Time Series Analysis

Trend

- **Trend** is the long term movement of data over time. This definition implies that time is the independent variable and the data or set of observations we are interested in is the dependent variable. When we track data purely as function of time, there are several possible scenarios. First, data may exhibit **no trend** as shown in the example below. In this case data remains constant and is unaffected by time.

Time Period	1	2	3	4	5
Data Value	30	30	30	30	30

- The second possible scenario is **linear trend.** In this case, data as a function of time has a linear relationship as shown in the example below. The table above shows that rate of increase data between successive time periods is a constant two units. This series is called an **Arithmetic Progression**. It should be noted that data can also exhibit negative linear trend with a rate of decrease between successive observations.

Time Period	1	2	3	4	5
Data Value	30	32	34	36	38

- The behavior of data over time may also exhibit a trend pattern that is **nonlinear** such as **exponential growth or decay** . An example of observations that have an exponential growth pattern is shown in the table below. In this case, each successive data value is twice its previous value. In this example each pair of successive observations have a common ratio. Such a series is called a **Geometric Progression**.
-

Time Period	1	2	3	4	5
Data Value	30	60	120	240	480

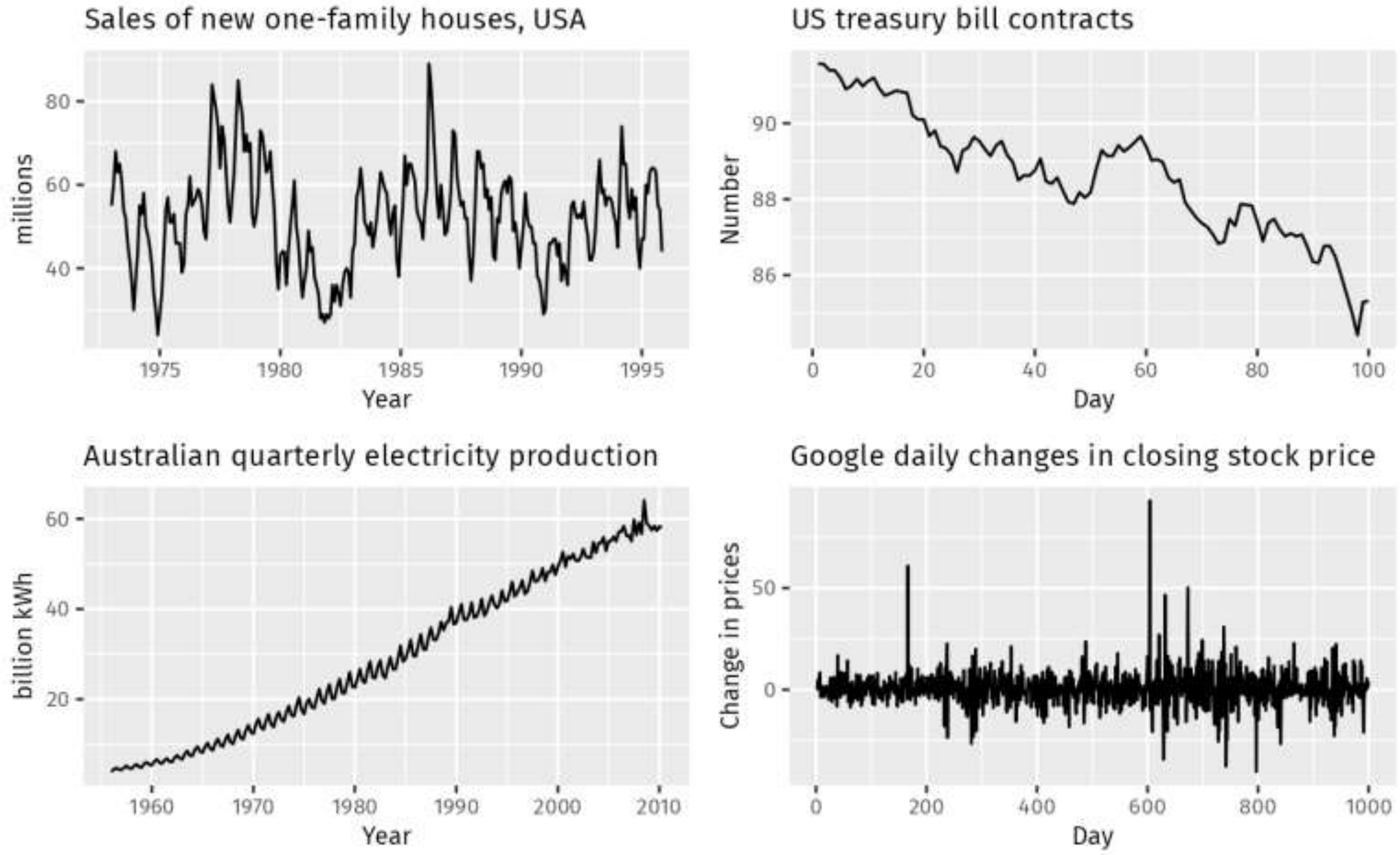
- In the case of exponential decay, each succeeding observation decreases by some constant factor. This is another form of the Geometric Progression and a sort of pattern that is observed with such phenomena as the decay in radiation levels from nuclear activity. The measure frequently used in exponential decay is the "half-life." It is the time it takes for the dependent variable to decay to half its original value. An example of exponential decay is presented in the table below. In this example of exponential decay, the half-life is one period.
-

Time Period	1	2	3	4	5
Data Value	400	200	100	50	25

- **Seasonal variations** can be another component of a time series. These are periodic, short term, fairly regular fluctuations in data caused by man-made or weather factors. The increase in demand for candies during the Christmas season is an example of seasonal variations in data. **Cyclical variations** in a time series are wave-like oscillations in data about the trend line and typically have more than one-year duration. These variations are often caused by economic or political factors. **Random variations** are variations in data not accounted for by any of the previous components of the time series. These variations cannot be easily predicted and are only after the fact. In forecasting, these variations are accounted for as an error term. The decrease in demand for a company's product due to a plant shutdown caused by a labor strike is an example of a random variation in demand.

- In addition to the qualitative and quantitative classification discussed above, forecasting methods can also be classified based on time frame. These are short term and intermediate term forecasting methods. Forecasting for the long term is typically done using by the qualitative methods discussed earlier in this lesson. We will now explore the various forecasting techniques that can be employed for the short and intermediate term forecasting.
-

