# Week 1: Introduction

Visual Data Analytics

University of Sydney

# Outline

- Why visualisation?
- Good and bad visualisation.
- Building a narrative with visualisation.
- Tools for visualisation.

# Why Visualisation

# Visual Data Analytics (VDA)

- The entire process leading to data graphing.
- Encompasses the preparation of data for graphing and exploratory data analysis methods.
- Not simply about making ‘pretty pictures’ but about comprehending features of the data that are otherwise hidden by summary statistics.
- VDA is an invaluable business intelligence tool that uncovers hidden opportunities, and informs clear decision making

# How is VDA used?

- To report data using visual means rather than tables, enabling faster comparisons.
- For exploratory analysis to uncover new questions, discover previously unknown patterns, identify extreme behaviour and understand relationships between variables.
- As a diagnostic tool following statistical estimation.

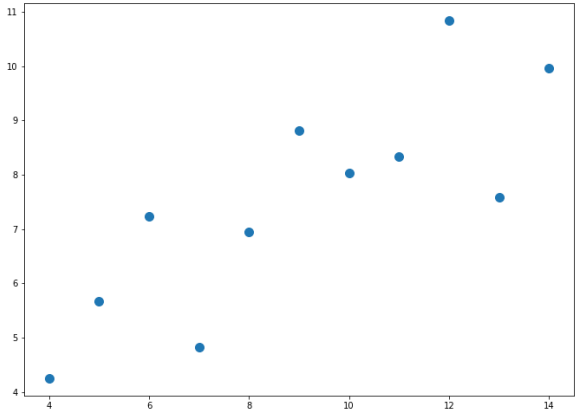
# Why do we visualise?

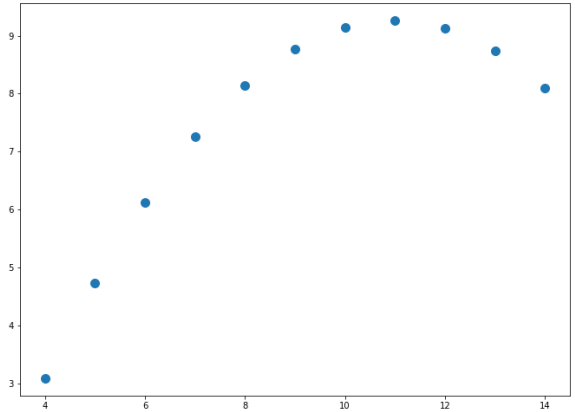
- More than 50% of the brain's neurons dedicated to vision.
- Nearly 10 million bits of information are processed per second through our eyes.
- Pre-attentive processing decodes information with high accuracy within 250 milliseconds (Healy and Enns, 2012).
- We have evolved to better decode information through visualisation.

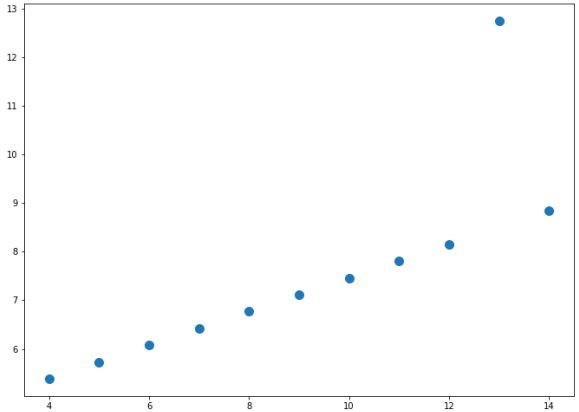
# Anscombe's Quartet

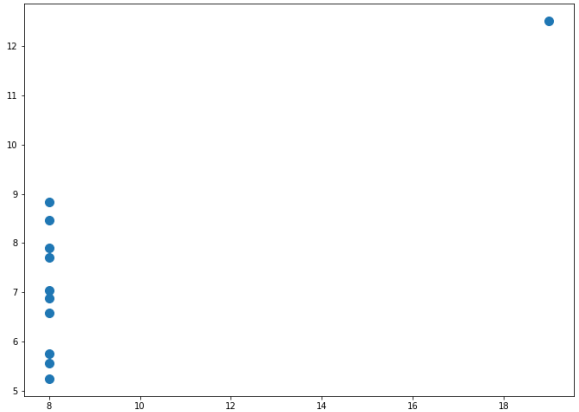
- Anscombe's quartet is a synthetic dataset of pairs of variables.
- For each data set, the means and variances and correlation between $x$ and $y$ are the same.
- However a simple scatterplot shows how different the datasets are:

