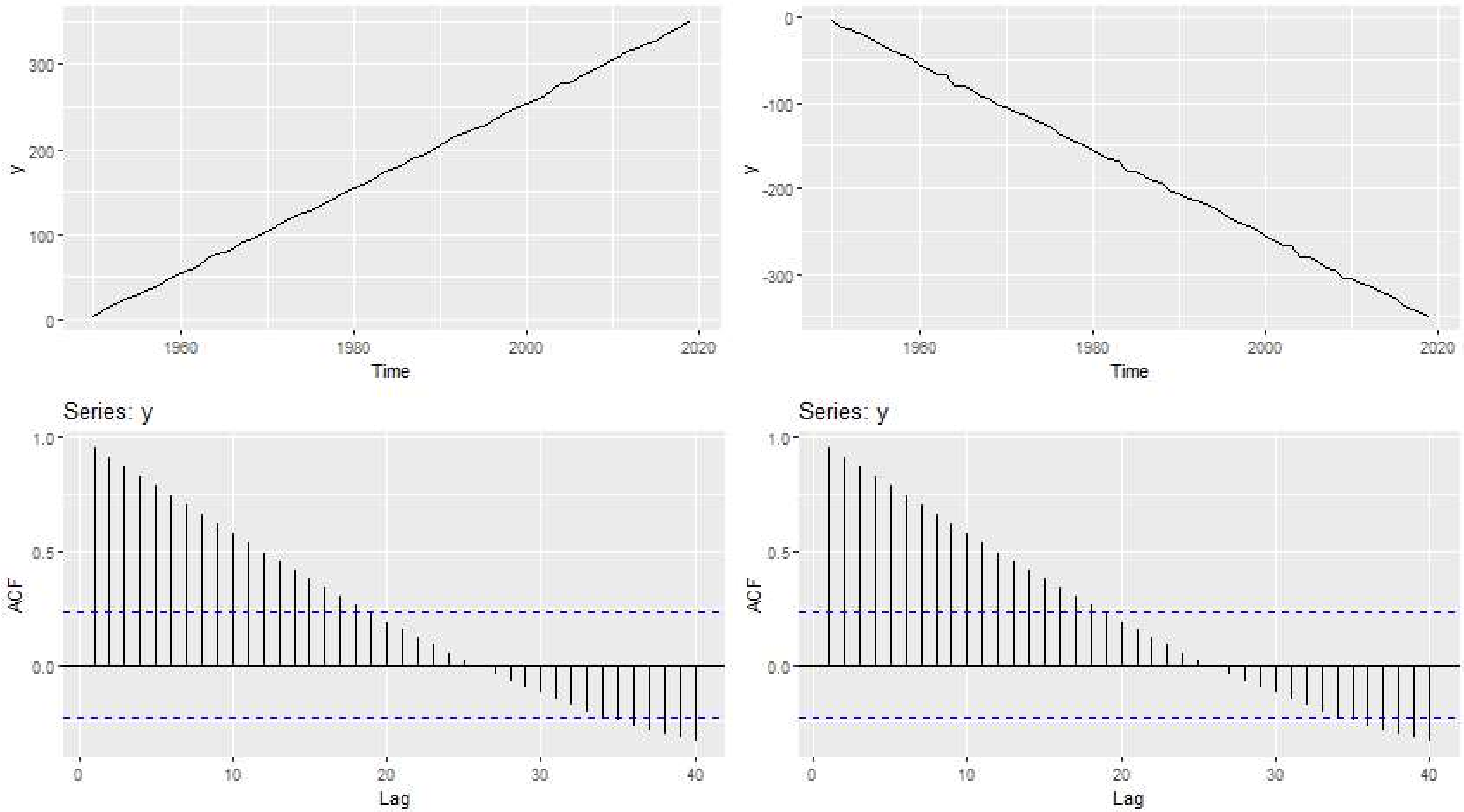


- r_2 is **more negative** than for the other lags because troughs and peaks tend to be 2 quarters apart.
  - The highest and the lowest productions are 2 quarters apart.
  - The spikes every 4 lags after this $r_6, r_{10}, \ldots$ decrease in size as the lag number increases.
- Together, the autocorrelations at lags $1, 2, \ldots,$ make up the **autocorrelation** or **ACF**.
- The plot is known as a **correlogram**.
- The dashed blue lines indicate whether the correlations are significantly different from zero.