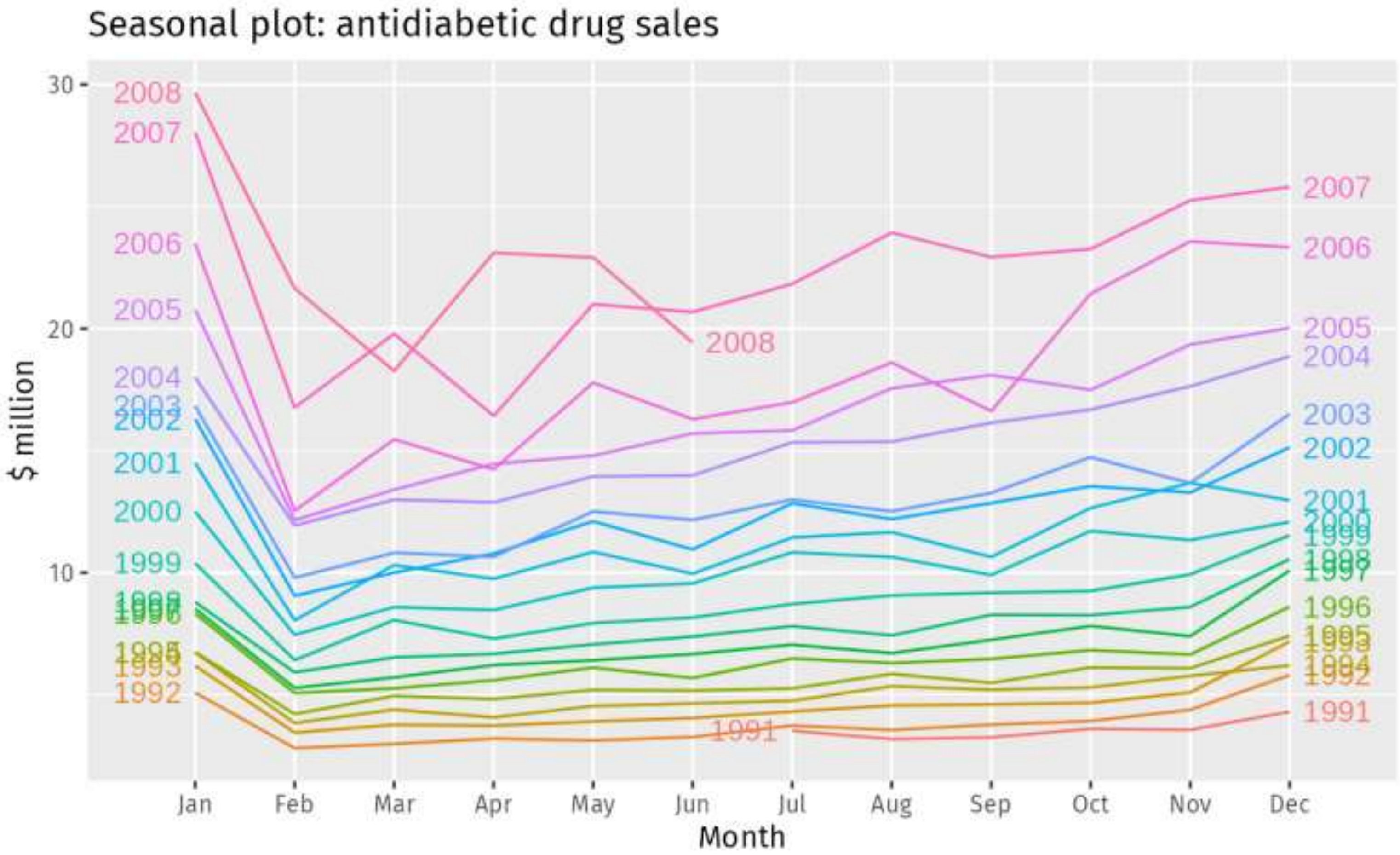
- Trend
- A trend exists when there is a long-term increase or decrease in the data. It does not have to be linear. Sometimes we will refer to a trend as “changing direction”, when it might go from an increasing trend to a decreasing trend. There is a trend in the antidiabetic drug sales data.
- Seasonal
- A seasonal pattern occurs when a time series is affected by seasonal factors such as the time of the year or the day of the week. Seasonality is always of a fixed and known frequency. The monthly sales of antidiabetic drugs above shows seasonality which is induced partly by the change in the cost of the drugs at the end of the calendar year.
- Cyclic
- A cycle occurs when the data exhibit rises and falls that are not of a fixed frequency. These fluctuations are usually due to economic conditions, and are often related to the “business cycle”. The duration of these fluctuations is usually at least 2 years.


Sales of new one-family houses, USA
millions
80
60
40
1975
1980
1985
1990
1995
Year
US treasury bill contracts
Number
90
88
86
0
20
40
60
80
100
Day
Australian quarterly electricity production
billion kWh
60
40
20
1960
1970
1980
1990
2000
2010
Year
Google daily changes in closing stock price
Change in prices
50
0
0
200
400
600
800
1000
Day

- The monthly housing sales (top left) show strong seasonality within each year, as well as some strong cyclic behaviour with a period of about 6–10 years. There is no apparent trend in the data over this period.
- The US treasury bill contracts (top right) show results from the Chicago market for 100 consecutive trading days in 1981. Here there is no seasonality, but an obvious downward trend. Possibly, if we had a much longer series, we would see that this downward trend is actually part of a long cycle, but when viewed over only 100 days it appears to be a trend.
- The Australian quarterly electricity production (bottom left) shows a strong increasing trend, with strong seasonality. There is no evidence of any cyclic behaviour here.
- The daily change in the Google closing stock price (bottom right) has no trend, seasonality or cyclic behaviour. There are random fluctuations which do not appear to be very predictable, and no strong patterns that would help with developing a forecasting model.
-

Seasonal plots

- A seasonal plot is similar to a time plot except that the data are plotted against the individual “seasons” in which the data were observed. An example is given below showing the antidiabetic drug sales.
- ggseasonplot(a10, year.labels=TRUE, year.labels.left=TRUE) +
- ylab("$ million") +
- ggtitle("Seasonal plot: antidiabetic drug sales")

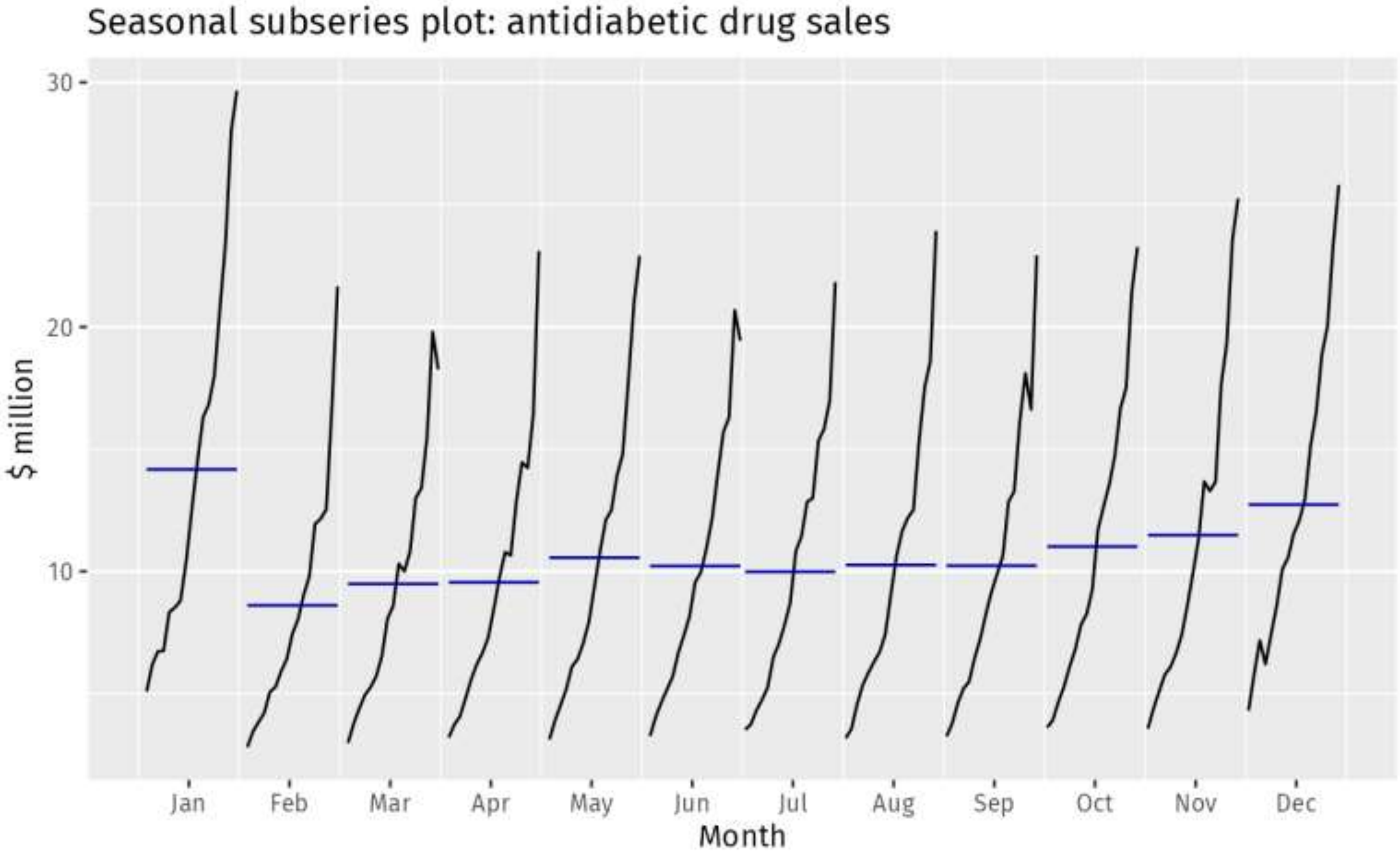


Seasonal plot of monthly antidiabetic drug sales in Australia.

Seasonal subseries plots

- An alternative plot that emphasises the seasonal patterns is where the data for each season are collected together in separate mini time plots.
- ggsubseriesplot(a10) +
- ylab("$ million") +
- ggtitle("Seasonal subseries plot: antidiabetic drug sales")

Figure 2.6: Seasonal subseries plot of monthly antidiabetic drug sales in Australia.

