Kristianstad University

Data Science II - Visualization

Lecture 7th of June 2024

Fredrik Frisk

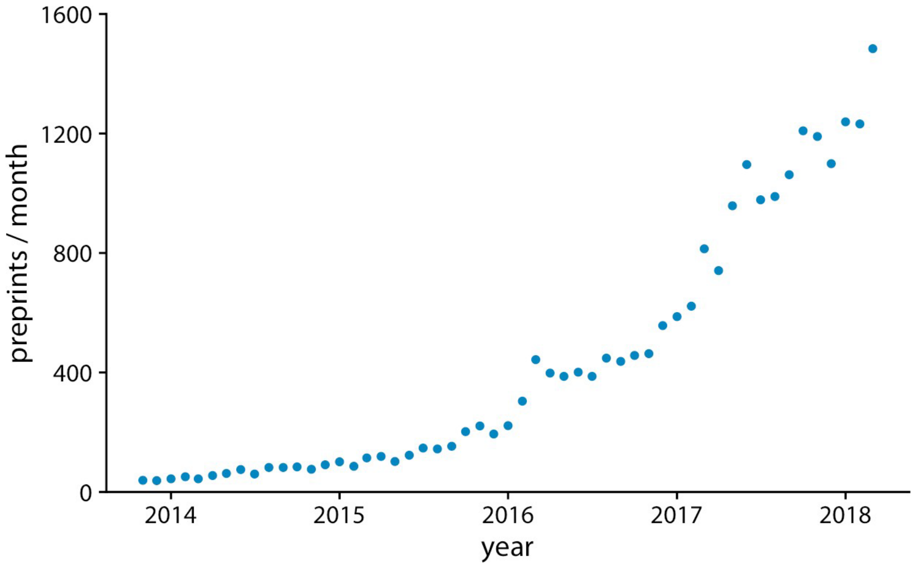
Kristianstad University

Some quizes may pop up

Time Series

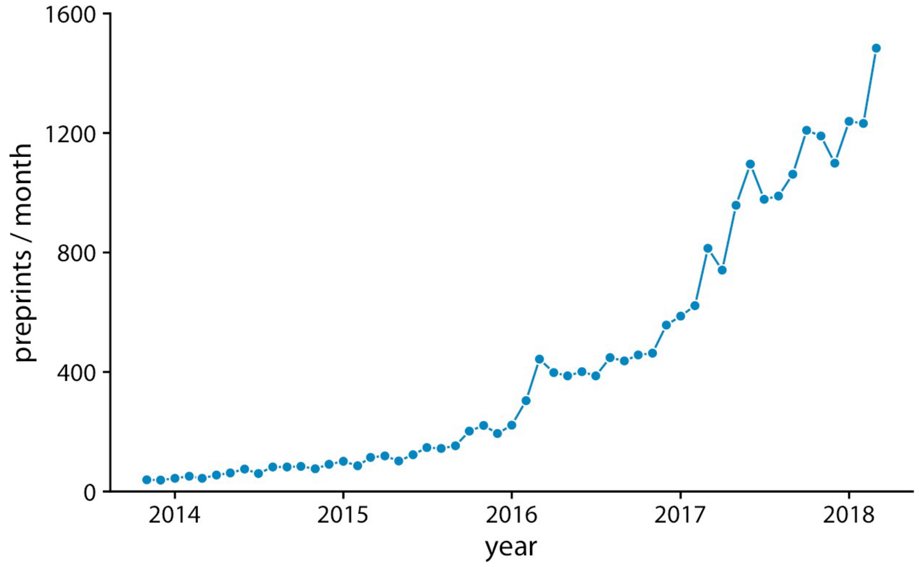
- in scatterplots we plotted one quantitative variable against another
- time series: one variable represents time and time imposes additional structure on the data — the data points are ordered in time
- we can visualize this order with line graphs
    - we can use line graphs whenever one variable imposes an ordering on the data (not limited to time series)
- as an alternative we can use a scatterplot and draw lines to connect the neighboring points in time

Time Series Visualization with Scatterplots



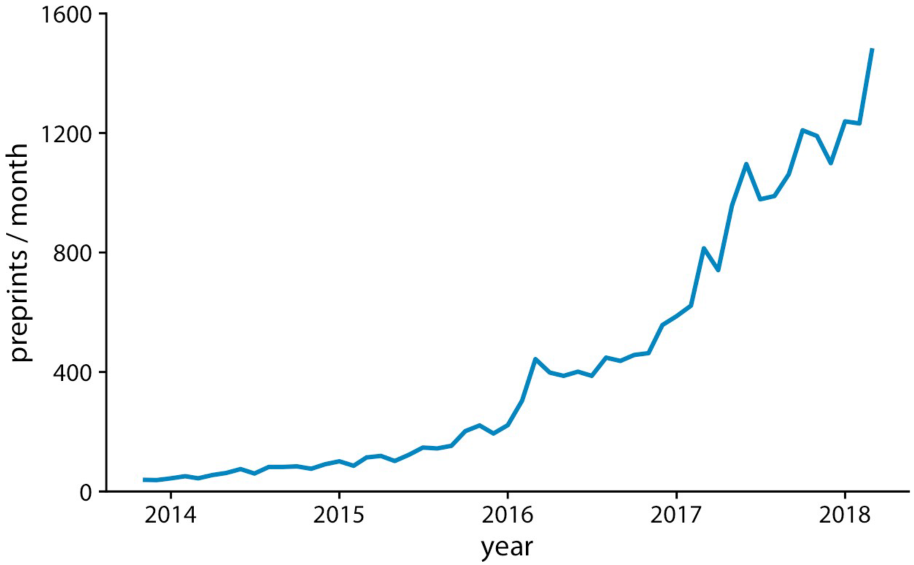
- notice how the dots are spaced evenly along the x-axis
- there is a defined order among the dots

Time Series Visualization with Scatterplots



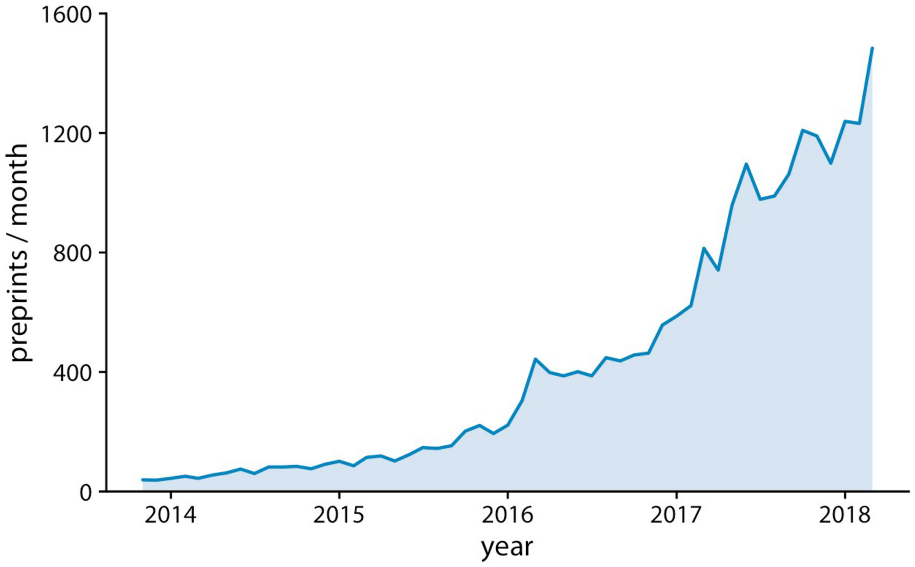
- the order is emphasized by connecting neighboring dots with lines
- note that lines represent made-up data (there are no data measurements at intermediate times), yet they may help with perception
- using lines to represent time series data is *generally accepted practice*, however dots are often omitted althogether

Time Series Visualization with Scatterplots



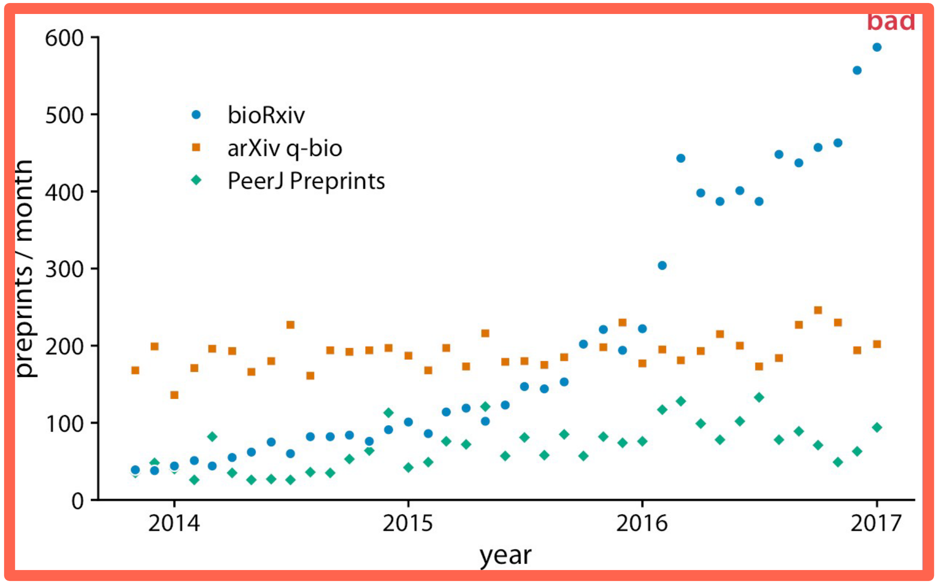
- line graph without dots
- omitting the dots emphasizes the overall temporal trend
- useful when time points are spaced very densely

Time Series Visualization with Scatterplots



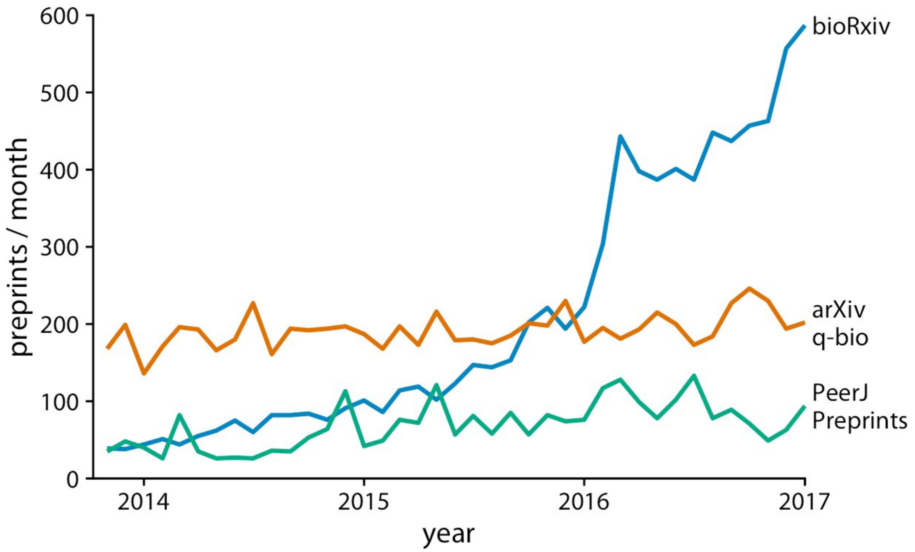
- area under the curve filled with solid color
- further emphasizes the trend in data by separating the area above and below the curve