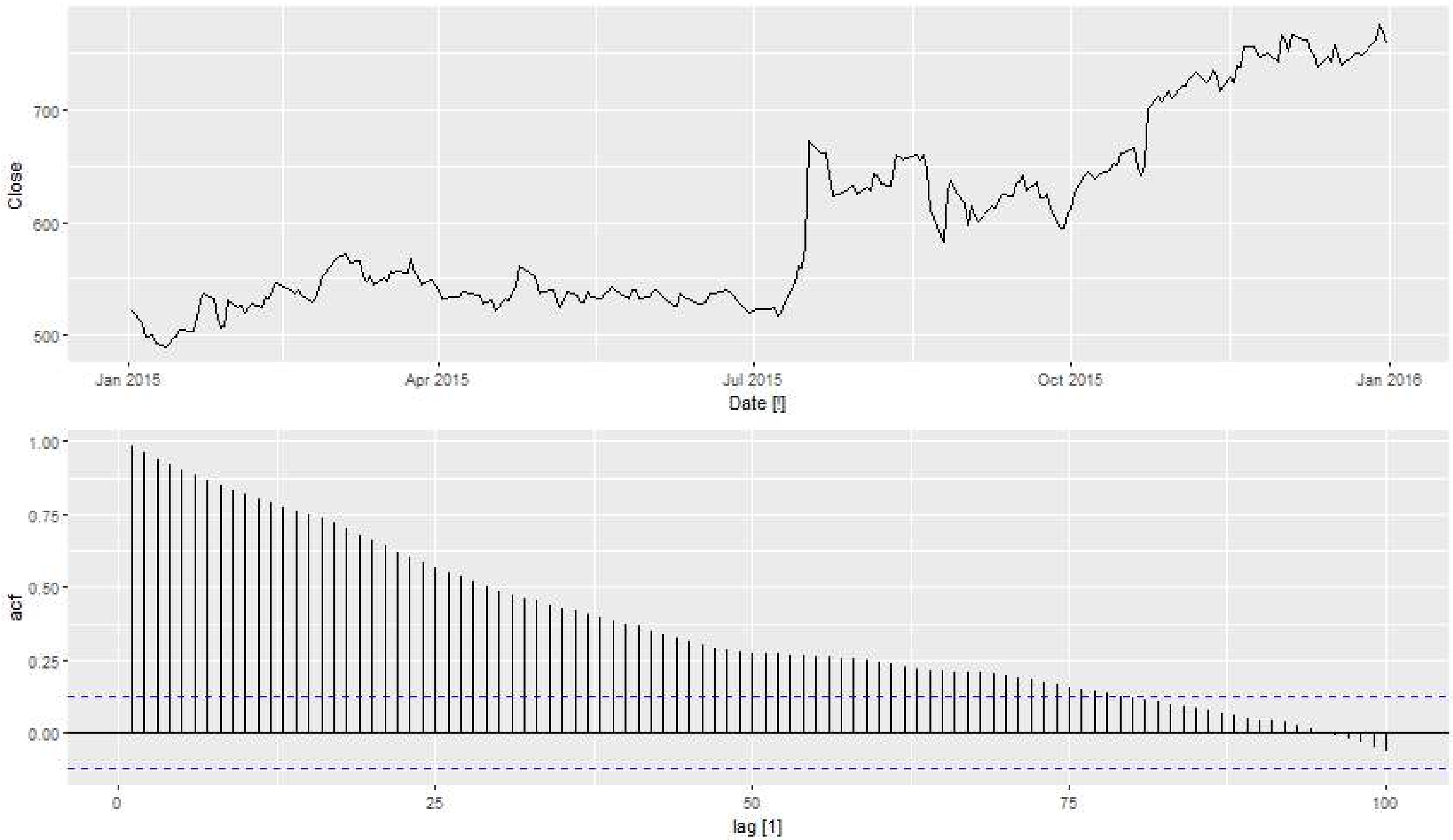
Trend and seasonality in ACF plots

- When data have a trend, the autocorrelations for small lags tend to be large and positive.
- When data are seasonal, the autocorrelations will be larger at the seasonal lags (i.e., at multiples of the seasonal frequency)
- When data are trended and seasonal, you see a combination of these effects