# Anscombe's Quartet



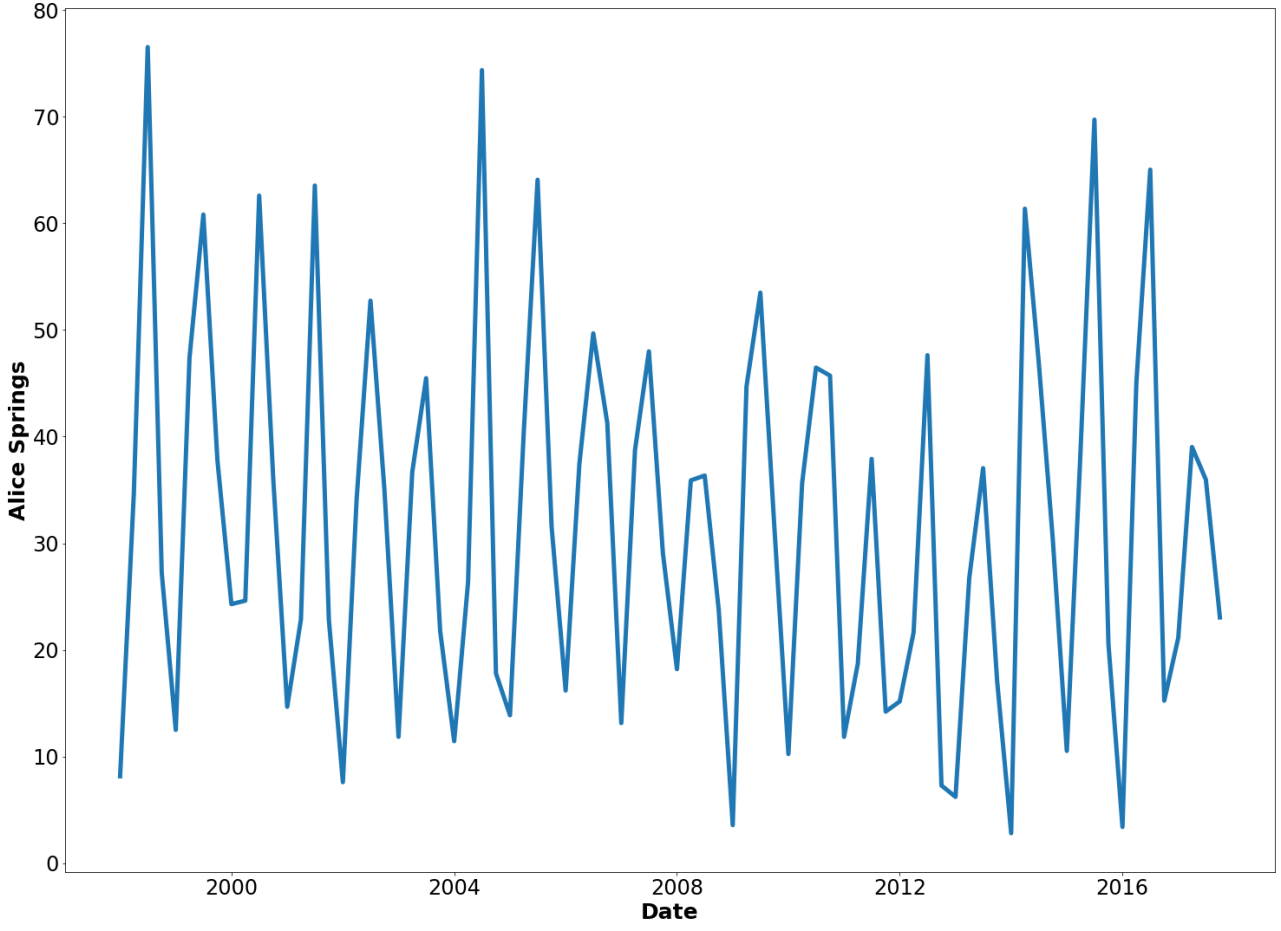
# Example

- Consider data on number of trips made for the purpose of holidays in two Australian regions:
  - Alice Springs in the Northern Territory,
  - The Wilderness West in Tasmania.
- We will look at these data in two different ways.
- May want to understand how demand evolves over the year to plan resourcing.