Visualizing Multiple Time Series



- a scatterplot for multiple time series is not a good idea
- figure can become confusing and difficult to read

Visualizing Multiple Time Series



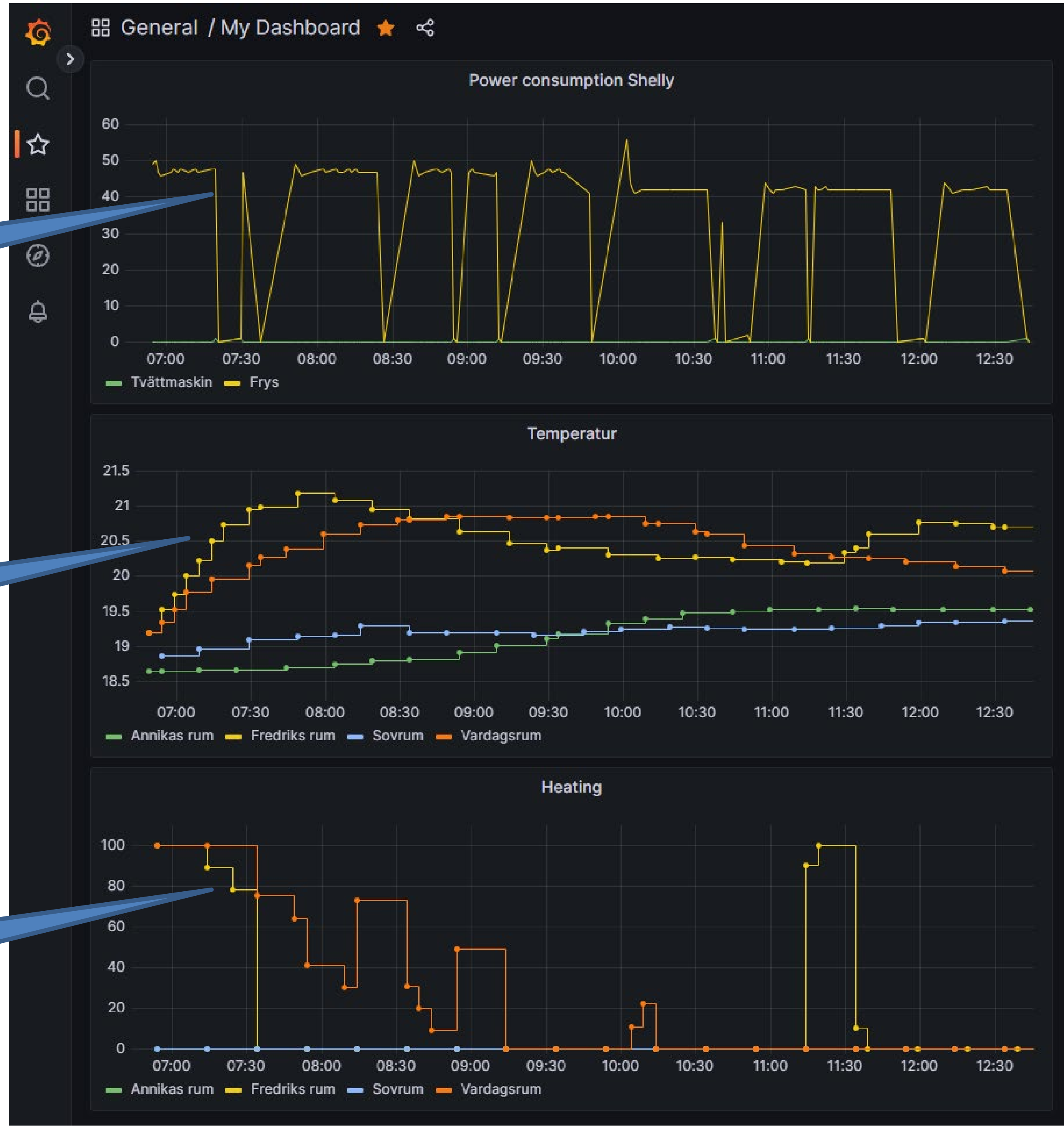
- connecting the dots (and omitting the dots altogether) helps with perception

Types of Lineplots

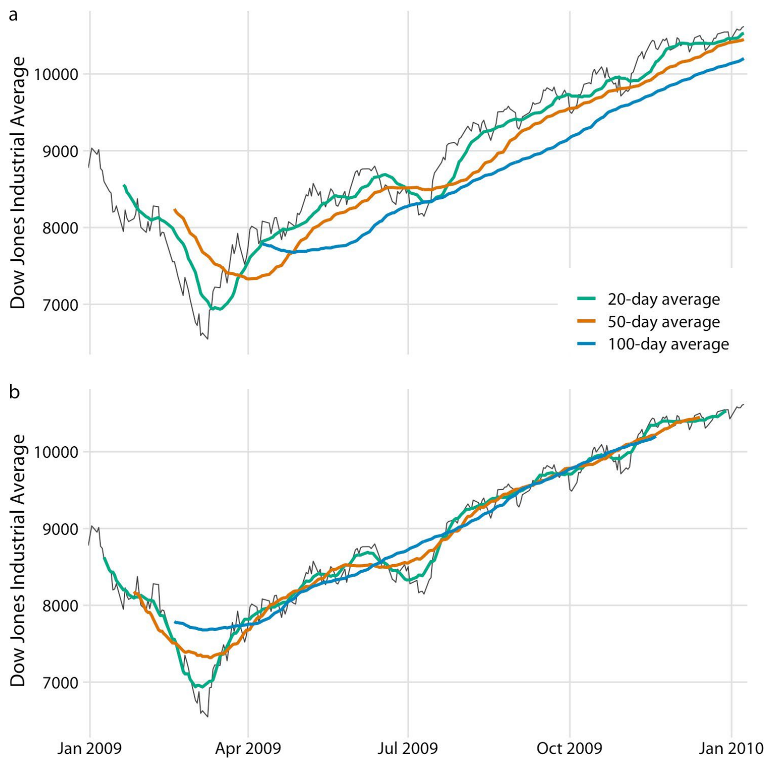
Interpolation

”Real” data only

”Real” data only

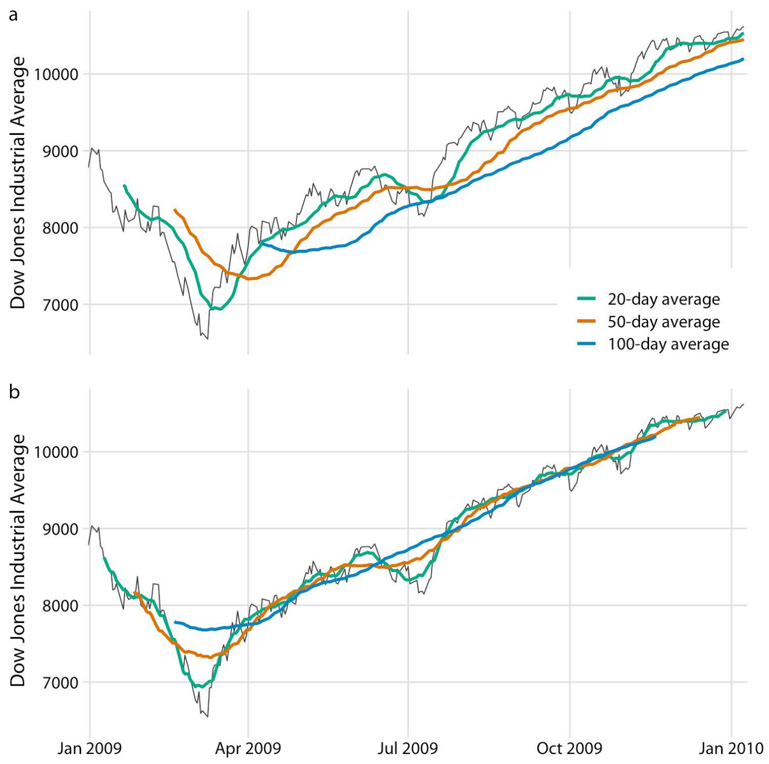


Visualizing Trends



- in lineplots or scatterplots we are often interested in the overarching trend of the data
- with a visualization of the trend on top of or instead of the actual data points we can help the reader see the key features of the data
- we want to visualize longer-term trends while deemphasizing the less important short-term fluctuations

Moving Average for Smoothing

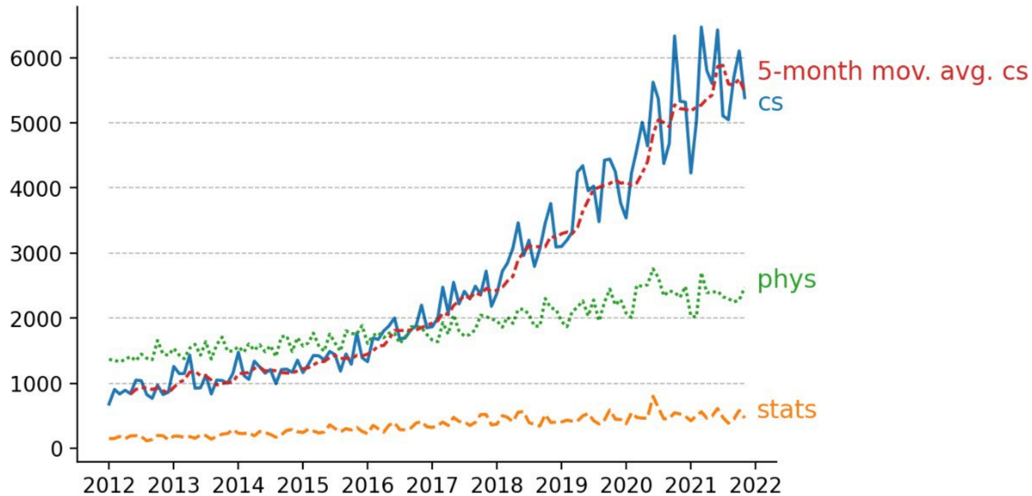


- to generate a moving average, we take a time window and calculate the average over that window, then we move that time window by one time-unit and repeat that procedure

- to plot this sequence of moving averages, we need to associate a specific time point with the average of each time window