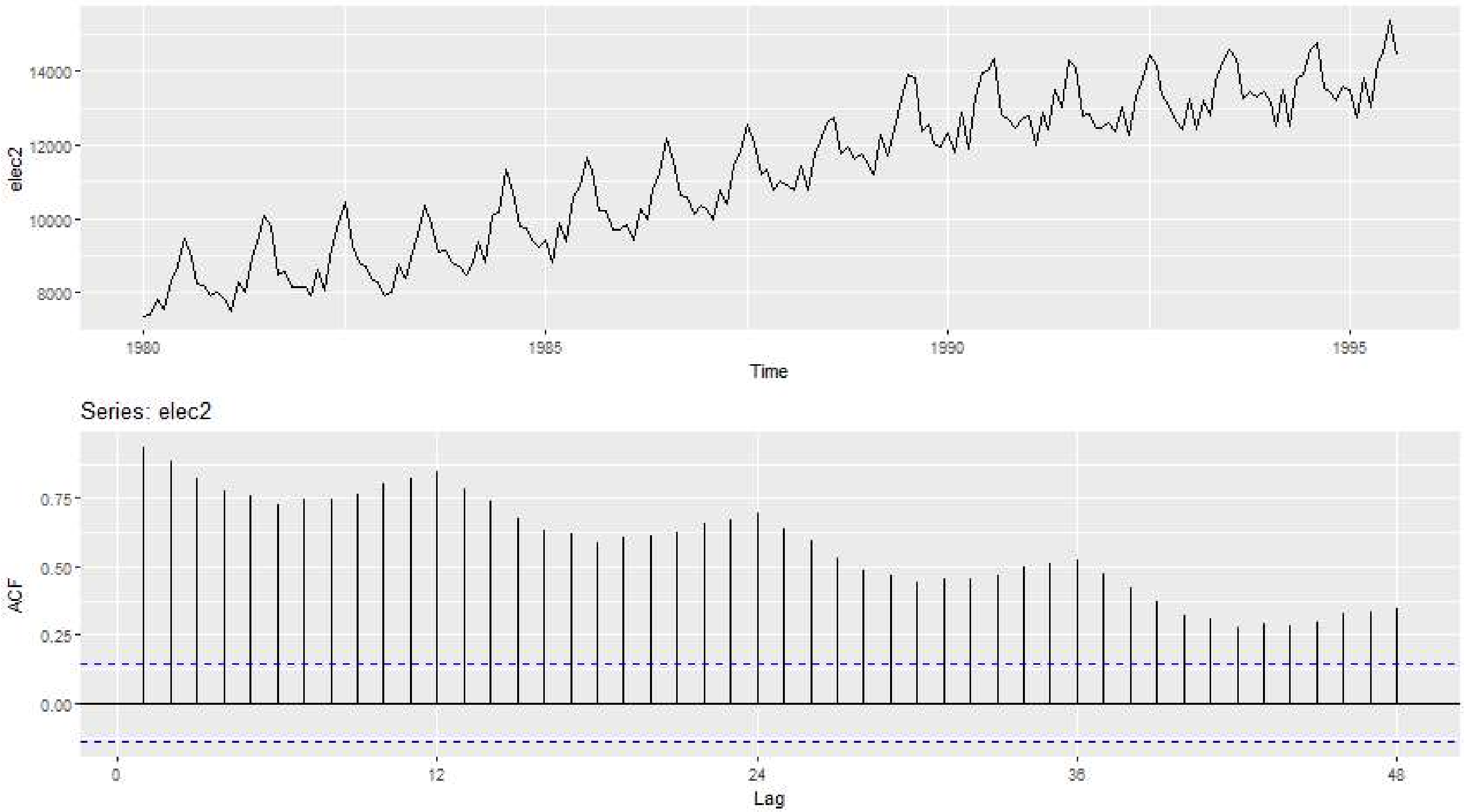
Trend



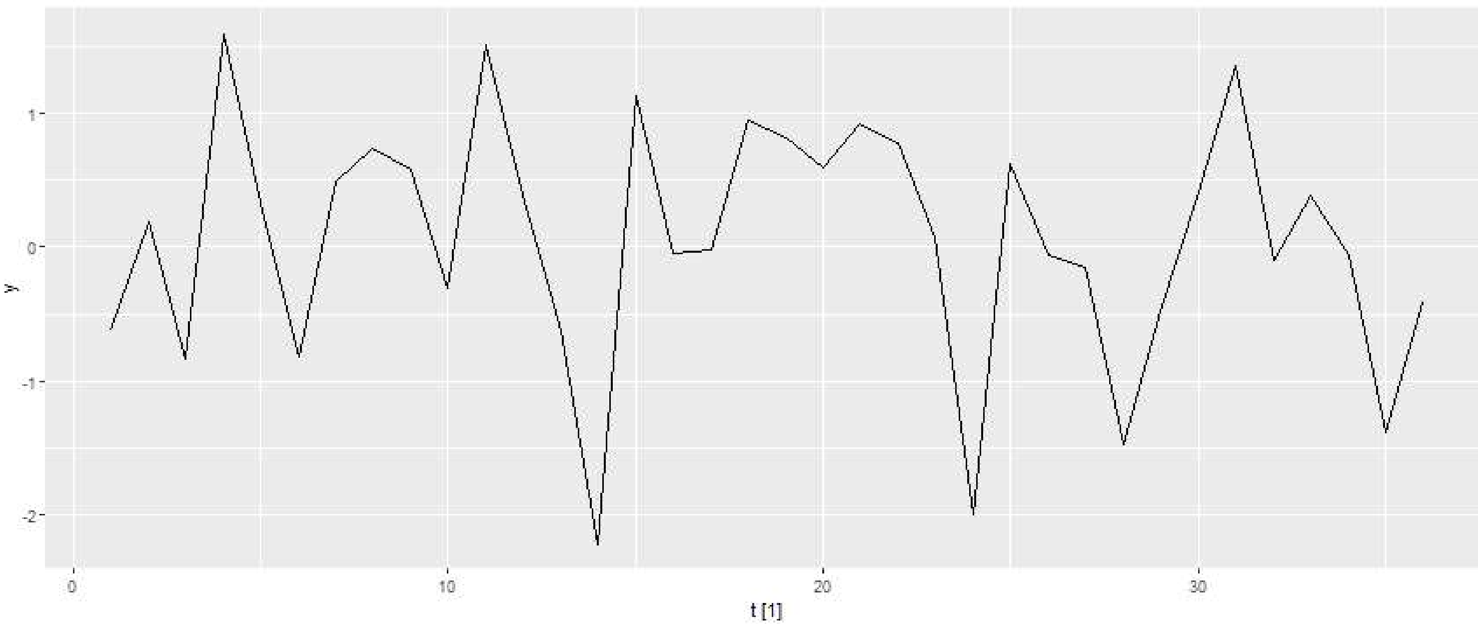
Google stock price



Australian monthly electricity production