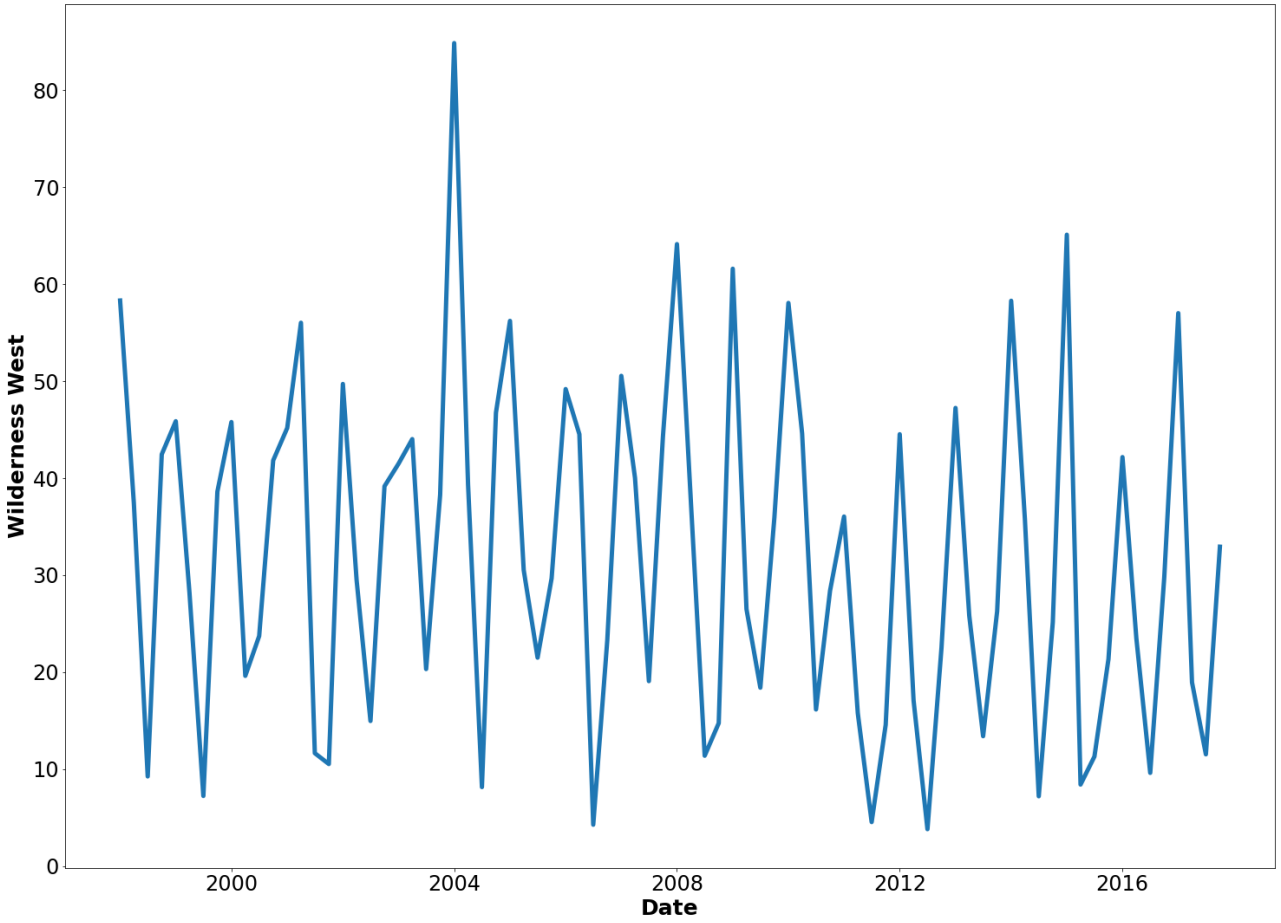
# As raw data

| Quarter | Alice Springs | Wilderness West |
|---|---|---|
| 1998-Q1 | 8.15 | 58.33 |
| 1998-Q2 | 34.66 | 37.46 |
| 1998-Q3 | 76.54 | 9.25 |
| 1998-Q4 | 27.22 | 42.46 |
| 1999-Q1 | 12.50 | 45.88 |
| 1999-Q2 | 47.38 | 28.17 |
| 1999-Q3 | 60.83 | 7.23 |
| 1999-Q4 | 37.81 | 38.62 |
| 2000-Q1 | 24.29 | 45.79 |
| 2000-Q2 | 24.62 | 19.62 |
| 2000-Q3 | 62.61 | 23.72 |
| 2000-Q4 | 36.12 | 41.84 |
| 2001-Q1 | 14.66 | 45.18 |
| 2001-Q2 | 22.87 | 56.04 |
| 2001-Q3 | 63.55 | 11.65 |