  - usually the end or the center of the window are chosen

Moving Average for Smoothing



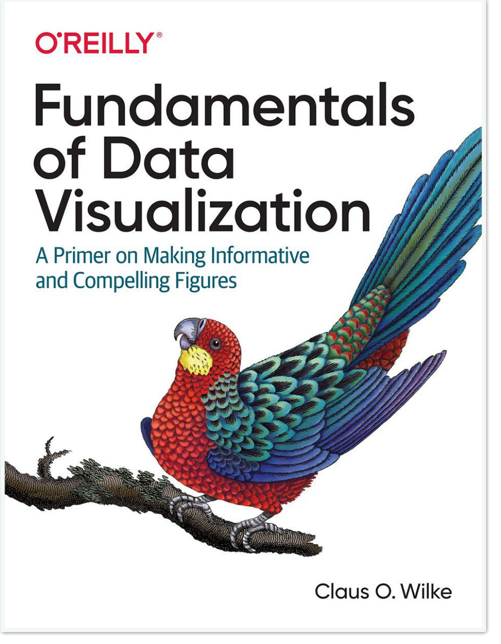
- compute the 5-month moving average for the cs related submissions to arxiv time series using

  pandas.DataFrame.rolling

- plot your result using `matplotlib`

- moving average is just one method for smoothing; there are other like LOWESS (implemented in statsmodels) that you can evaluate for your visualization

Literature

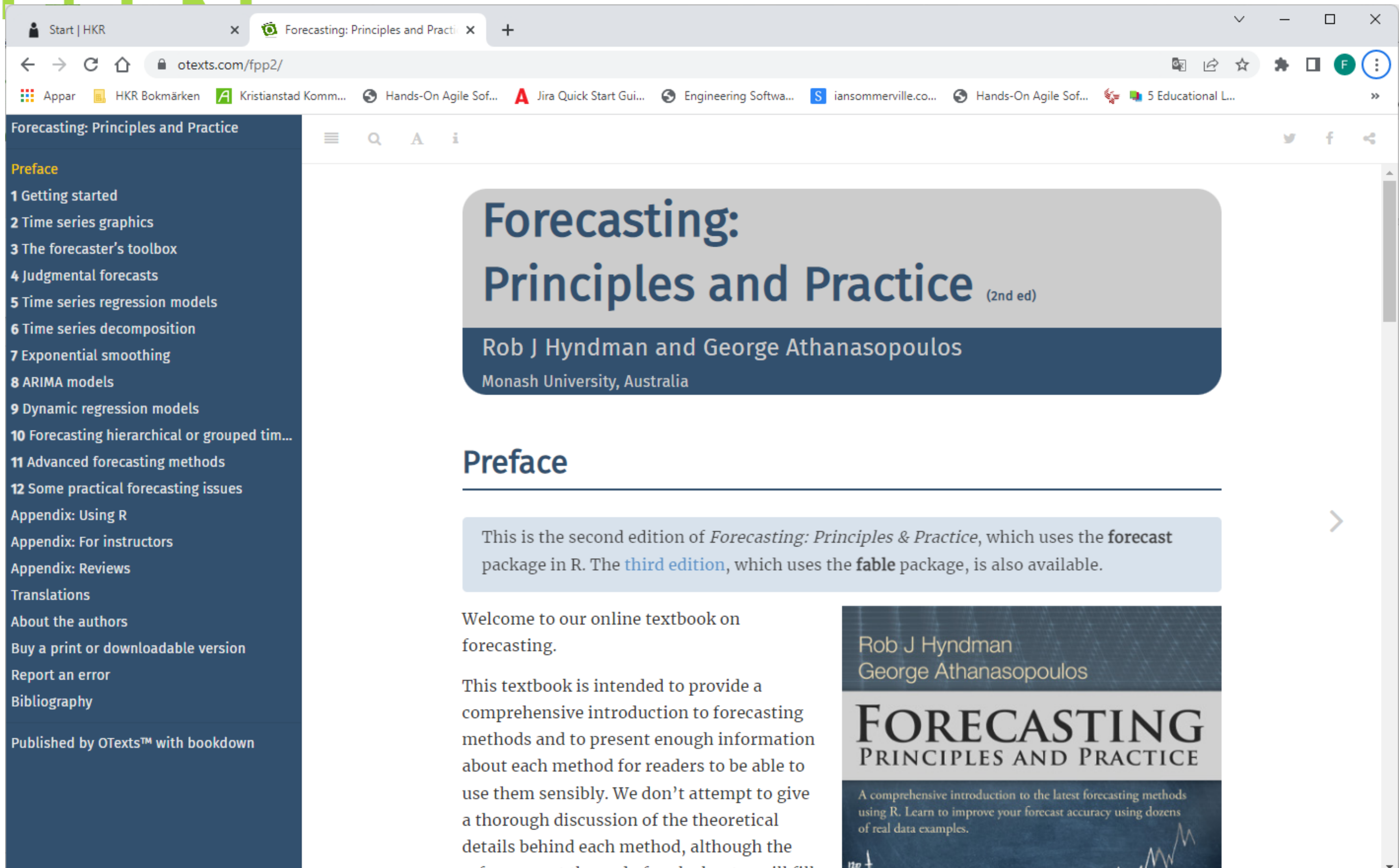

O'REILLY
Fundamentals of Data Visualization
A Primer on Making Informative and Compelling Figures
Claus O. Wilke

References

- Slide 4-9,13,14; Image Source: Claus O. Wilke - Fundamentals of Data Visualization, O'Reilly

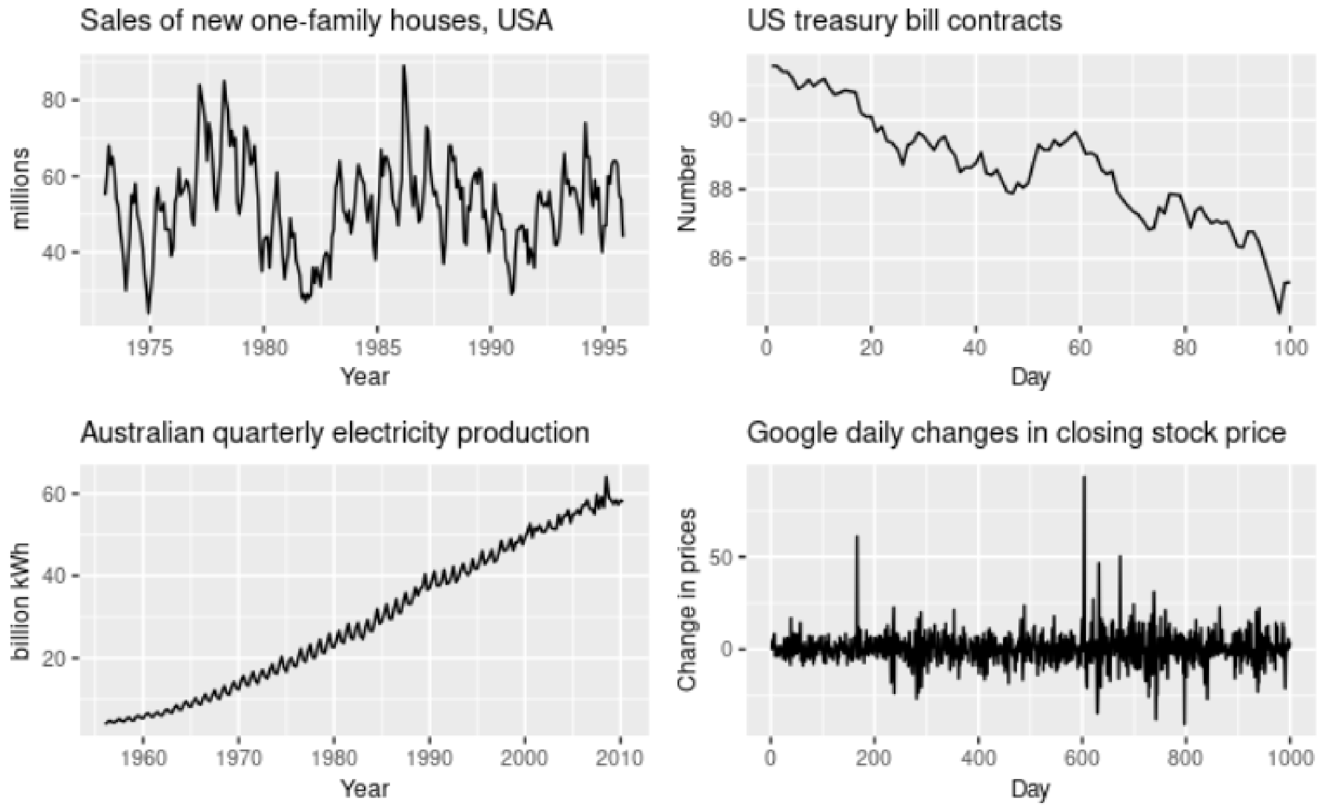
Kristianstad University

Following pictures from



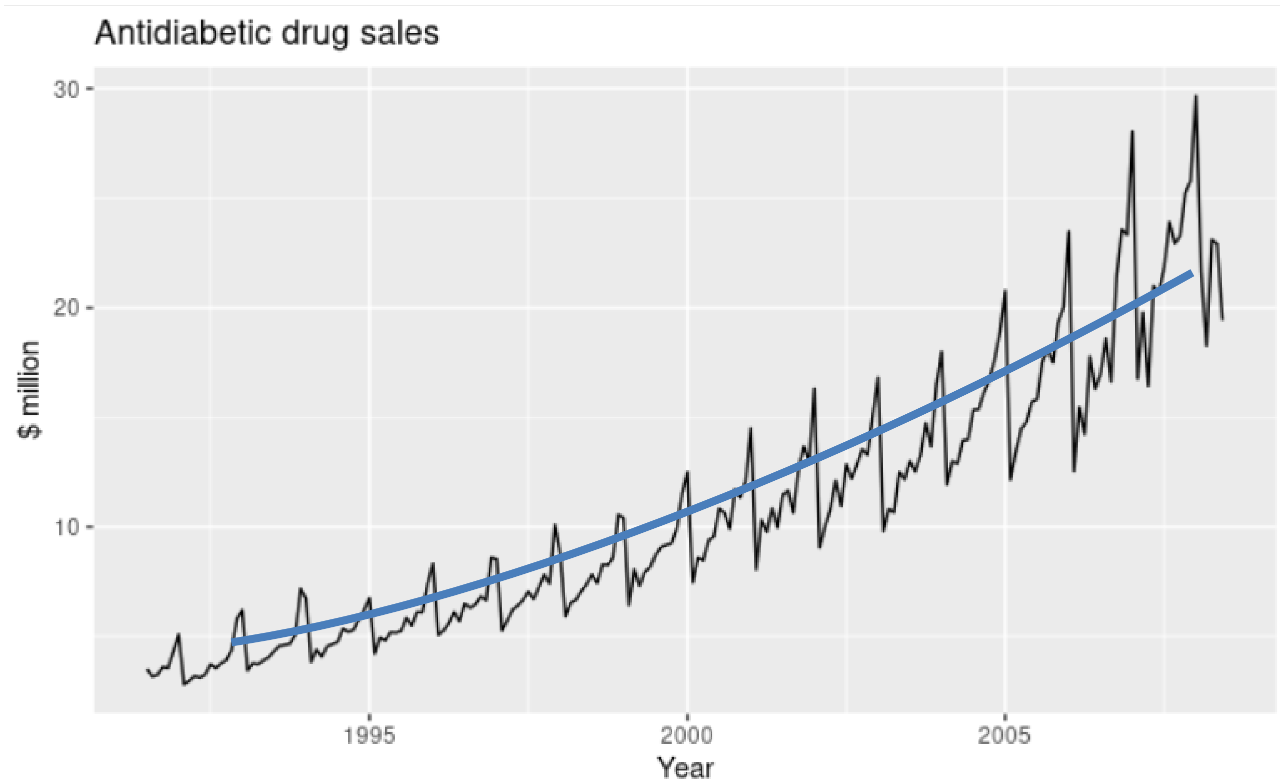
Time series patterns (ch 2.3)