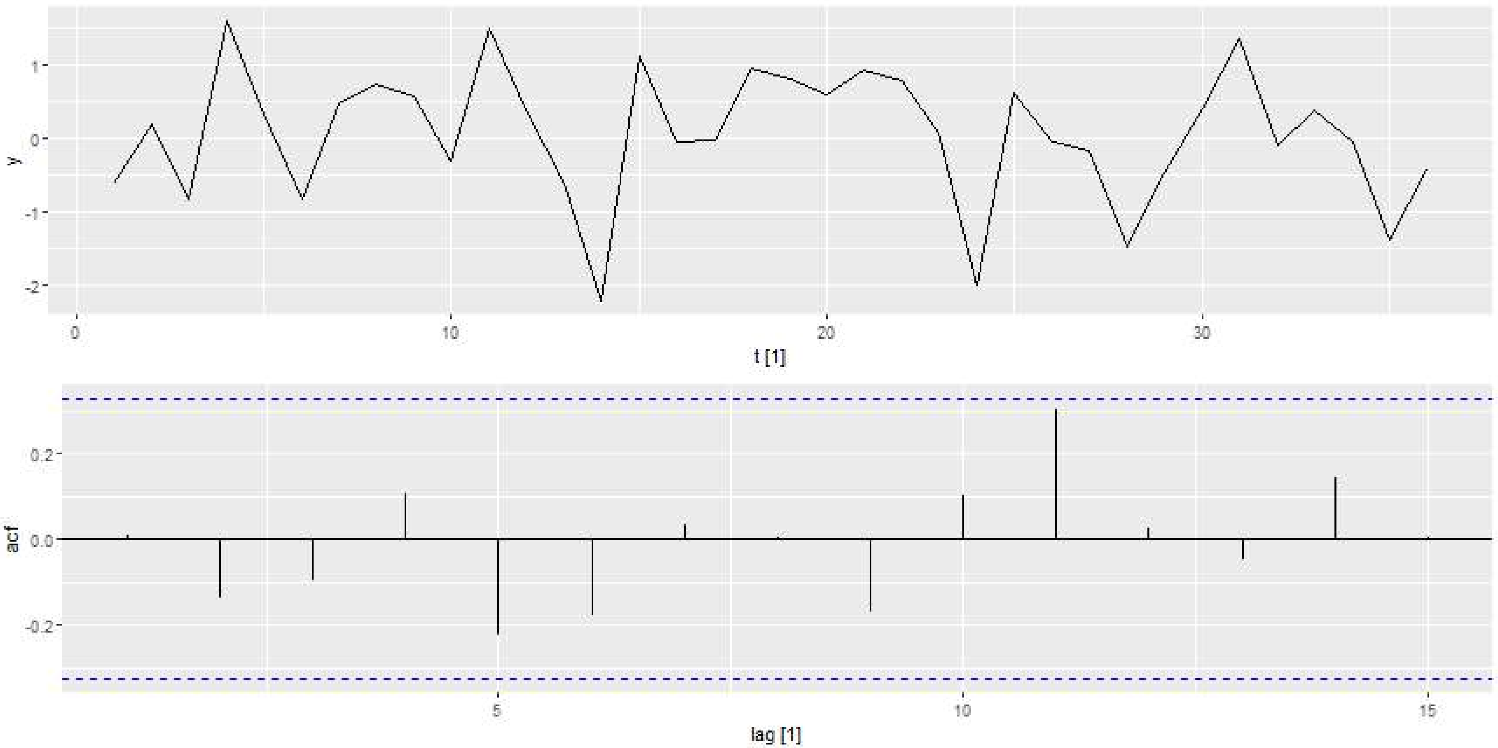


White noise



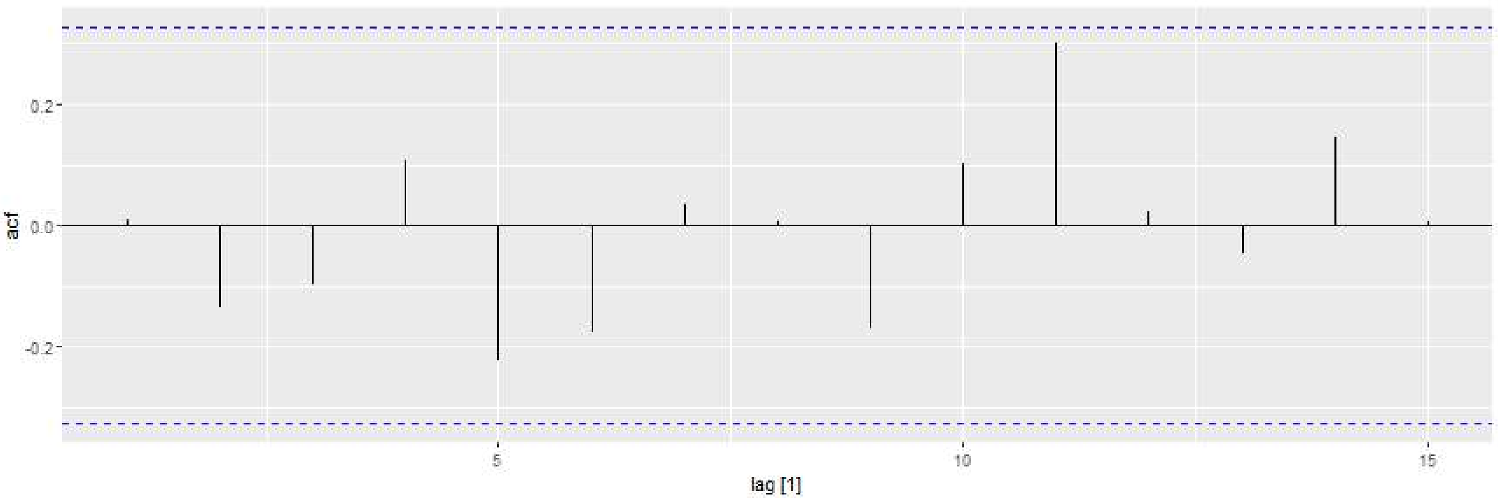
- **White noise data** is uncorrelated across