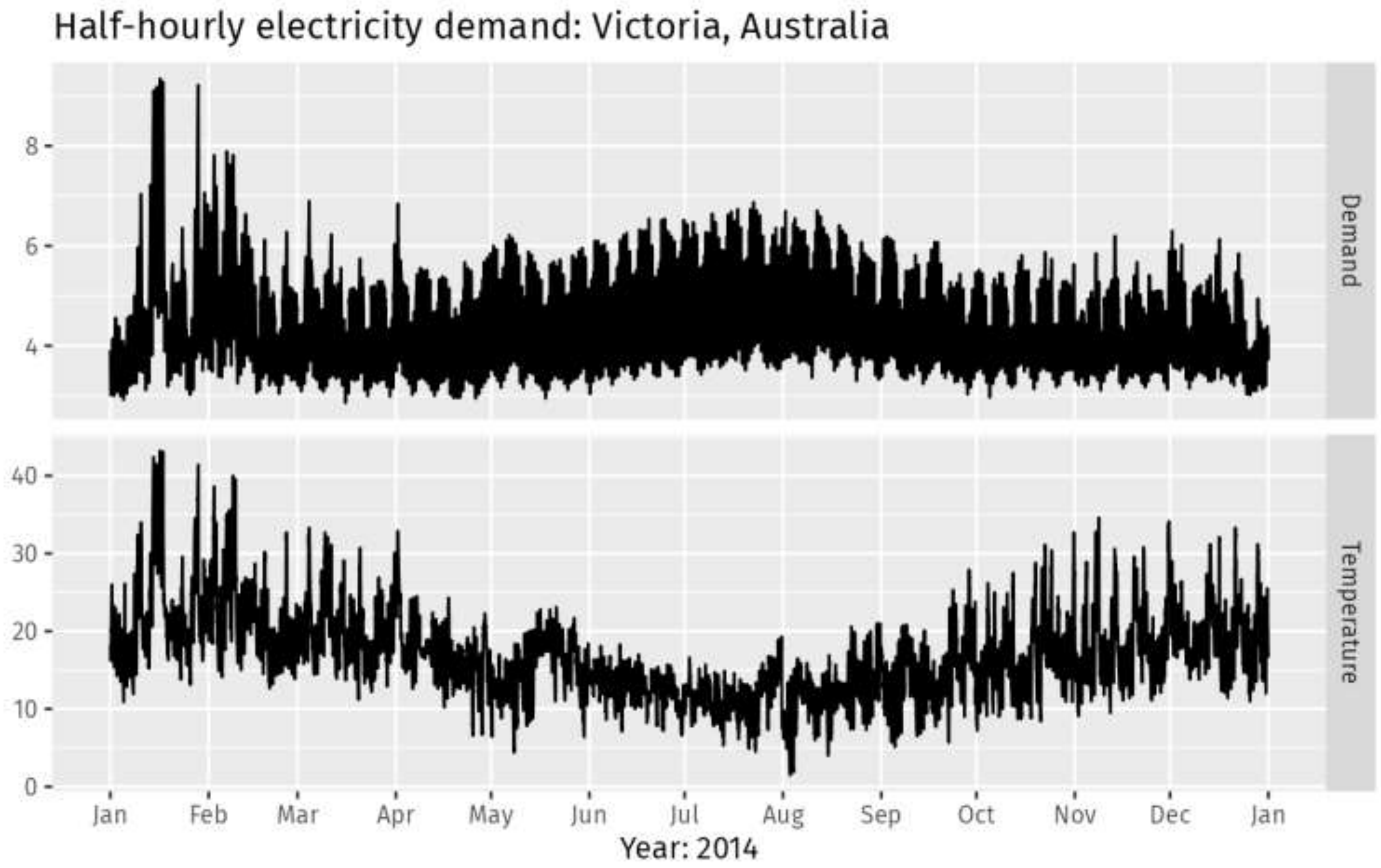


Scatterplots

- It is also useful to explore relationships *between* time series.

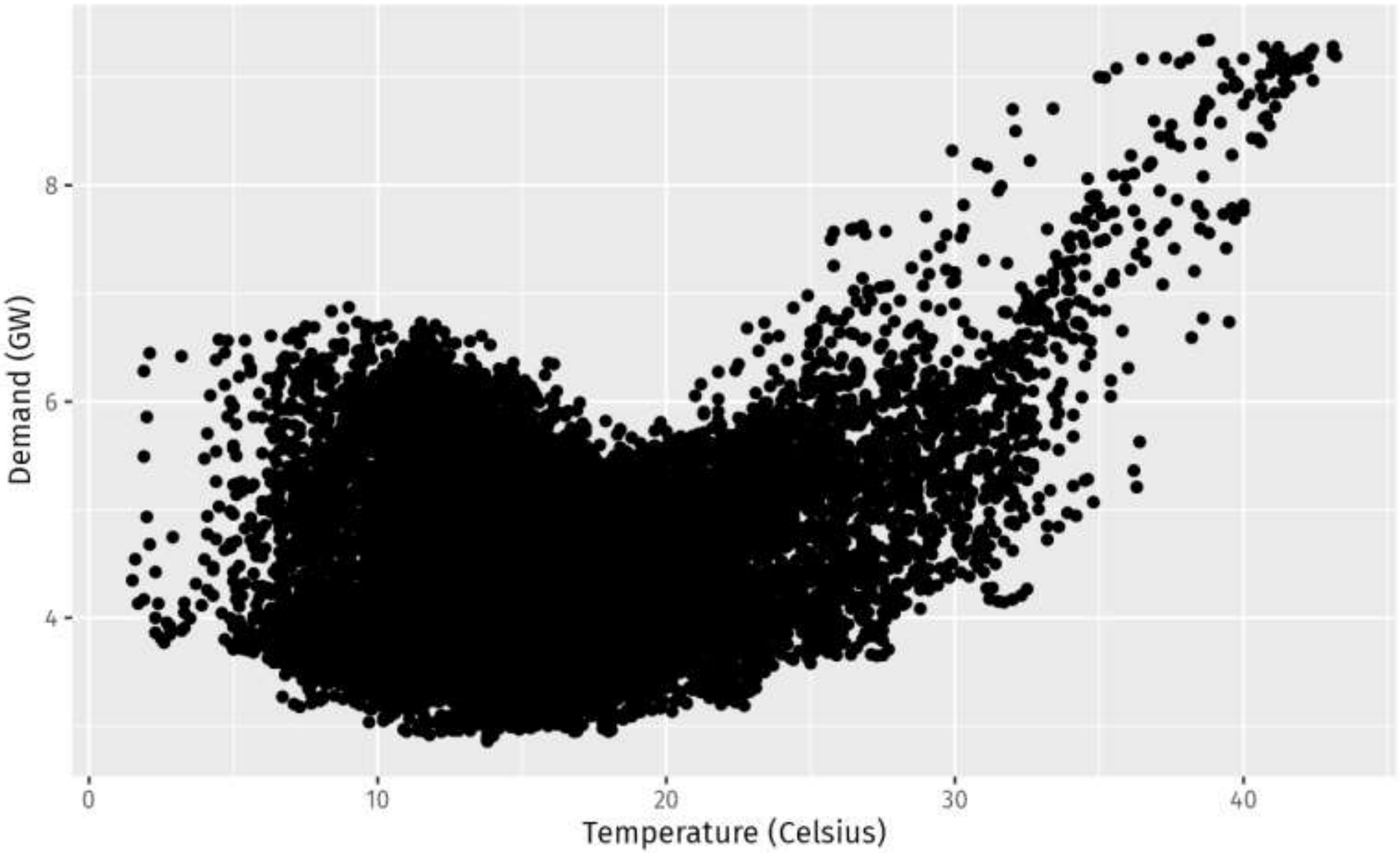
- half-hourly electricity demand (in Gigawatts) and temperature (in degrees Celsius), for 2014 in Victoria, Australia. The temperatures are for Melbourne, the largest city in Victoria, while the demand values are for the entire state.
- autoplot(elecdemand[,c("Demand","Temperature")], facets=TRUE) +
- xlab("Year: 2014") + ylab("") +
- ggtitle("Half-hourly electricity demand: Victoria, Australia")

Half hourly electricity demand and
temperatures in Victoria, Australia, for 2014.