- Trend
- Seasonal
- Cyclic



Trend (from ch 2.3)

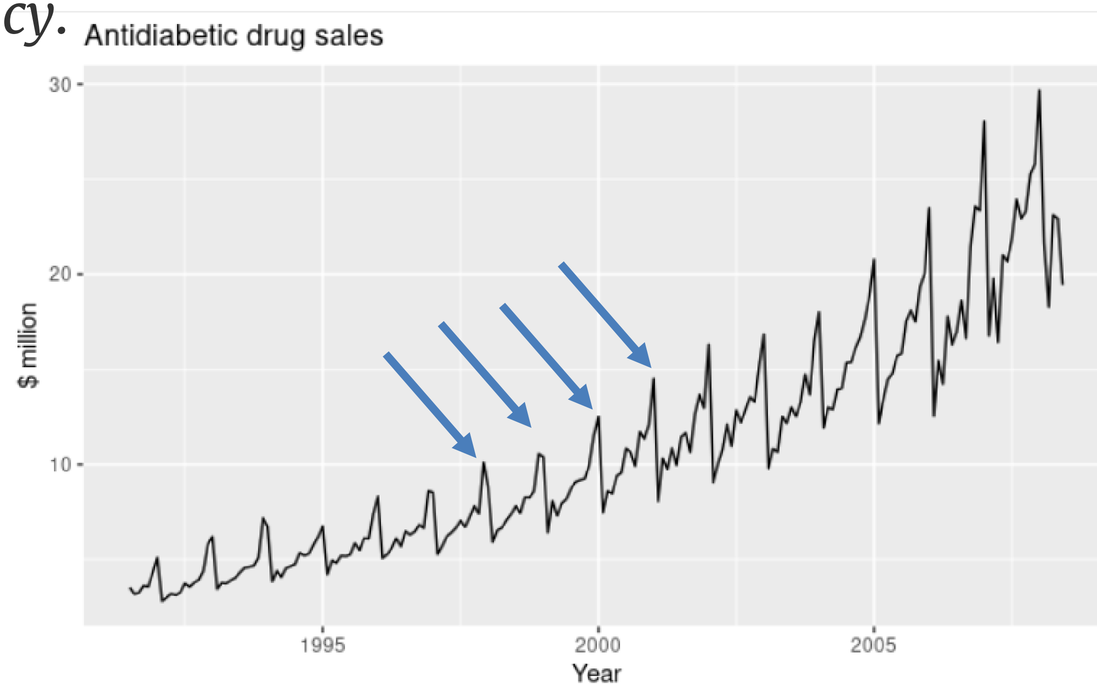
A *trend* exists when there is a long-term increase or decrease in the data.



Kristianstad University

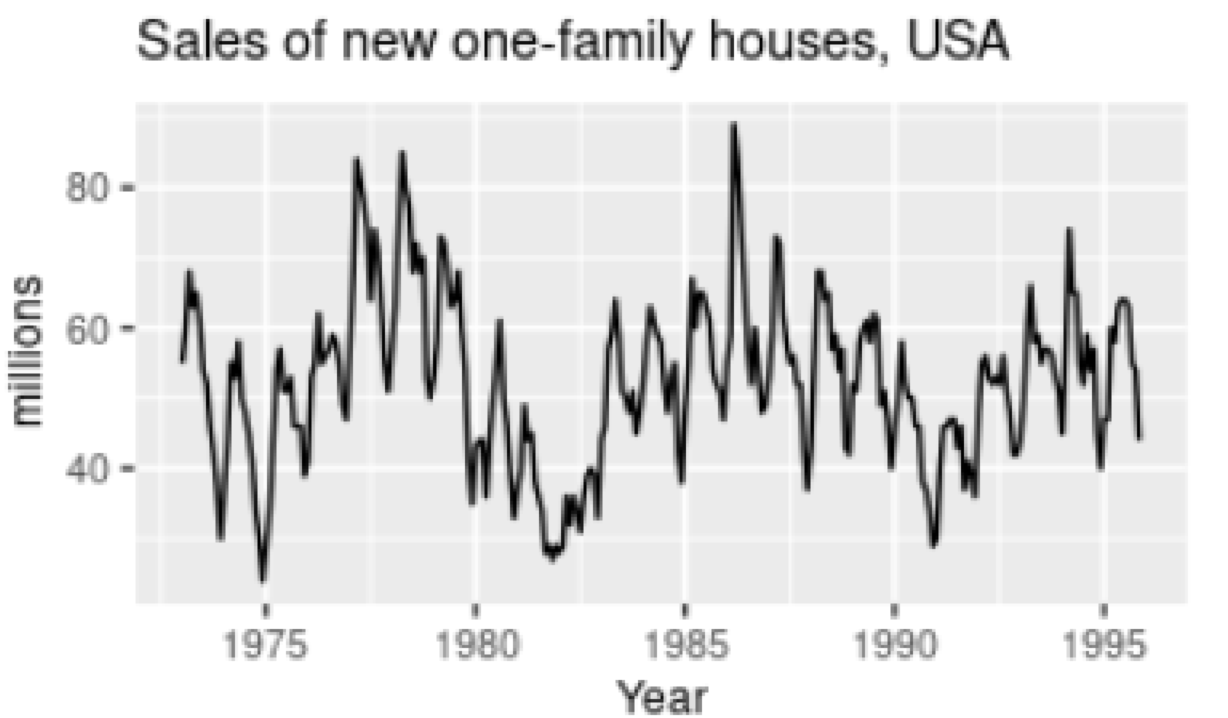
Seasonal (from ch 2.3)

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*.



Cyclic (from ch 2.3)

A *cycle* occurs when the data exhibit rises and falls that are not of a fixed frequency.



Random behaviour (ch. 2.3)

- No trend, seasonality or cyclic behaviour.
- Random fluctuations
- Hard to predict

Kristianstad University

Relations between Time Series
(ch 2.6)

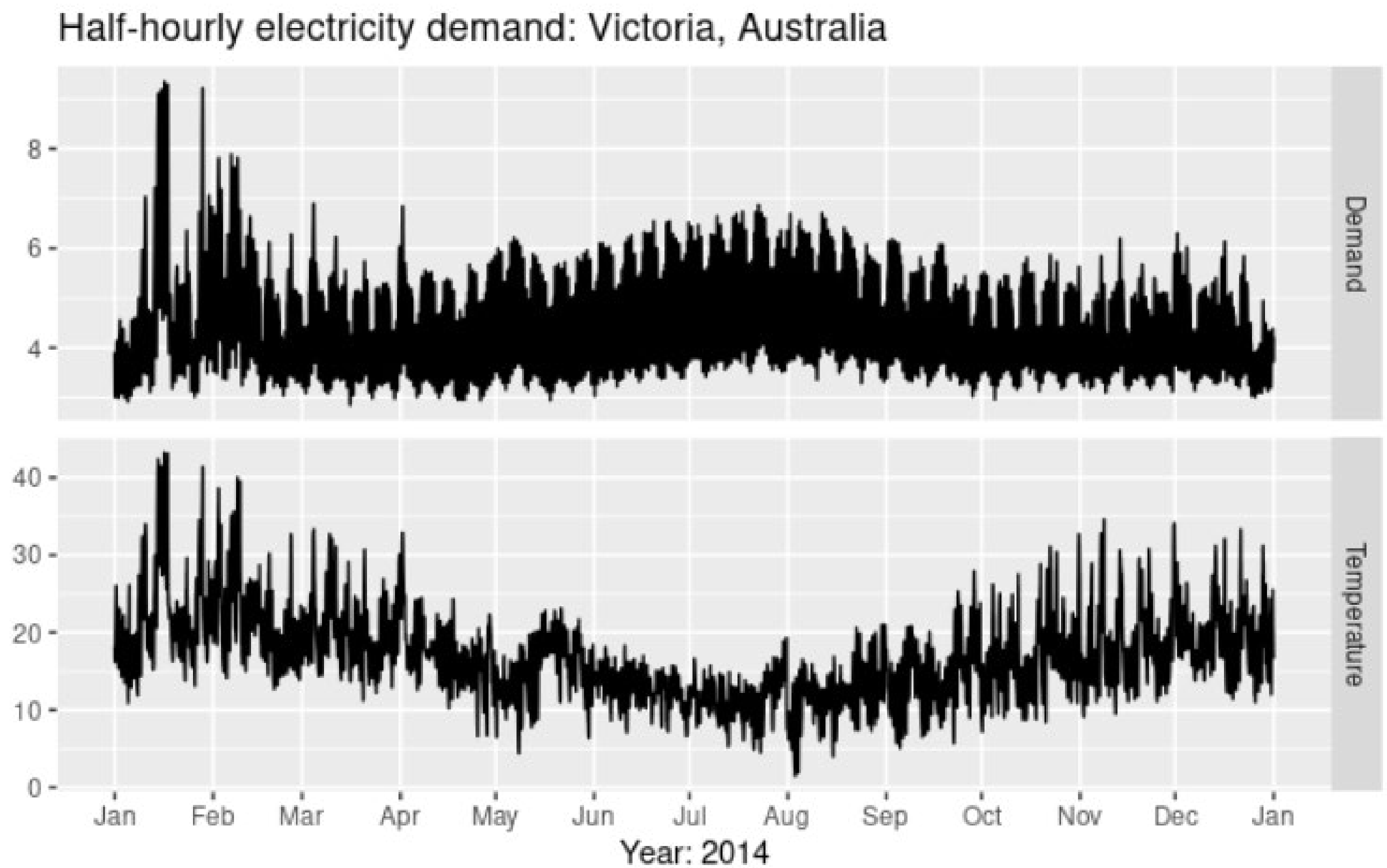


Figure 2.7: Half hourly electricity demand and temperatures in Victoria, Australia, for 2014.