time with zero mean and constant variance.
- Technically, we require independence as well.

Sample autocorrelations for white noise series



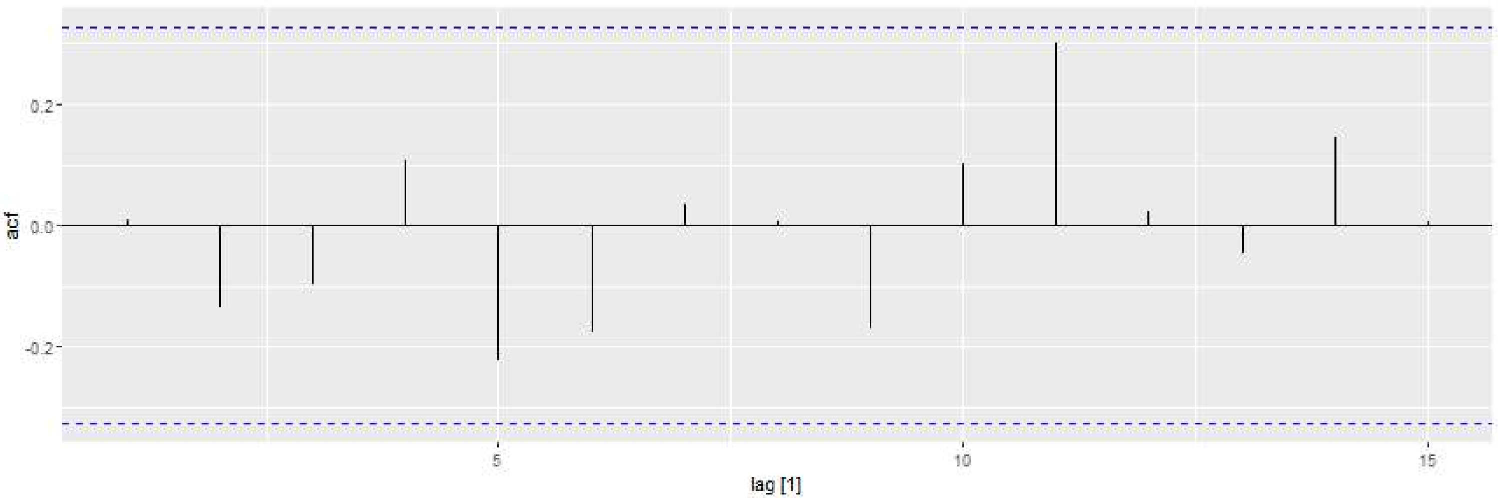
- For uncorrelated data, we would expect each one to be close to zero.
- Blue lines show 95% **critical values**.