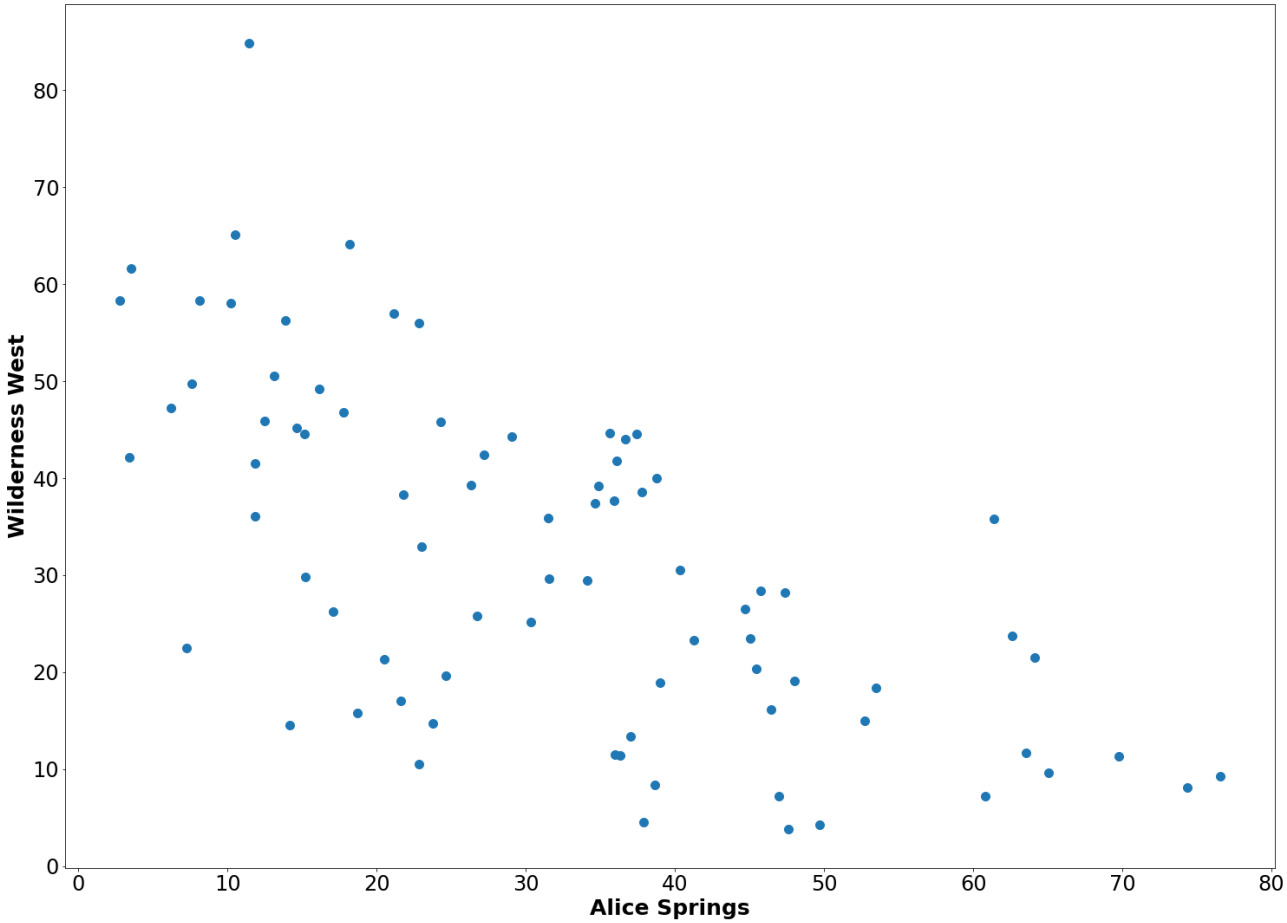
# As a plot (Alice Springs)



# As a plot (Wilderness West)



# As a plot (Both)



# Insights

- Both time series have a seasonal pattern.
  - There are some times of the years more popular for holidays.
- The seasonal patterns are different.
  - Alice Springs and Tasmanian Wilderness West have very different climate.
- It is easier to make these insights by visualisation rather than looking at raw numbers.