Kristianstad University

Scatterplot (ch 2.6)

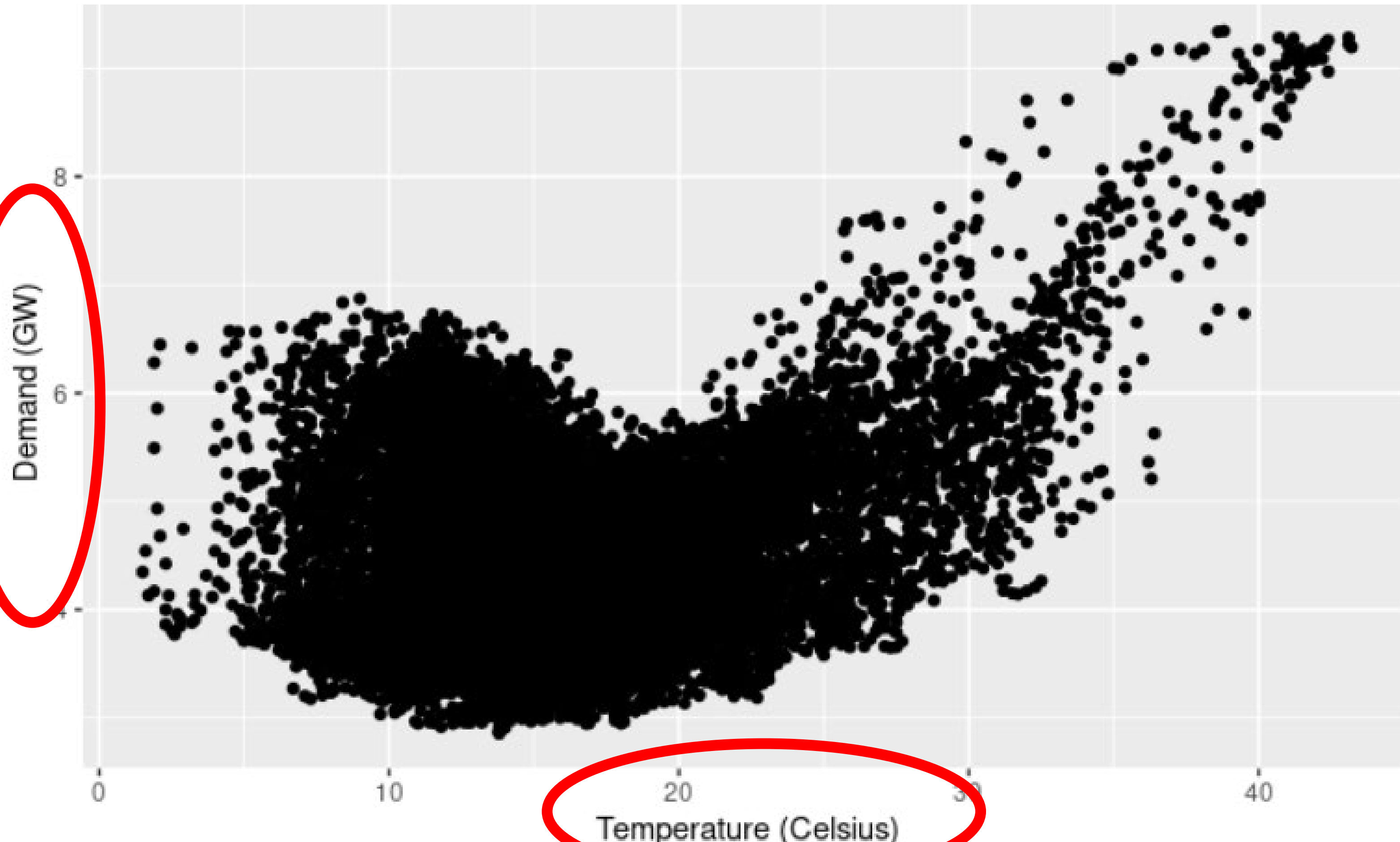
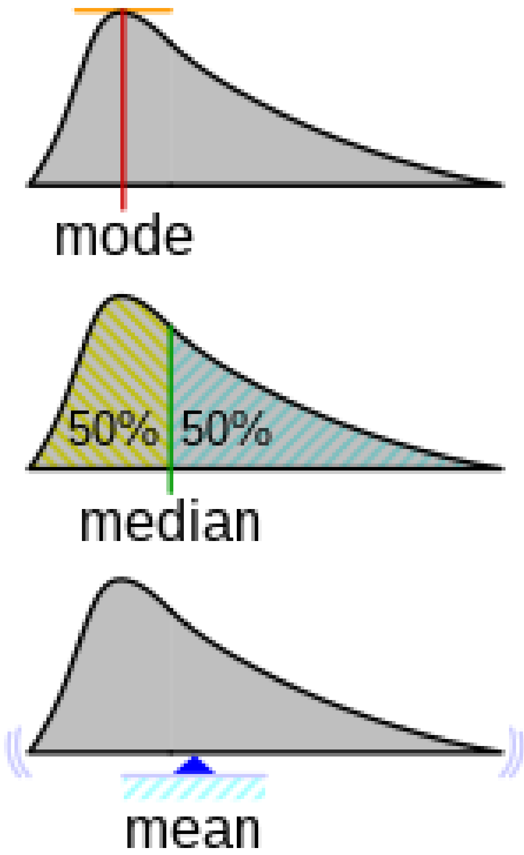


Figure 2.8: Half-hourly electricity demand plotted against temperature for 2014 in Victoria, Australia.

Kristianstad University

Some basic properties of pdf



- How many data points is found in a pdf?

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

From: https://en.wikipedia.org/wiki/Probability_density_function is the sample mean

Population vs. Sample set

Kristianstad University

- Infinite #datapoints

- μ is the population mean
- σ^2 is the population variance

- Finite #datapoints
- $\bar{x}$ is the sample mean
- s^2 is the sample variance

Used by Pandas

From https://en.wikipedia.org/wiki/Bessel's_correction
https://pandas.pydata.org/docs/reference/api/pandas.DataFrame.std.html