Sampling distribution of autocorrelations



- Sampling distribution of r_k for white noise data is asymptotically $N(0, 1/T)$.
- 95% of all r_k for white noise must lie within $\pm 1.96/\sqrt{T}$.
- If this is not the case, the series is probably not WN.
- Common to plot lines at $\pm 1.96/\sqrt{T}$ when plotting ACF. These are the **critical values**.

Sampling distribution of autocorrelations

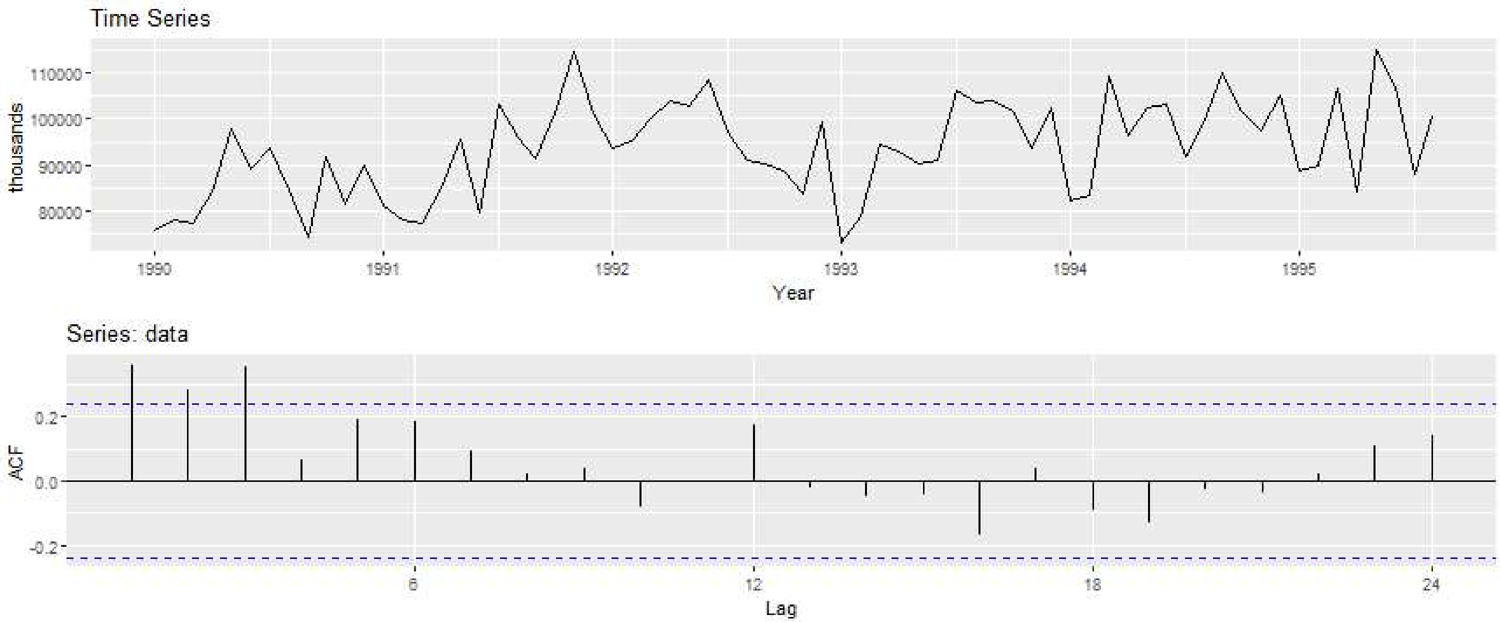


Example:

- $T = 36$ and so critical values at $\pm 1.96/\sqrt{36} = \pm 0.327$.
- All autocorrelation coefficients lie within these limits, confirming that the data are white noise.