- the relationship between demand and temperature by plotting one series against the other.
- as.data.frame(elecdemand) |>
- ggplot(aes(x=Temperature, y=Demand)) +
- geom_point() +
- ylab("Demand (GW)") + xlab("Temperature (Celsius)")

Half-hourly electricity demand plotted against temperature for 2014 in Victoria, Australia.



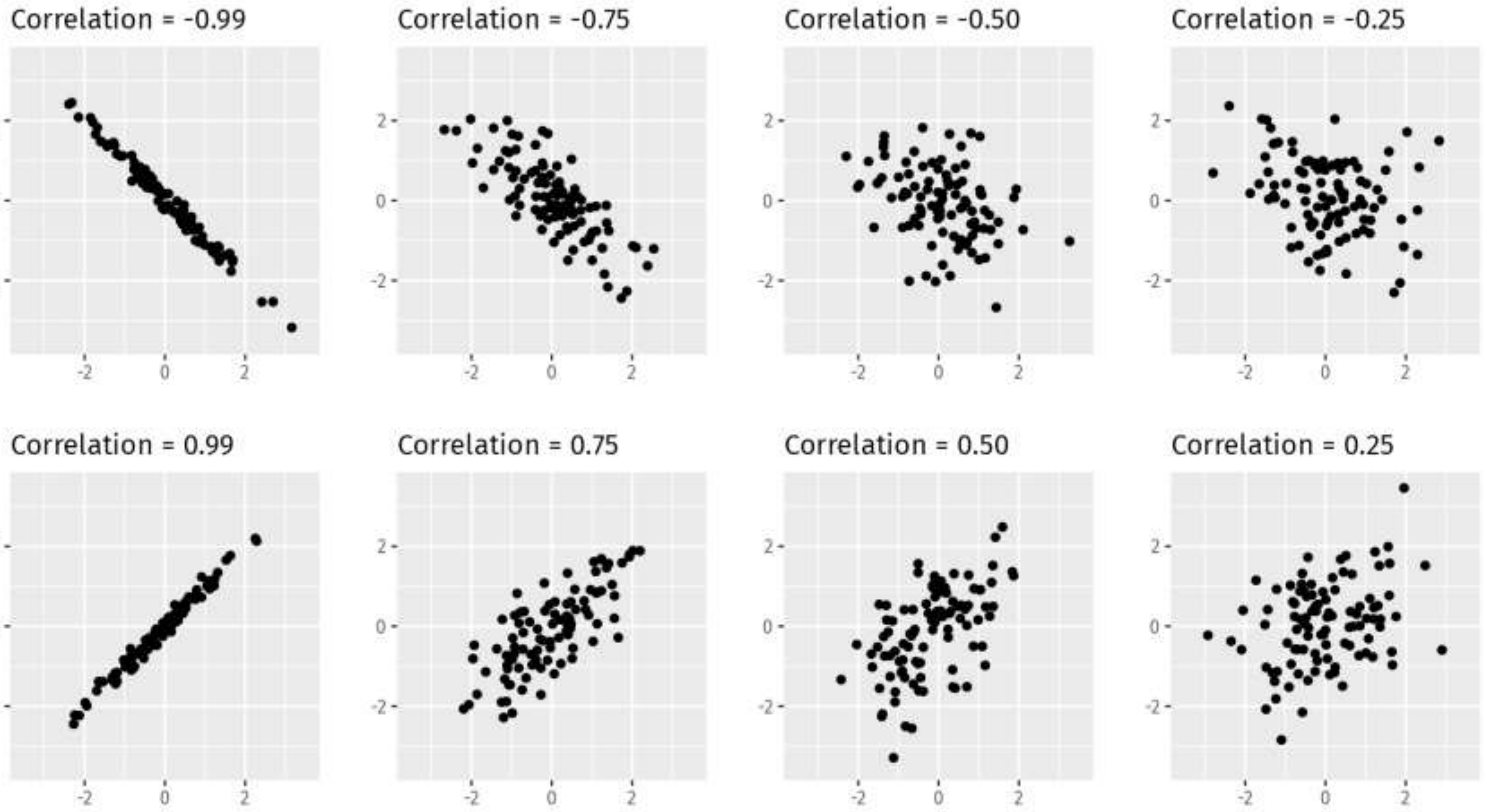
Correlation

- It is common to compute *correlation coefficients* to measure the strength of the relationship between two variables. The correlation between variables x and y is given by

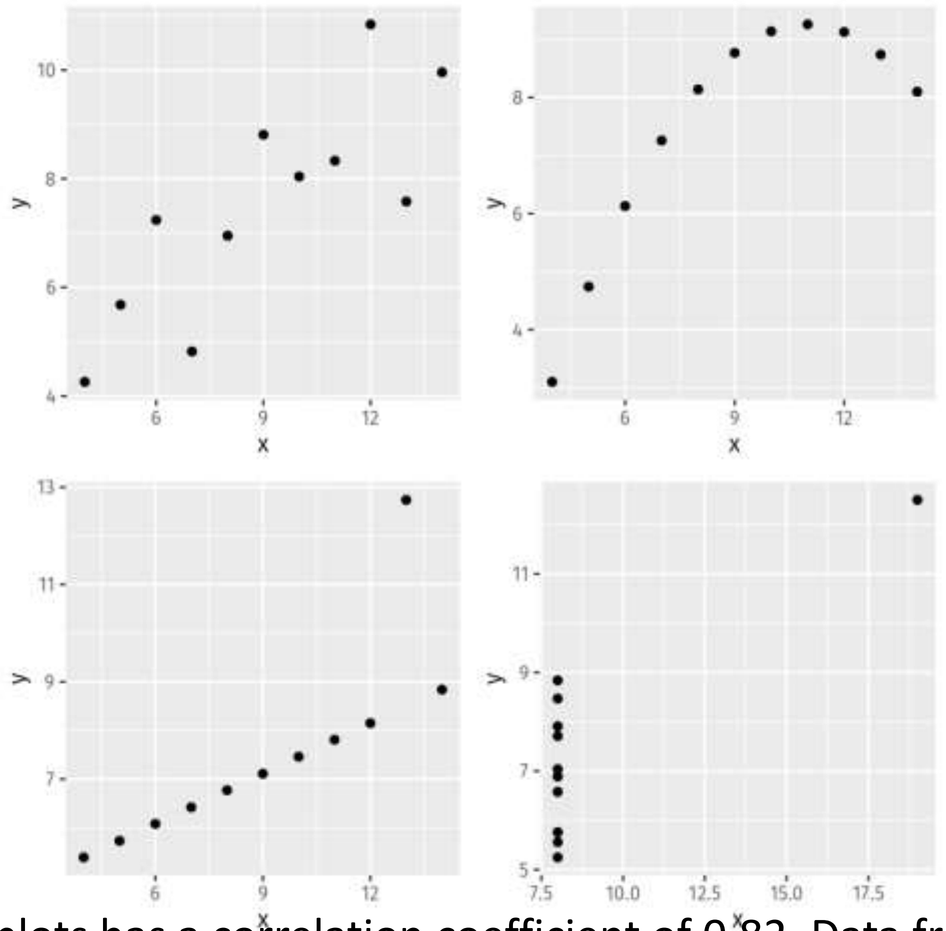
$$r = \frac{\sum(x_t - \bar{x})(y_t - \bar{y})}{\sqrt{\sum(x_t - \bar{x})^2}\sqrt{\sum(y_t - \bar{y})^2}}.$$

- The value of r always lies between −1 and 1 with negative values indicating a negative relationship and positive values indicating a positive relationship. The graphs in Figure show examples of data sets with varying levels of correlation.

Examples of data sets with different levels of correlation.



- The plots in Figure all have correlation coefficients of 0.82, but they have very different relationships. This shows how important it is to look at the plots of the data and not simply rely on correlation values.