Variance

Population

$$\sigma^2 = \sum_i \frac{(x_i - \mu)^2}{n}$$

Sample set

$$s^2 = \sum_{i=1}^{n} \frac{(x_i - \bar{x})^2}{n-1}$$

Standard deviation: σ resp. s

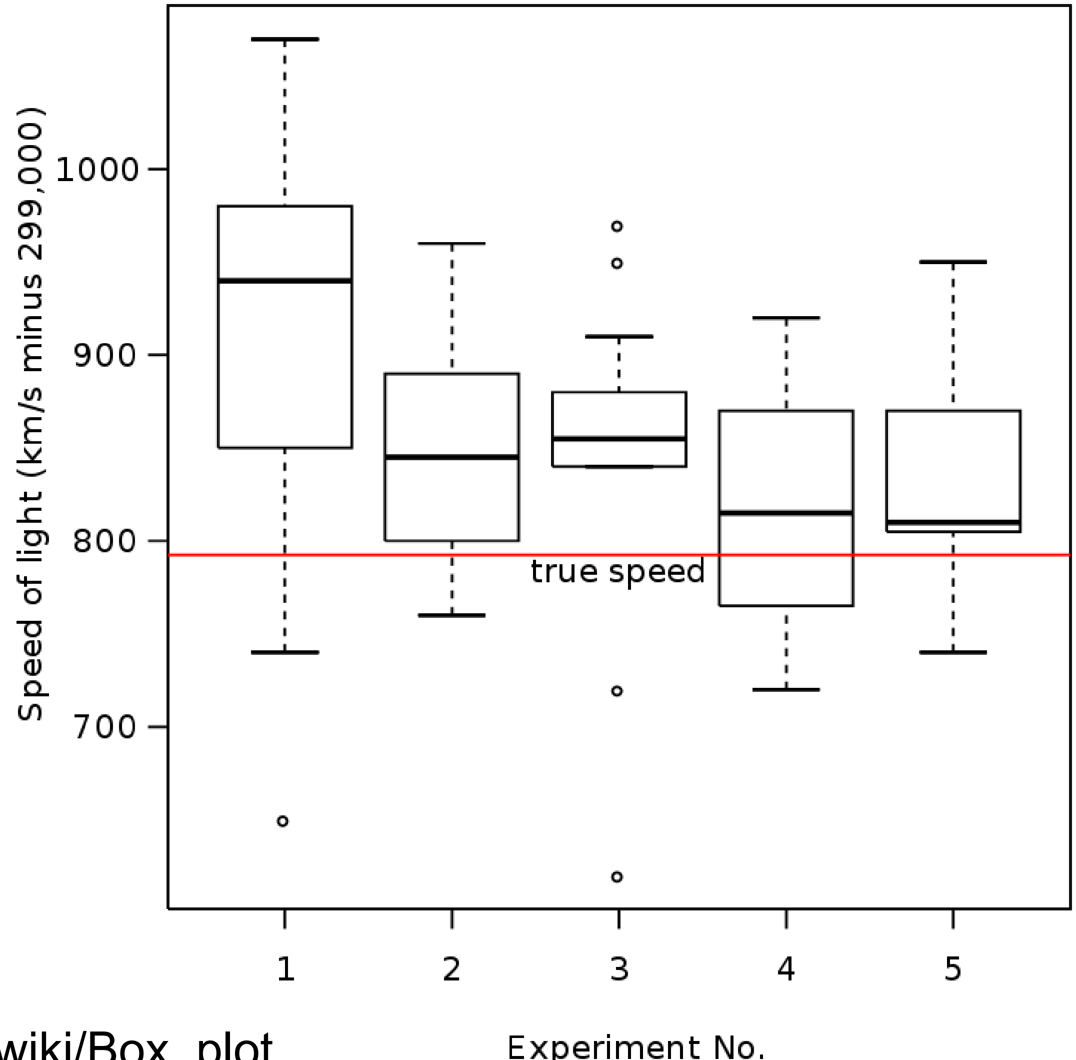
Kristianstad University

Statistics recap

Example of an outlier

Box plot

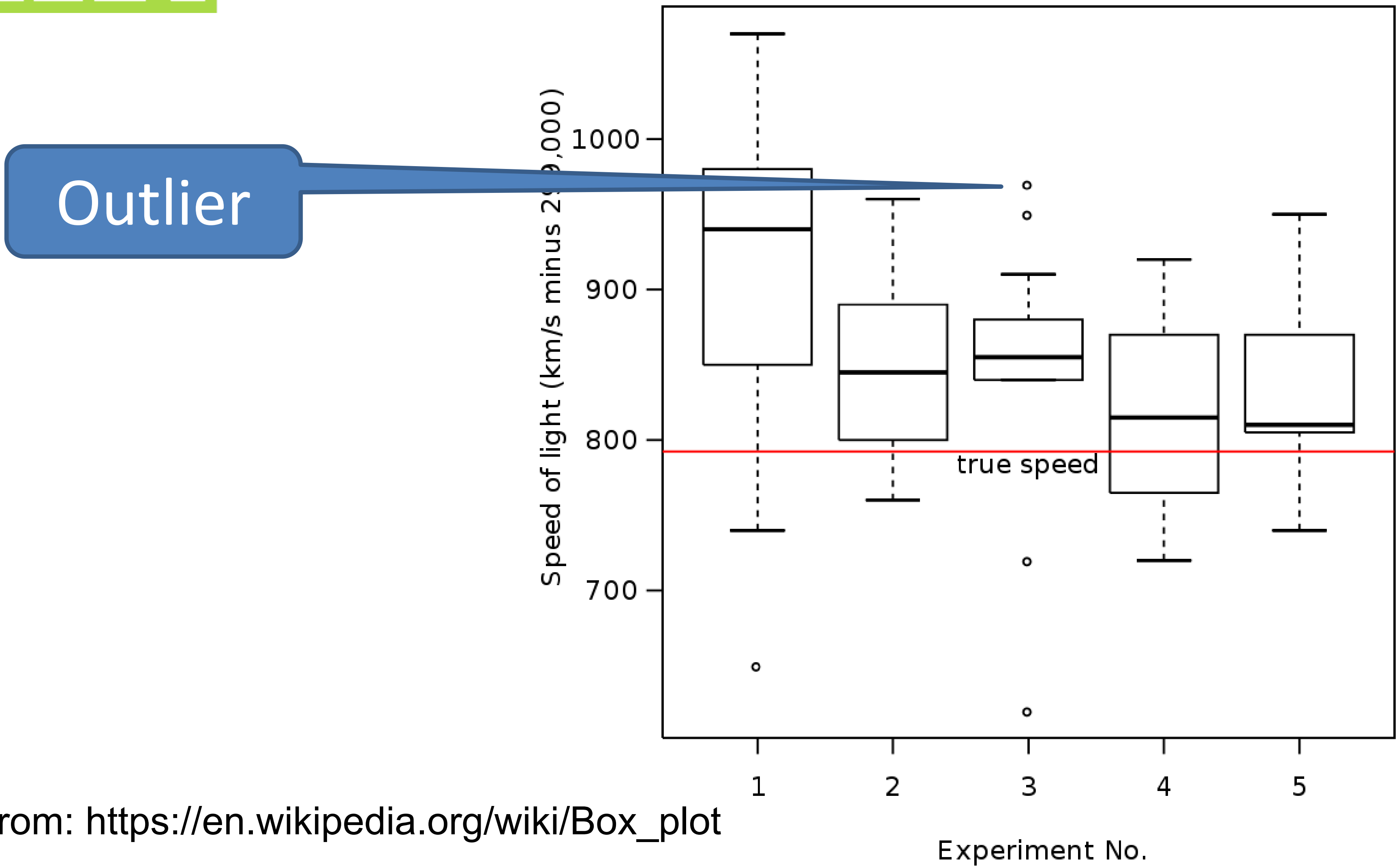
- Median
- First quartile
- Third quartile
- Minimum
- Maximum
- Outliers



From: https://en.wikipedia.org/wiki/Box_plot

Kristianstad University

Box plot



From: https://en.wikipedia.org/wiki/Box_plot

outliers

Run the file – outlier.ipynb

Kristianstad University

NIST

National Institute of Standards and Technology (USA)

Definition of outliers

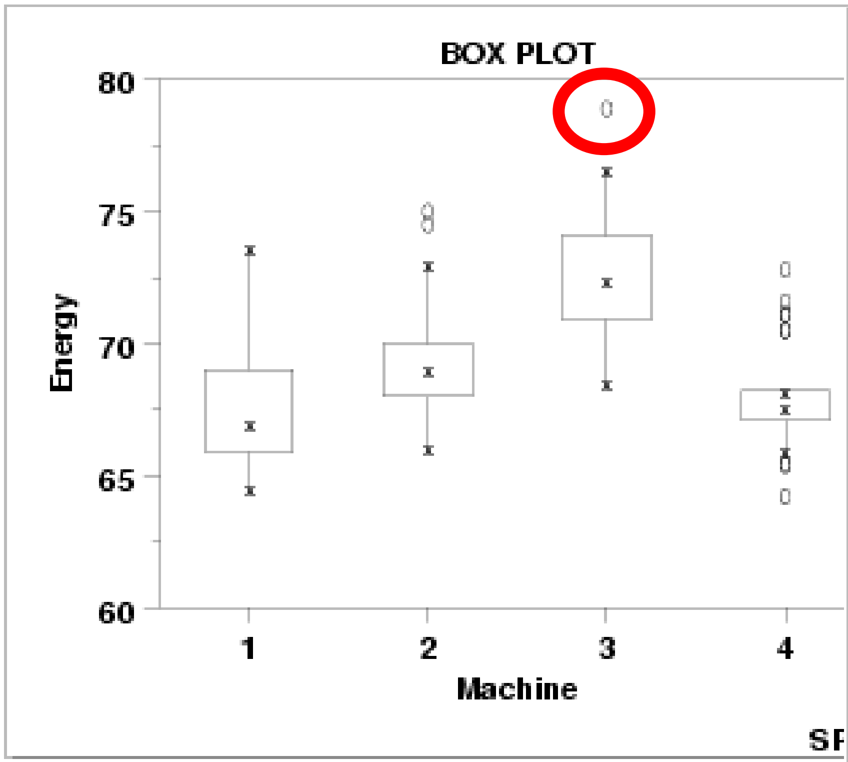
An *outlier* is an observation that lies an ***abnormal distance*** from other values in a random sample from a population. In a sense, this definition leaves it *up to the analyst* (or a consensus process) to *decide what will be considered abnormal*. Before abnormal observations can be singled out, it is necessary to characterize normal observations.

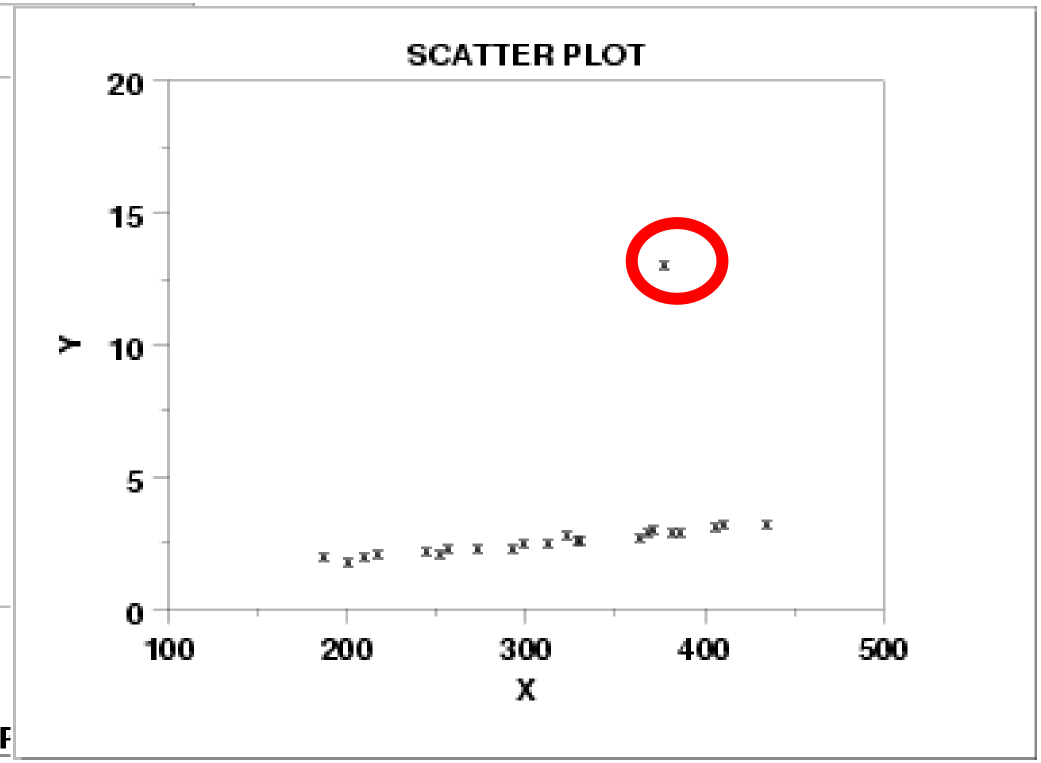
From https://www.itl.nist.gov/div898/handbook/prc/section1/prc16.htm

NIST

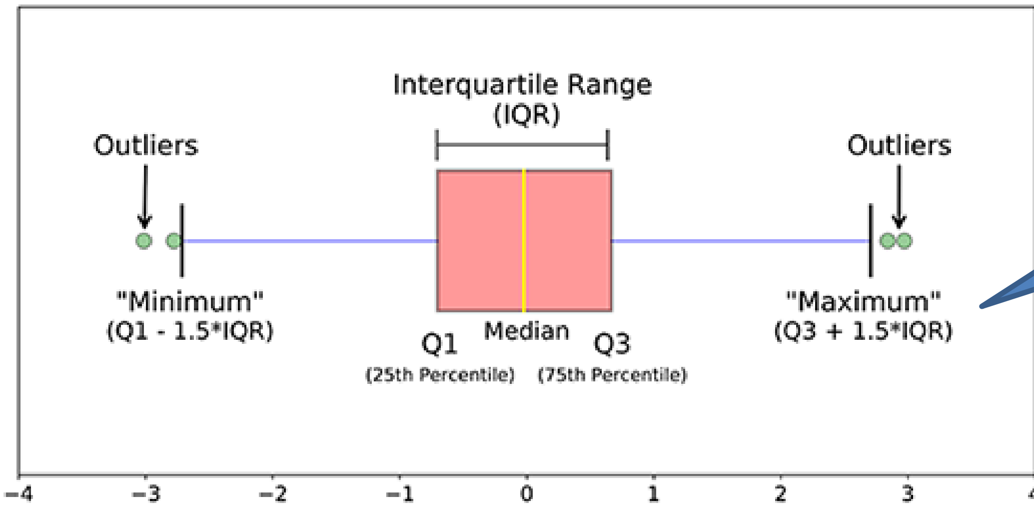
Recommend two graphical methods

- Box plots
- Scatter plots





NIST



Mild outlier

Mild outlier: outside Q3-1.5*IQR or Q3+1.5*IQR

Extreme outlier: outside Q3 - 3*IQR or Q3 + 3*IQR

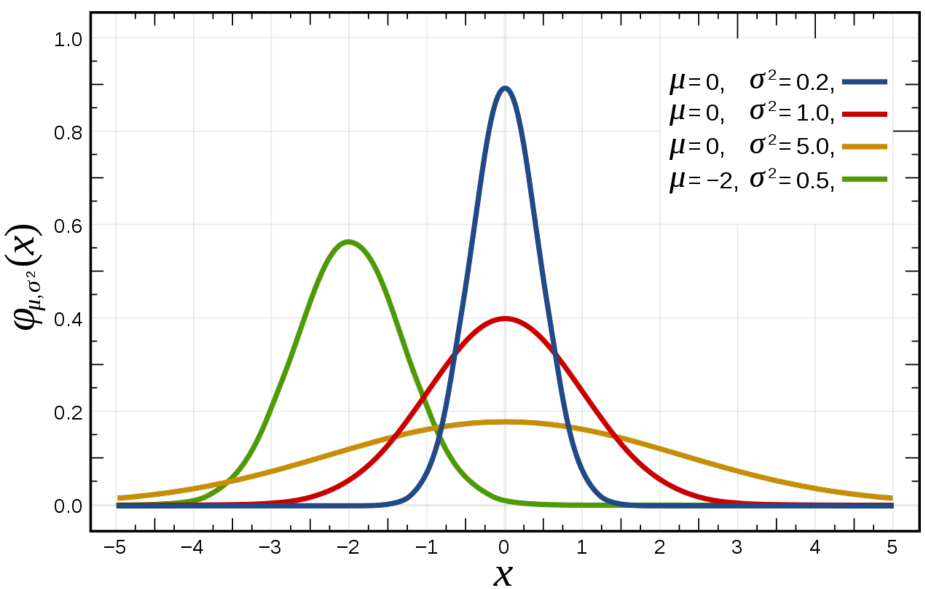
https://www.itl.nist.gov/div898/handbook/prc/section1/prc16.htm

Outliers and Z-score

- Outliers and Z-score:
- Values with z-scores greater than 3 or less than -3 are often considered outliers.
- This means that these values are more than three standard deviations away from the mean.
- Used to identify and analyze unusual data points in a dataset.