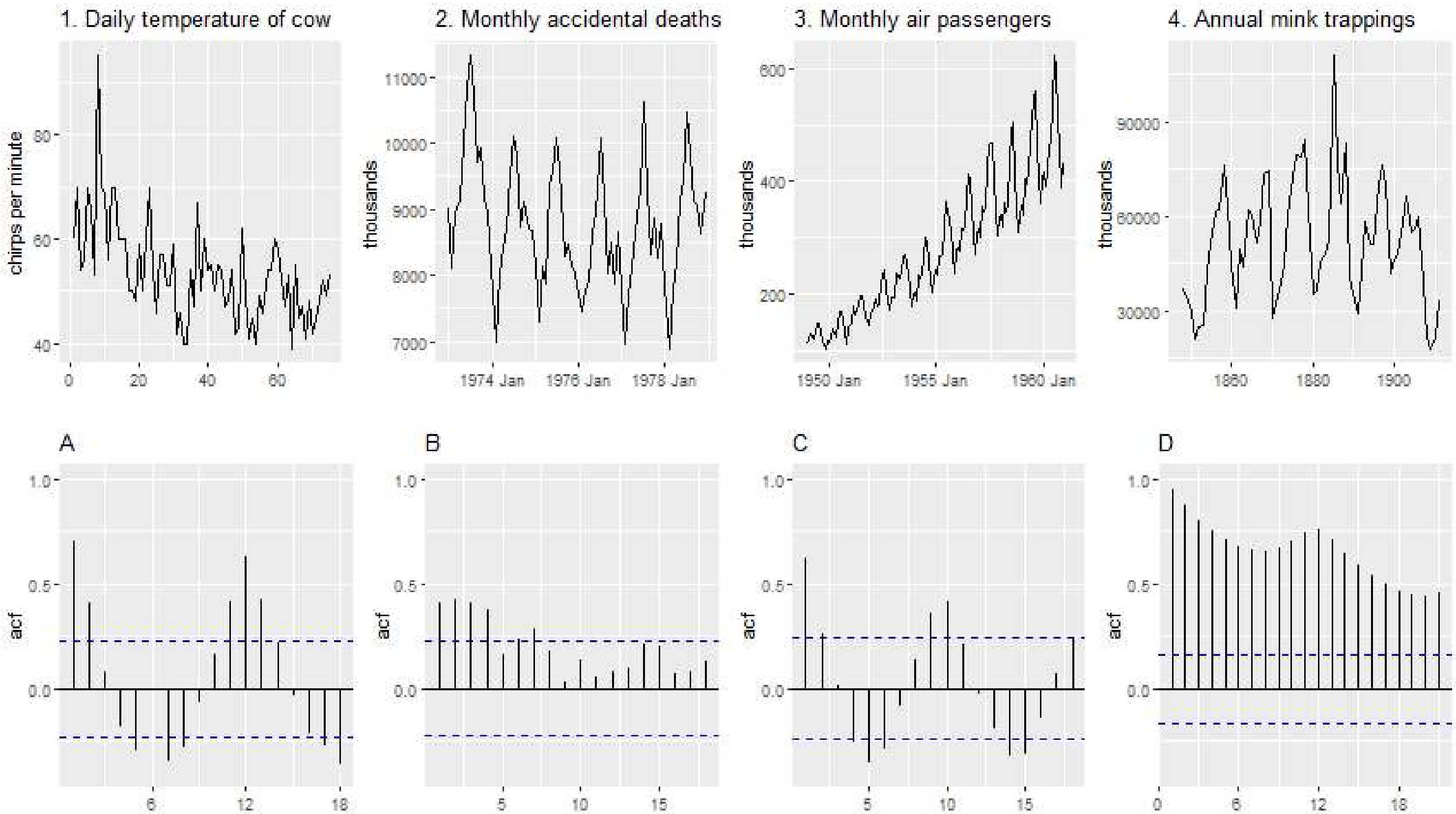
(More precisely, the data cannot be distinguished from white noise.

Example



- Difficult to detect pattern in time plot.
- ACF shows some significant autocorrelation at lags 1, 2, and 3.
- r_{12} relatively large although not significant.This may indicate some slight seasonality.
- These show the series is **not a white noise series**.

Which is which?



References

- Hyndman, R. J., & Athanasopoulos, G. (2018). Forecasting: principles and practice. OTexts.
- Mathai, A. M., & Haubold, H. J. (2008). Applications to Stochastic Process and Time Series. Special Functions for Applied Scientists, 247-295.