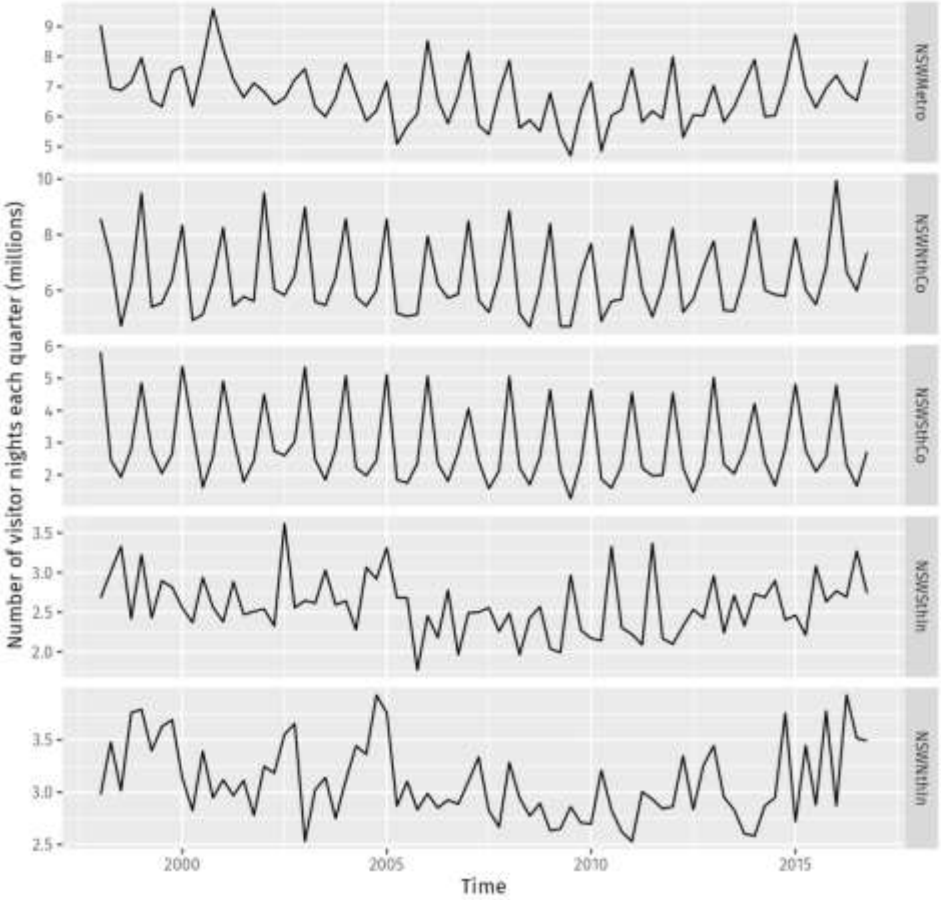


Each of these plots has a correlation coefficient of 0.82. Data from FJ Anscombe (1973) Graphs in statistical analysis. *American Statistician*, **27**, 17–21.

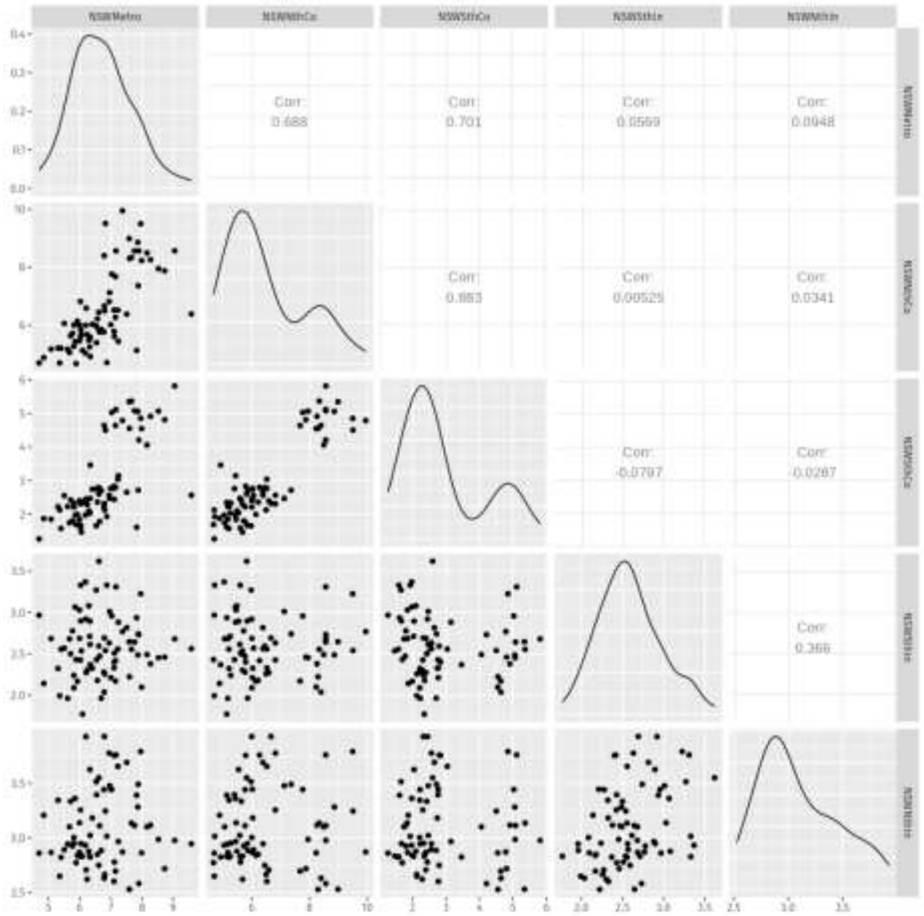
Scatterplot matrices

- When there are several potential predictor variables, it is useful to plot each variable against each other variable. Consider the five time series shown in Figure , showing quarterly visitor numbers for five regions of New South Wales, Australia.
- autoplot(visnights[,1:5], facets=TRUE) + ylab("Number of visitor nights each quarter (millions)")

Quarterly visitor nights for various regions of NSW, Australia.



- To see the relationships between these five time series, we can plot each time series against the others. These plots can be arranged in a scatterplot matrix, as shown in Figure (This plot requires the GGally package to be installed.)
- `GGally::ggpairs(as.data.frame(visnights[,1:5]))`

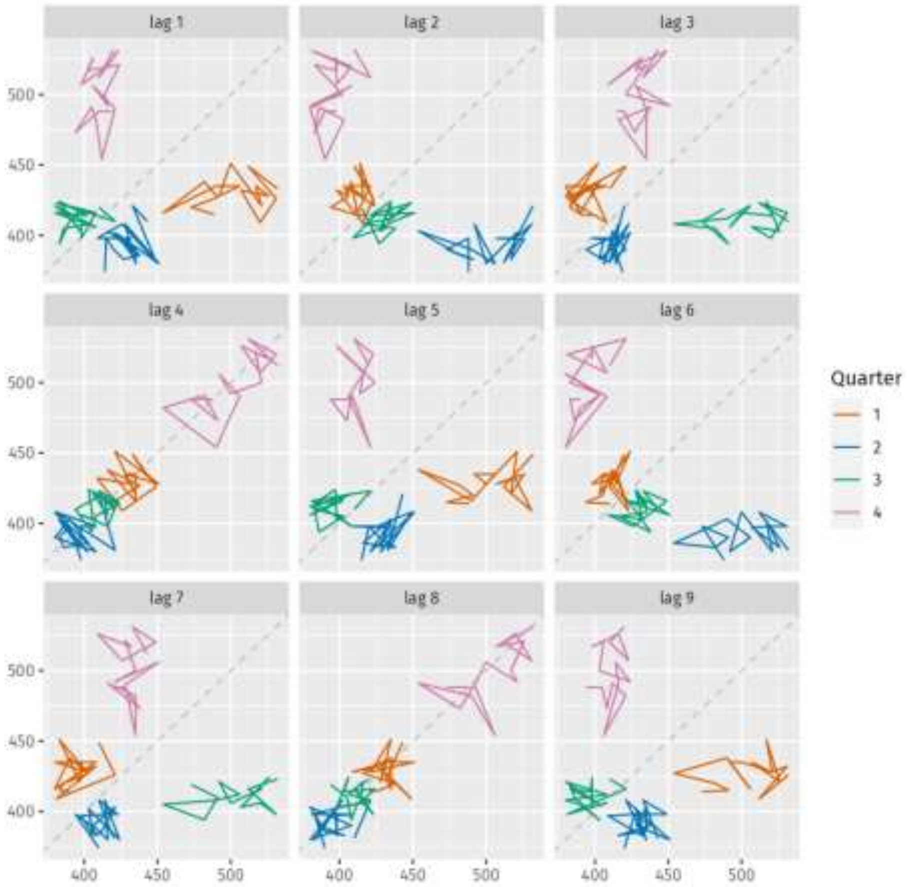


A scatterplot matrix of the quarterly visitor nights in five regions of NSW, Australia.