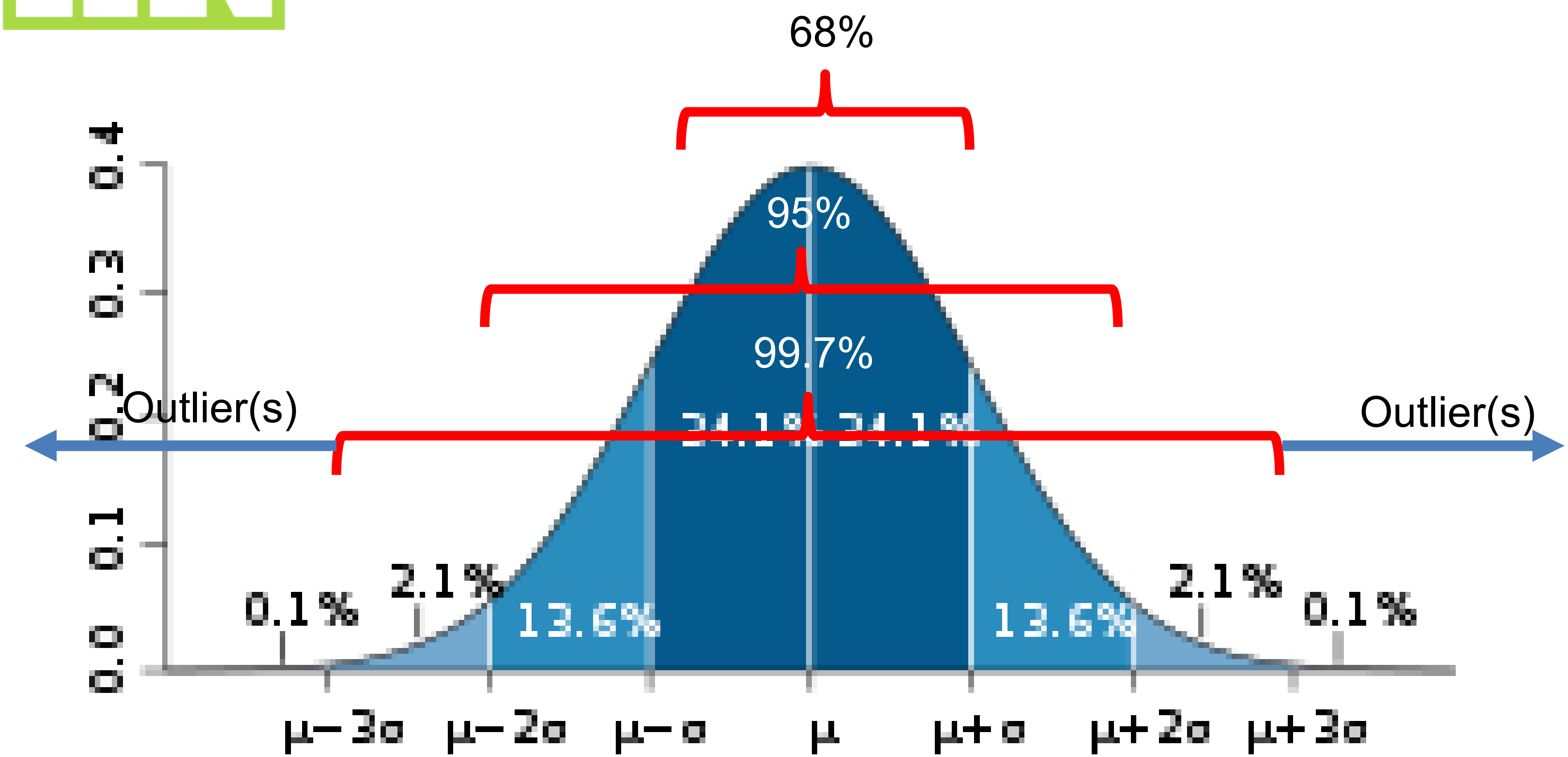
Normal Distribution

$$f(x) = \frac{1}{\sqrt{2\pi}}\, e^{-x^2/2} \qquad f(x;\mu,\sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$



Two parameters:

- Mean - μ
- Standarddeviation - σ



From: https://en.wikipedia.org/wiki/Outlier

The Formula for Z-score

- Formula:
- $Z = (X - \mu) / \sigma$

- X: Individual value
- μ: Mean of the distribution
- σ: Standard deviation

Pearson Correlation coefficient

The correlation between variables x and y is given by

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

Kristianstad University

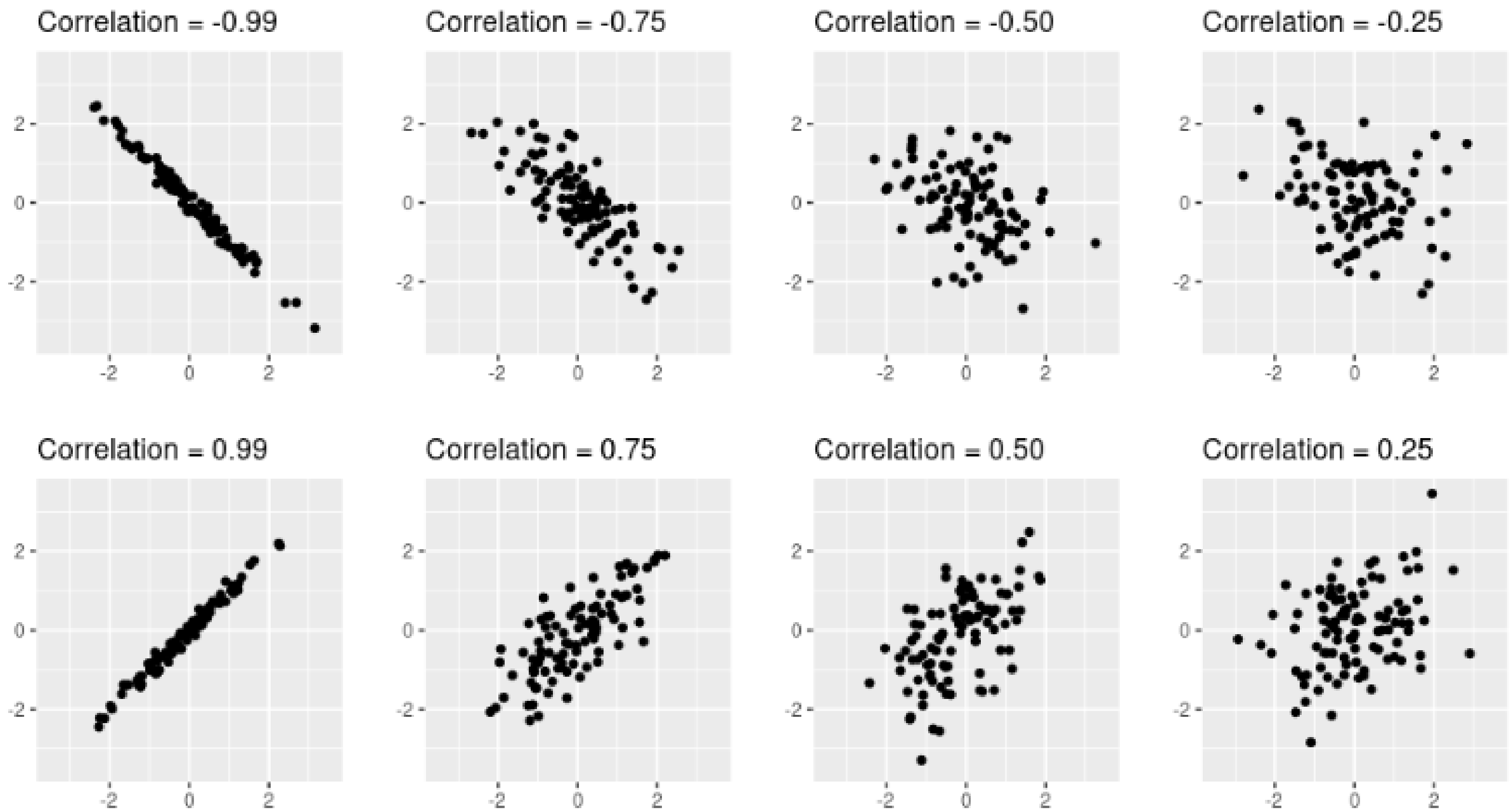


Figure 2.9: Examples of data sets with different levels of correlation.

r = 0.82 for all plots

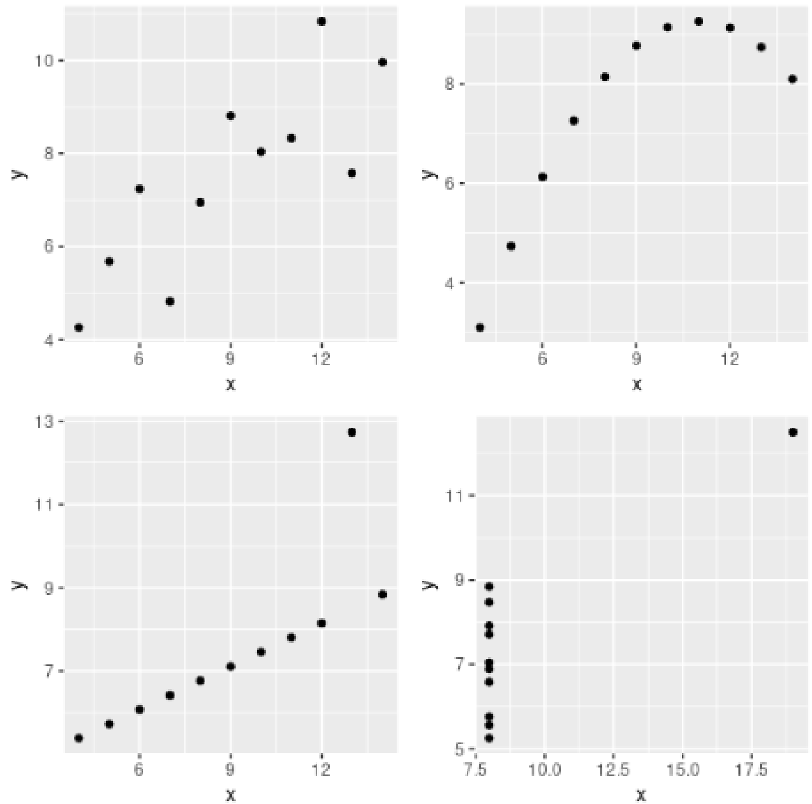


Figure 2.10: Each of these plots has a correlation coefficient of 0.82. Data from FJ Anscombe (1973) Graphs in statistical analysis. *American Statistician*, **27**, 17–21.