# The good, the bad and the ugly

# Tufte's Principles

- Principles of good practice in data visualisation are outlined in *The Visual Display of Quantitative Information* by Edward Tufte. These include:
  - Avoid distorting what the data have to say,
  - Present many numbers in a small space,
  - Make large data sets coherent,
  - Encourage the eye to compare different pieces of data.

# Iliinsky's four pillars

A visualisation is not just a picture

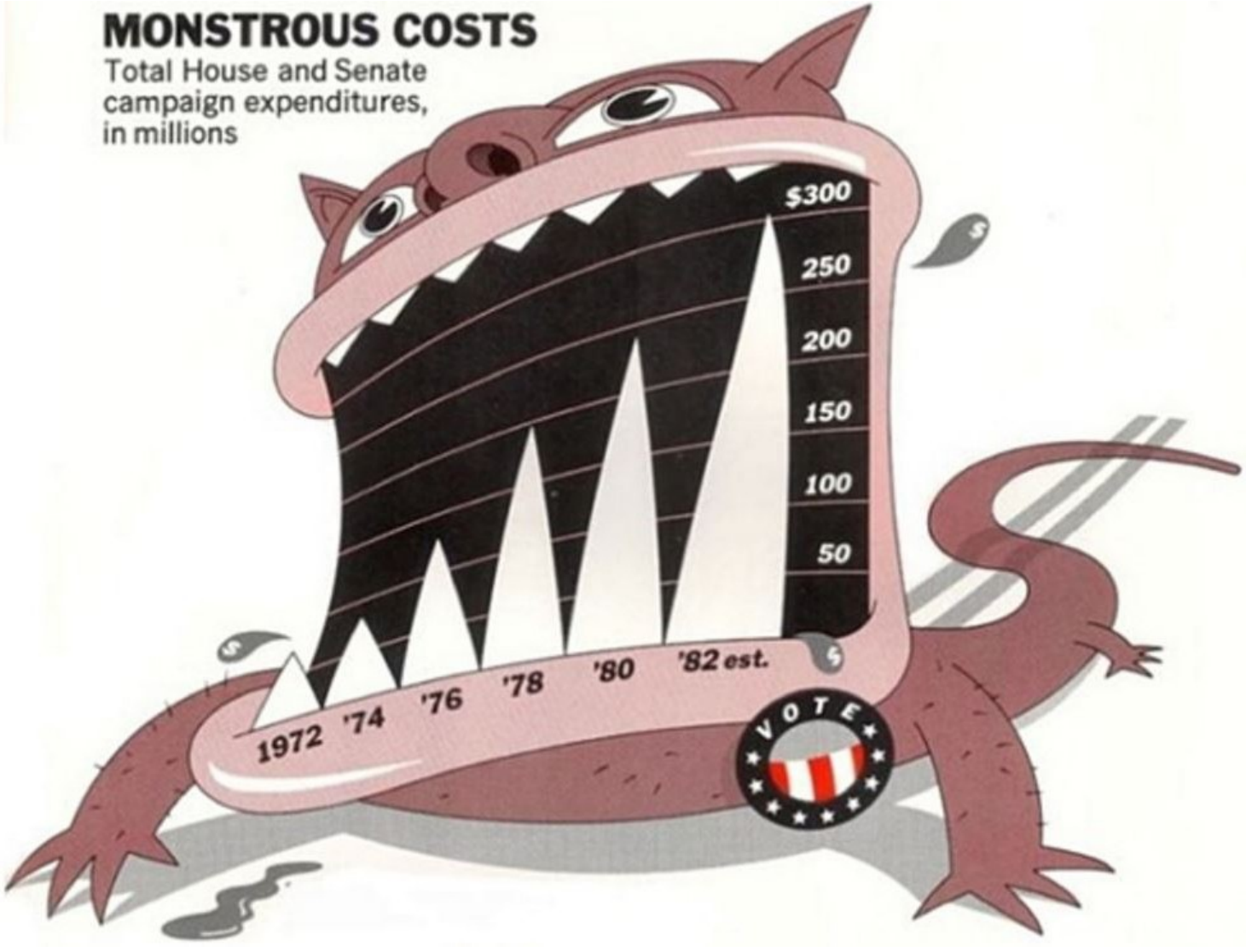
- Purpose (the why): have clear focus.
- Content (the what): contain correct and useful information.
- Structure (the how): what graph to choose.
- Formatting (everything else): bring focus.