Lag plots

- Figure displays scatterplots of quarterly Australian beer production, where the horizontal axis shows lagged values of the time series. Each graph shows yt plotted against yt−k for different values of k.
- beer2 <- window(ausbeer, start=1992)
- gglagplot(beer2)

Lagged scatterplots for quarterly beer production.



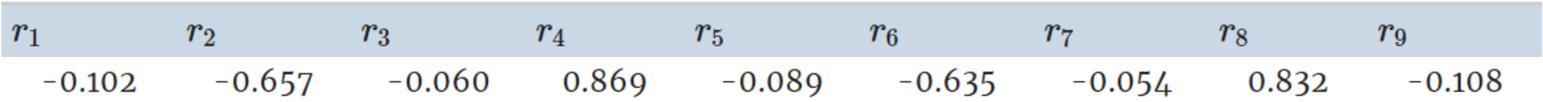
Autocorrelation

- Just as correlation measures the extent of a linear relationship between two variables, autocorrelation measures the linear relationship between *lagged values* of a time series.
- There are several autocorrelation coefficients, corresponding to each panel in the lag plot. For example, r1 measures the relationship between yt and yt−1, r2 measures the relationship between yt and yt−2, and so on.
- The value of rk can be written as

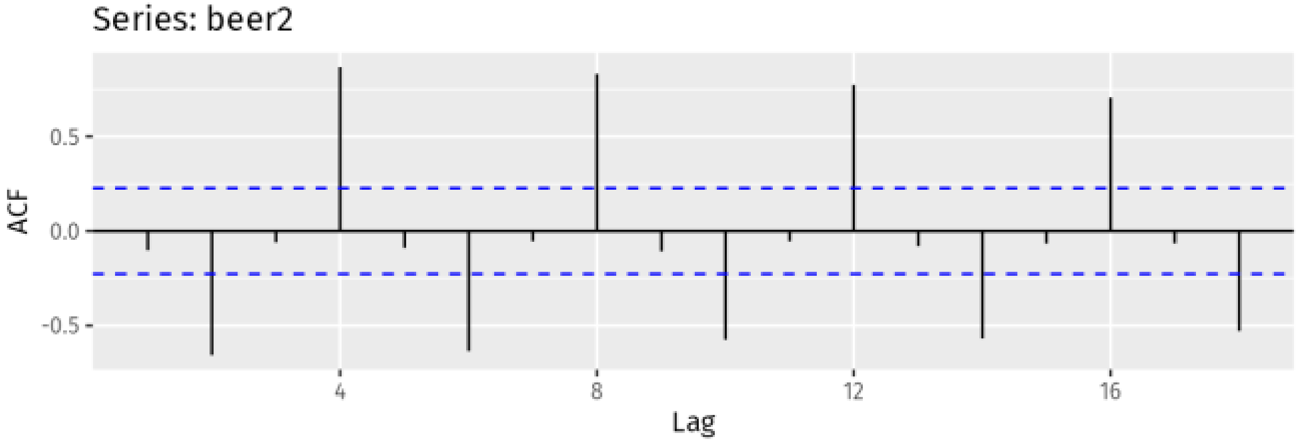
$$r_k = \frac{\sum\limits_{t=k+1}^{T} (y_t - \bar{y})(y_{t-k} - \bar{y})}{\sum\limits_{t=1}^{T} (y_t - \bar{y})^2}$$

- where T is the length of the time series.

- The first nine autocorrelation coefficients for the beer production data are given in the following table.



r_1	r_2	r_3	r_4	r_5	r_6	r_7	r_8	r_9
-0.102	-0.657	-0.060	0.869	-0.089	-0.635	-0.054	0.832	-0.108

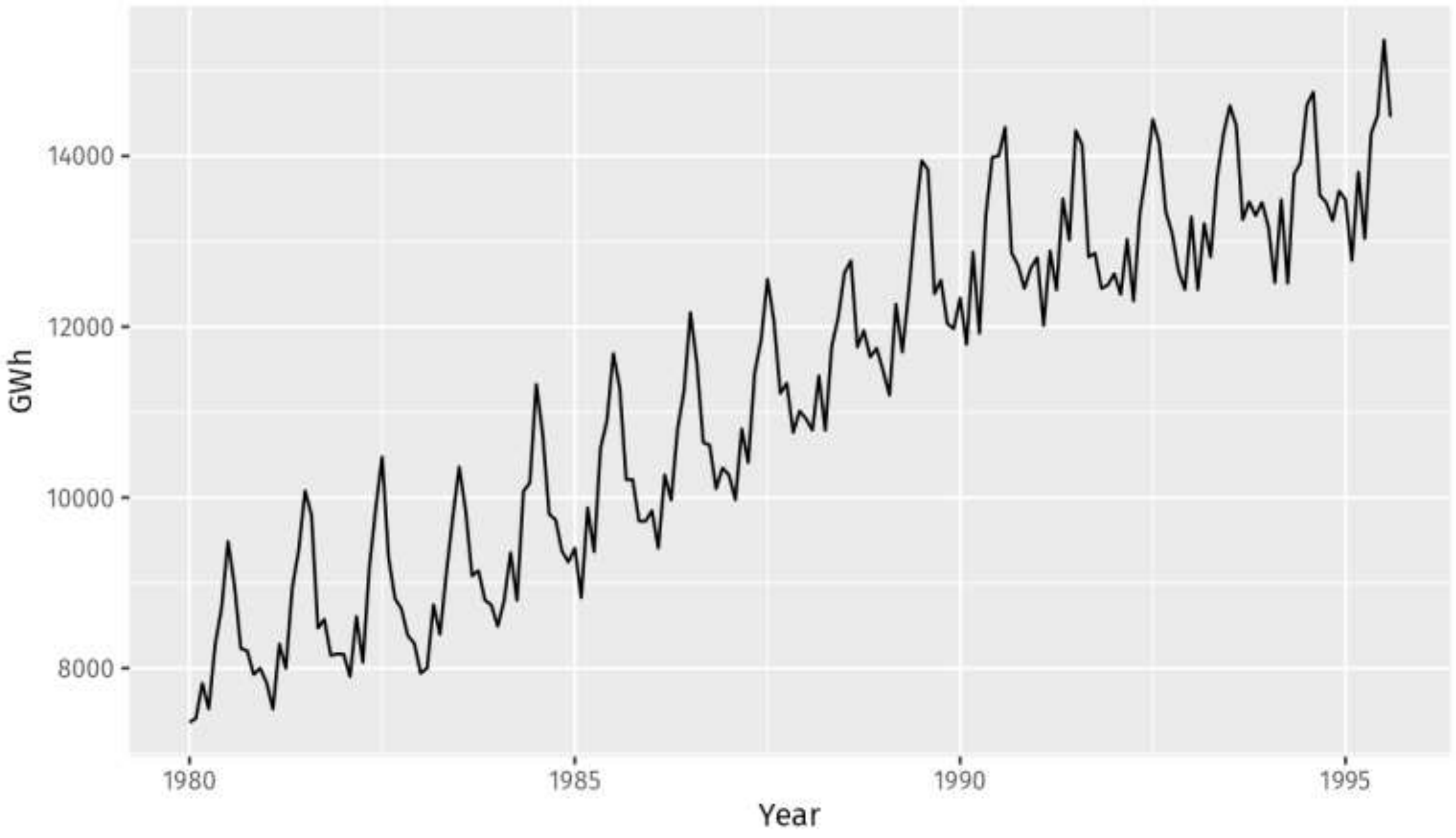
- These correspond to the nine scatterplots in Figure 2. The autocorrelation coefficients are plotted to show the *autocorrelation function* or ACF. The plot is also known as a *correlogram*.
- ggAcf(beer2)
- 

Autocorrelation function of quarterly beer production.