Kristianstad University

Example NBA players

	Name	Team	Number	Position	Age	Height	Weight	College	Salary
0	Avery Bradley	Boston Celtics	0.0	PG	25.0	6-2	180.0	Texas	7730337.0
1	Jae Crowder	Boston Celtics	99.0	SF	25.0	6-6	235.0	Marquette	6796117.0
2	John Holland	Boston Celtics	30.0	SG	27.0	6-5	205.0	Boston University	NaN
3	R.J. Hunter	Boston Celtics	28.0	SG	22.0	6-5	185.0	Georgia State	1148640.0
4	Jonas Jerebko	Boston Celtics	8.0	PF	29.0	6-10	231.0	NaN	5000000.0
5	Amir Johnson	Boston Celtics	90.0	PF	29.0	6-9	240.0	NaN	12000000.0
6	Jordan Mickey	Boston Celtics	55.0	PF	21.0	6-8	235.0	LSU	1170960.0
7	Kelly Olynyk	Boston Celtics	41.0	C	25.0	7-0	238.0	Gonzaga	2165160.0
8	Terry Rozier	Boston Celtics	12.0	PG	22.0	6-2	190.0	Louisville	1824360.0
9	Marcus Smart	Boston Celtics	36.0	PG	22.0	6-4	220.0	Oklahoma State	3431040.0

From: https://www.geeksforgeeks.org/python-pandas-dataframe-corr/

Kristianstad University

Correlation matrix

[5]:
```python
# To find the correlation among
# the columns using pearson method
df2 = df[["Number","Age","Weight","Salary"]]
df2.corr(method ='pearson')
```

[5]:

	Number	Age	Weight	Salary
Number	1.000000	0.028724	0.206921	-0.112386
Age	0.028724	1.000000	0.087183	0.213459
Weight	0.206921	0.087183	1.000000	0.138321
Salary	-0.112386	0.213459	0.138321	1.000000

Kristianstad University

Run the file – nba.csv

jupyter nba Last Checkpoint: för 3 timmar sedan (autosaved)

Logout

File Edit View Insert Cell Kernel Help

Trusted

Python 3 (ipykernel)

Run

Code

In [1]:
```
# importing pandas as pd
import pandas as pd
```

In [3]:
```
# Making data frame from the csv file
df = pd.read_csv("nba.csv")
```

In [5]:
```
 # Printing the first 10 rows of the data frame for visualization
df[:10]
```

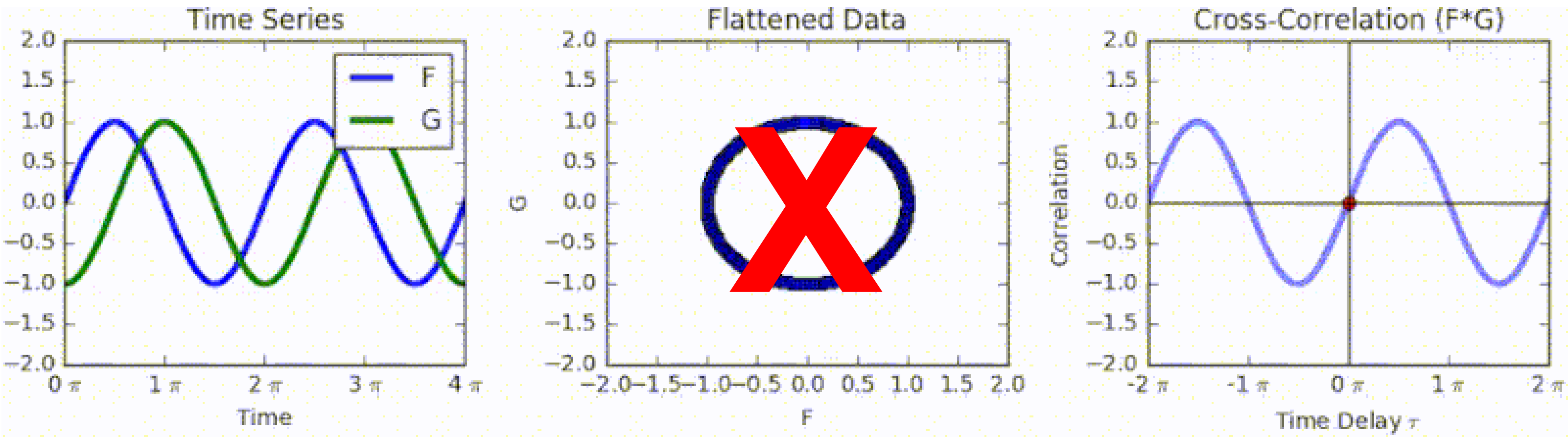
Out[5]:

	Name	Team	Number	Position	Age	Height	Weight	College	Salary
0	Avery Bradley	Boston Celtics	0.0	PG	25.0	6-2	180.0	Texas	7730337.0
1	Jae Crowder	Boston Celtics	99.0	SF	25.0	6-6	235.0	Marquette	6796117.0
2	John Holland	Boston Celtics	30.0	SG	27.0	6-5	205.0	Boston University	NaN
3	R.J. Hunter	Boston Celtics	28.0	SG	22.0	6-5	185.0	Georgia State	1148640.0
4	Jonas Jerebko	Boston Celtics	8.0	PF	29.0	6-10	231.0	NaN	5000000.0
5	Amir Johnson	Boston Celtics	90.0	PF	29.0	6-9	240.0	NaN	12000000.0
6	Jordan Mickey	Boston Celtics	55.0	PF	21.0	6-8	235.0	LSU	1170960.0
7	Kelly Olynyk	Boston Celtics	41.0	C	25.0	7-0	238.0	Gonzaga	2165160.0
8	Terry Rozier	Boston Celtics	12.0	PG	22.0	6-2	190.0	Louisville	1824360.0
9	Marcus Smart	Boston Celtics	36.0	PG	22.0	6-4	220.0	Oklahoma State	3431040.0

Autocorrelation

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

Auto-Correlation



Australian electricity demand

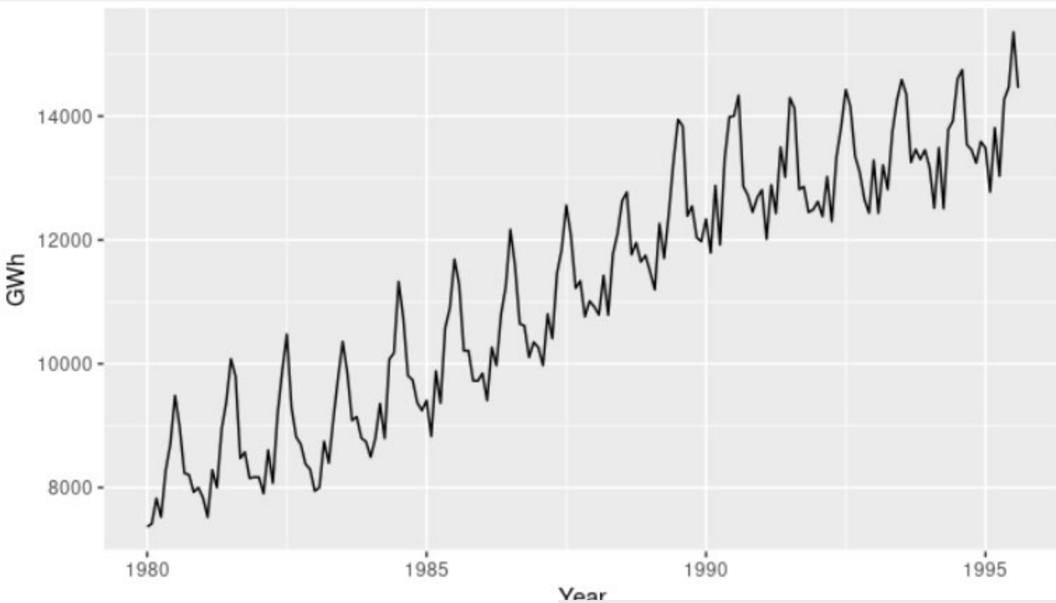


Figure 2.15: Monthly Australian

- Positive trend
- Seasonal (1 year)

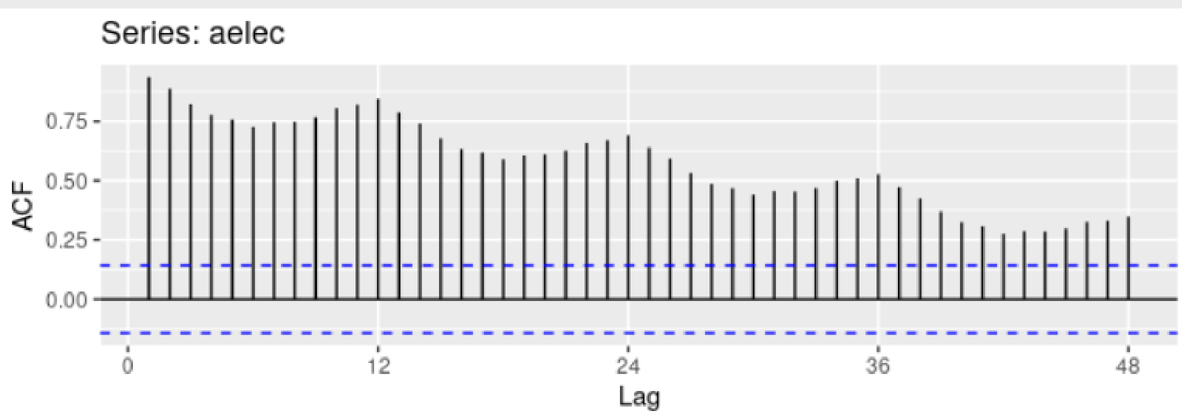


Figure 2.16: ACF of monthly Australian electricity demand.