For more see the video here

# Bad plots

- According to Healey's *Data Visualisation* plots may be bad due to:
  - Bad taste,
  - Bad perception,
  - Bad data.
- In the following examples, think about:
  - How the data is *encoded* into a visualisation, i.e. data $\rightarrow$ visuals,
  - How the data are *decoded* by the person interpreting the plot , i.e. visuals $\rightarrow$ insight.

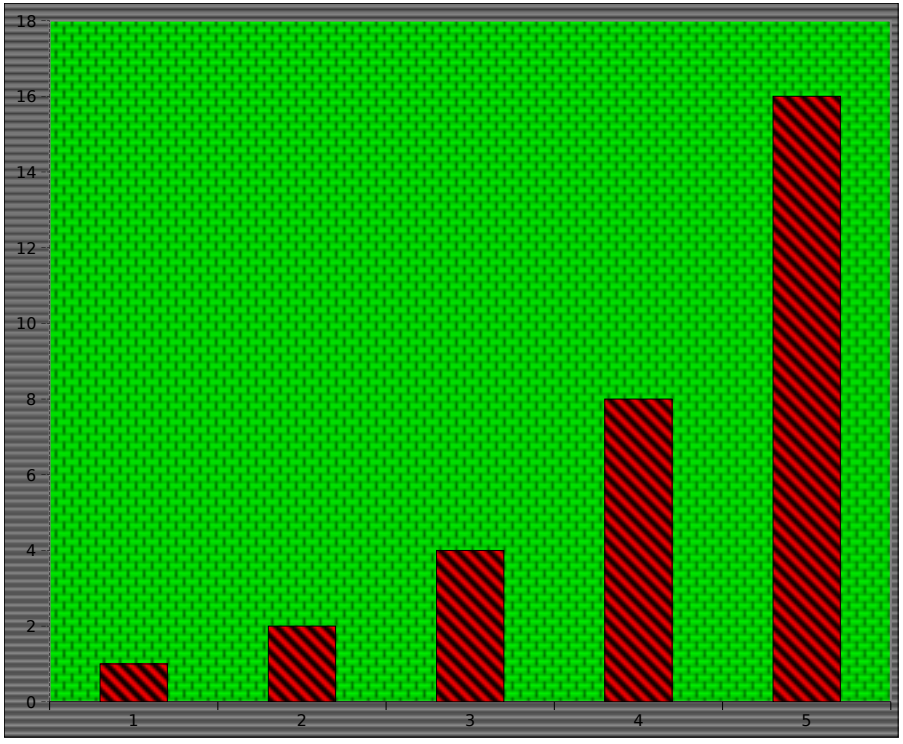
# Poor taste


MONSTROUS COSTS
Total House and Senate campaign expenditures, in millions
$300
250
200
150
100
50
1972
'74
'76
'78
'80
'82 est.
VOTE

# Chartjunk

- Chartjunk is the inclusion of elements that are not necessary to communicate the information.
- The inclusion of the following can be considered chartjunk:
  - Heavy gridlines,
  - Unnecessary text,
  - Pictures within the chart.
- These are not incorrect but can be misleading due to a lack of objectivity.
- Also examples of chartjunk that do not mislead.