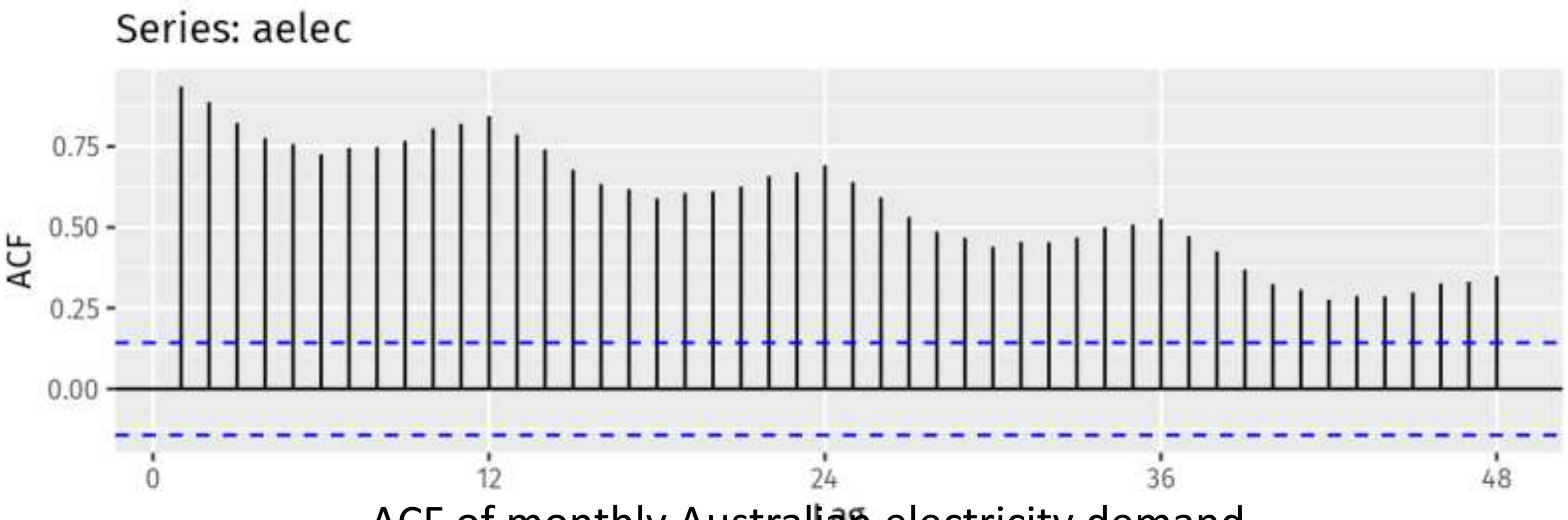
Trend and seasonality in ACF plots

- When data have a trend, the autocorrelations for small lags tend to be large and positive because observations nearby in time are also nearby in size. So the ACF of trended time series tend to have positive values that slowly decrease as the lags increase.
- When data are seasonal, the autocorrelations will be larger for the seasonal lags (at multiples of the seasonal frequency) than for other lags.
- When data are both trended and seasonal, you see a combination of these effects. The monthly Australian electricity demand series plotted in Figure shows both trend and seasonality. Its ACF is shown in Figure
- aelec <- window(elec, start=1980)
- autoplot(aelec) + xlab("Year") + ylab("GWh")

Monthly Australian electricity demand from 1980–1995.



- ggAcf(aelec, lag=48)



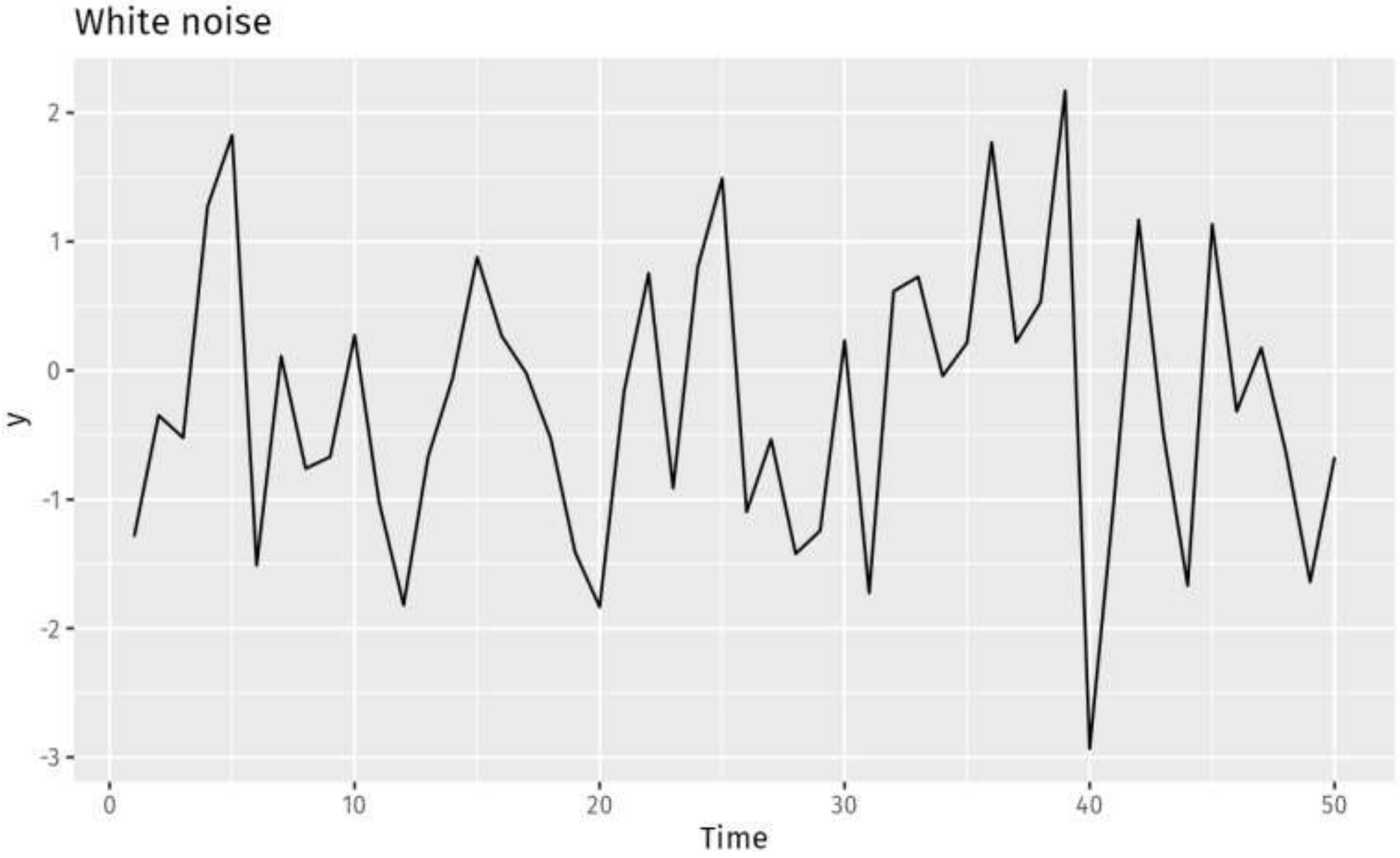
ACF of monthly Australian electricity demand.

The slow decrease in the ACF as the lags increase is due to the trend, while the "scalloped" shape is due to the seasonality.