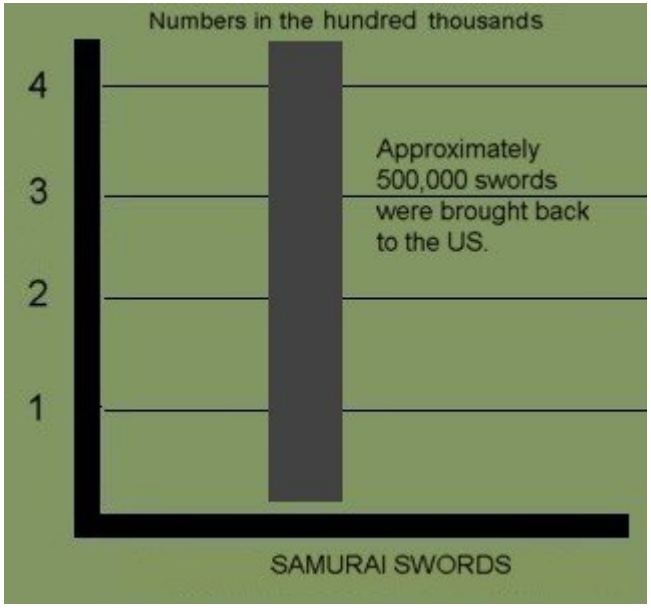
# Data-ink ratio



# Data-ink ratio

- One way to think about the design of a visualisation is using the *data-ink* ratio.
- The idea is to show the most data with the least amount of 'ink'.
- In the previous plot, the stripes on the bars, the color in the background do not convey any information about the data.
- This is an example of chatjunk that is not misleading.

# Data density



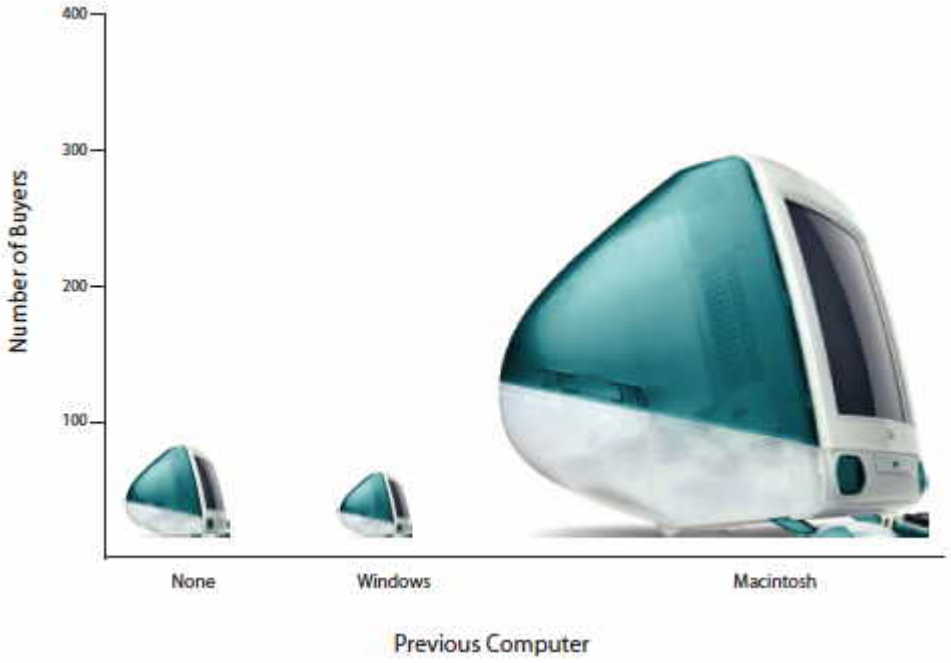
# Data density

- In the previous plot there is only one data point.
- Visualisation is not misleading.
- However is a visualisation necessary here?
- Visualisations that convey more data are said to have a high *data density*.
- In general, try to avoid low data density.

# Perceptually misleading

- Human perception is a broad field that takes in ideas from psychology and philosophy.
- For data visualisation we can perceive:
  - Length/Area/Volume,
  - Shape,
  - Position,
  - Color,
  - Angle.
- Now some examples where things go wrong.