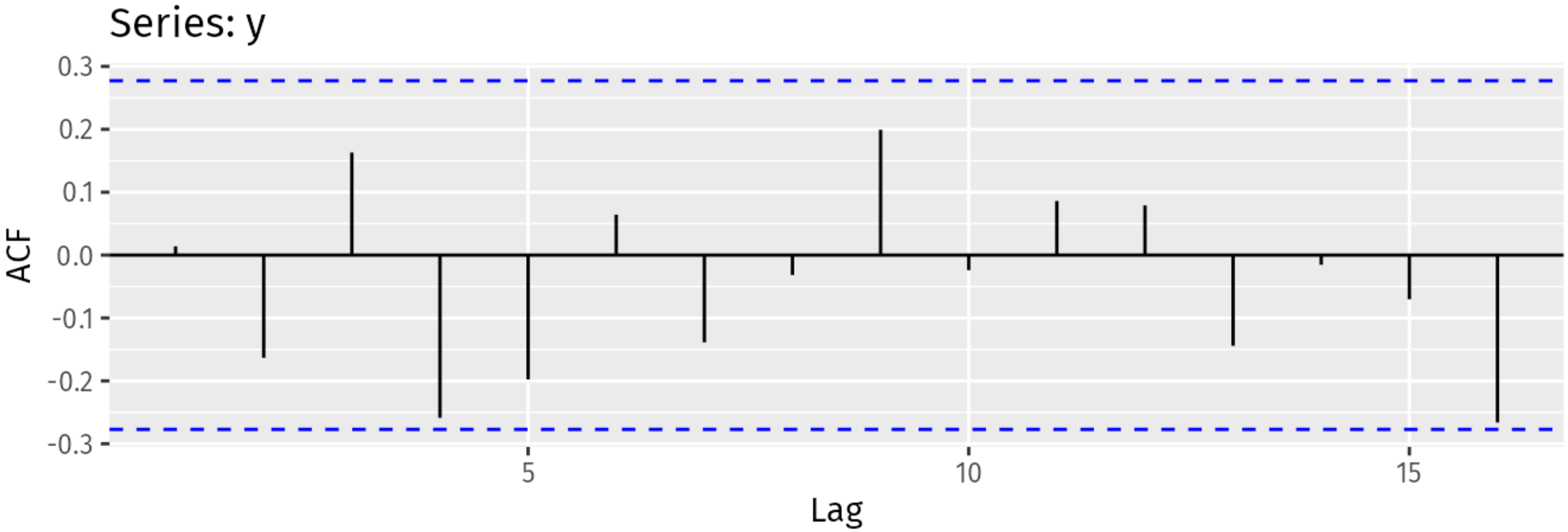
White noise

- Time series that show no autocorrelation are called **white noise**.
- set.seed(30)
- y <- ts(rnorm(50))
- autoplot(y) + ggtitle("White noise")

A white noise time series



```
ggAcf(y)
```



Autocorrelation function for the white noise series.

Naïve Approach

- In this approach, the forecast for the current period is the value of the previous observation of the time series. This approach to forecasting has found wide use due to its simplicity. It can be used with a time series that may be stable, has seasonal variations or has a trend component. In a project situation, this approach, in the absence of any other information, could be used for predicting the number of staff available to perform activities in the next reporting period. Also, in the case of a resource scheduling routine in use with a reporting period as short as one week, the naive forecast may be the most appropriate forecasting method for planning next week's work and allocating staff to tasks. Using this approach, the forecast for period *t*+1 is,
- $F_{t+1} = A_t$, Where
- F_{t+1}, is the forecast value for period *t*+1, and A_t, is the actual value at time *t*.

Simple Averages

- In simple averages, the next period's forecast is the average of all previous actual values.

$$F_{t+1} = (\sum_{t=1}^{n} A_t) /n = (A_1 + A_2 + A_3 + \cdots\cdots\cdots\cdots + A_{n-1} + A_n) /$$

as the data series becomes increasingly long, it will become increasingly less sensitive to any recent movements in data. It would be most appropriate to use this approach where there are considerable random variations in the observed values but no long term evidence of either a rising or falling trend. The averaging techniques in such cases smooth out the time series as the individual high and lows cancel out each other. Consequently, the forecast value over time will become increasingly stable. The biggest disadvantage, however, is that if trend is present in the time series data, the averaging technique will lag the forecast. In other words, in the presence of an increasing trend, the use of the simple averaging technique will understate the actual value; and in the presence of a negative trend, it will overstate the actual value. Projects, however, mostly encounter situations that are not usually stable and hence this method might not be an appropriate forecasting technique for a typical project situation.

Moving Averages

- In this method the next period's forecast is the average of the previous *n* actual values.
- F_{t+1} = ∑actual data values for n previous periods / *n*)
- i.e., $F_{t+1} = (A_t + A_{t-1} + A_{t-2} +$ -------------- $+A_{t-(n-1)}) / n$

- With this method the assumption is that the most recent events are the best indicators of the future with significant random fluctuations in the time series. This approach produces a moving average that is relatively more sensitive to recent movements in data and forecast responsiveness can be increased by reducing the value *n*. As this method uses only the most recent periods that are relevant, it greatly reduces the problem of forecast lag inherent in the simple averaging technique. The choice of the number of data values to be included in the moving average is arbitrary and is left to the judgment of the forecaster.
- It should be noted, however, that while the moving average method uses the data from most recent periods, it still assigns equal importance to all periods of data included in the base of the moving average. Consequently, even with this method there is bound to be some forecast lag. This problem can be resolved to a certain extent by using an extension of the moving average called the **weighted moving average.** In this method, the forecaster assigns more weight to most recent values in the time series. For example, the most immediate observation might be assigned a value of 0.5, the next most recent value a weight of 0.3, and so on. The sum of the weights, however, should be equal to 1. For example, the forecast using a weighted moving average with four recent periods (n = 4) using weights of $w_1 = 0.5$, $w_2 = 0.3$, $w_3 = 0.2$, $w_4 = 0.1$, is given by:
- $F_{t+1} = F_5 = w_1A_4 + w_2A_3 + w_3A_2 + w_4A_1 = 0.5A_4 + 0.3A_3 + 0.2A_2 + 0.1A_1$

Trend-Adjusted Exponential Smoothing

- The exponential smoothing approach discussed above is an appropriate forecasting technique, if the time series exhibits a horizontal pattern (i.e. No trend) with random fluctuations. However, if the time-series exhibits trend, forecasts based on simple exponential smoothing will lag the trend. In such cases, a variation of simple exponential smoothing called the trend-adjusted Exponential smoothing can be used as a forecasting technique. "The trend-adjusted forecast (TAF) has two components:
- **A smoothed error**
- **A trend factor**
- **$TAF_t = S_{t-1} + T_{t-1}$, where**
- **S_{t-1} = Previous period smoothed forecast**
- **T_{t-1} = Previous period trend estimate**
- **TAF_t = Current period's trend-adjusted forecast**
- **$S_t = TAF_t + \propto (A_t - TAF_t)$**
- **$T_t = T_{t-1} + \propto(TAF_t - TAF_{t-1} - T_{t-1})$,** where $\propto$ and β are smoothing constants
- In order to use this method, one must select values of and (usually through trial and error) and make a starting forecast and an estimate of the trend" (Stevenson, 2005).

Example of Trend-Adjusted Exponential Smoothing

- For the data given below, generate a forecast for period 11 through 13 using trend-adjusted exponential smoothing. Use $\propto = 0.4$ and $\beta = 0.3$
-

Period	1	2	3	4	5	6	7	8	9	10
Data values	500	524	520	528	540	542	558	550	570	575

- **Solution:** To use trend adjusted exponential smoothing, we first need an initial estimate of the trend. This initial estimate can be obtained by calculating the net change from the three changes in the data that occurred through the first four periods.
- Initial Trend Estimate = (528 - 500)/3 = 28/3 = 9.33
- Using this initial trend estimate and the actual data value for period 4, we compute an initial forecast for period 5.
- Initial Forecast for period 5 = 528 + 9.33 = 537.33.
- The forecasts and the associated calculations are shown in the table below.

Period	Actual	$S_{t-1} + T_{t-1} = TAF_t$	$TAF_t + 0.3(A_t - TAF_t) = S_t$	$T_{t-1} + 0.2(TAF_t - TAF_{t-1} - T_{t-1}) = T_t$
5	540	528 + 9.33 = 537.33	537.33 + 0.3(540 - 537.33) = 538.13	9.33
6	542	538.13 + 9.33 = 547.46	547.46 + 0.3(542 - 547.46) = 545.82	9.33 + 0.2(547.46 - 537.33 - 9.33) = 9.49
7	558	545.82 + 9.49 = 555.31	555.31 + 0.3(558 - 555.31) = 556.12	9.49 + 0.2(555.31 - 547.46 - 9.49) = 9.16
8	550	556.12 + 9.16 = 565.28	565.28 + 0.3(550 - 565.28) = 560.70	9.16 + 0.2(565.28 - 555.31 - 9.16) = 9.32
9	570	560.70 + 9.32 = 570.02	570.02 + 0.3(570 - 570.02) = 570.01	9.32 + 0.2(570.02 - 565.28 - 9.32) = 8.41
10	575	570.01 + 8.41 = 578.42	578.42 + 0.3(575 - 578.42) = 577.40	8.41 + 0.2(578.42 - 570.02 - 8.41) = 8.41
11		577.40 + 8.41 = 585.81		

The forecast for period 11 is 585.81.

- https://otexts.com/fpp2/graphics-exercises.html