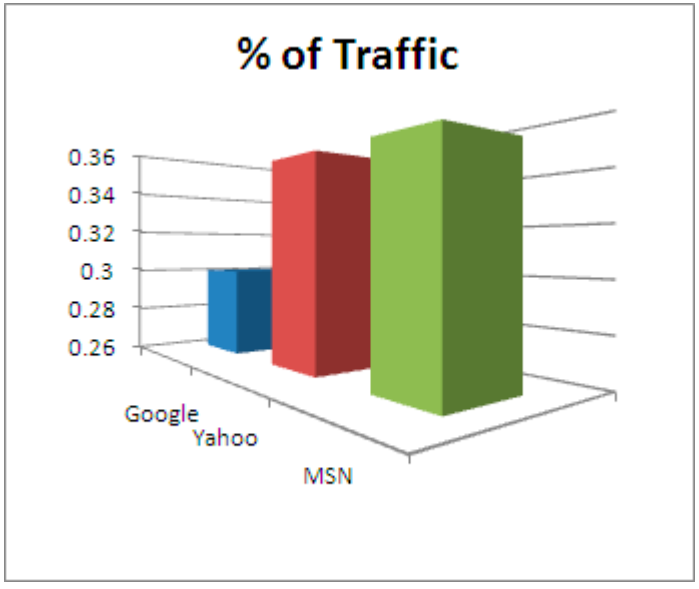
# Confusing length and area

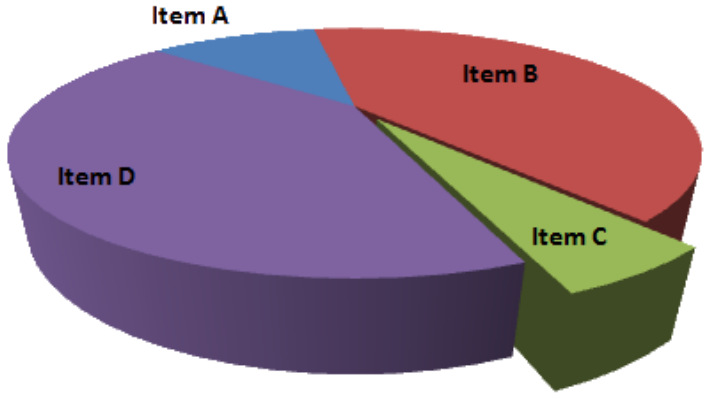


# Confusing Length and Area

- On the previous plot, the number of customers is mapped to length (height of computer)
- The area of the 2D pictures of computers scale up more than their heights.
- The picture leads us to imagine a 3D computer making this effect worse.
- The value for Mac is only about 3 to 4 times more than for None but we perceive the difference to be much more.

# Beware 3D

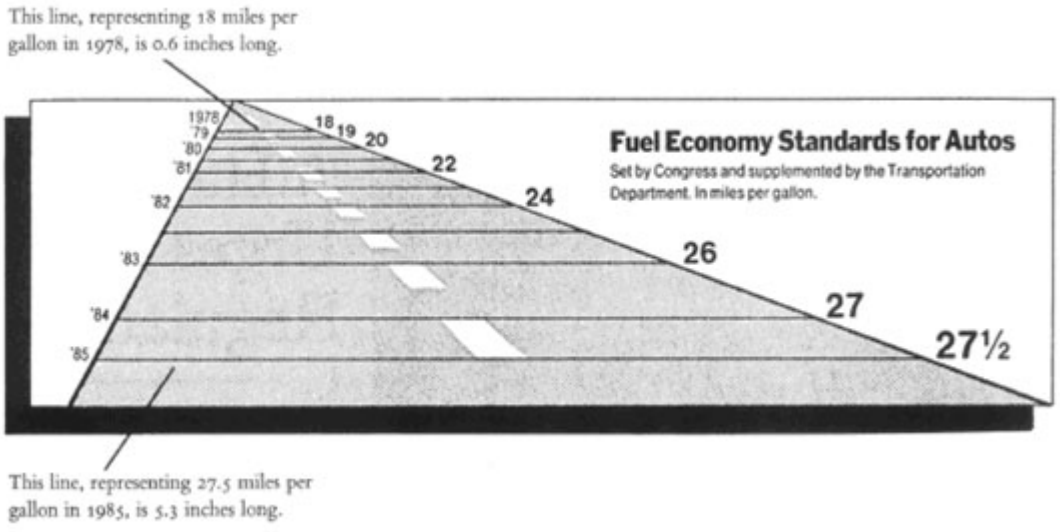

% of Traffic
0.36
0.34
0.32
0.3
0.28
0.26
Google
Yahoo
MSN


Item A
Item B
Item D
Item C

# Beware 3D

- Difficult to line up heights of bars with actual values
- Closer green bar (MSN) looks bigger.
  - Do not use 3 dimensions when 2 work well.
- On the pie chart the green segment looks bigger.
- In general angles are difficult to perceive
  - Experts in visualisation prefer not to use pie charts, but they are popular in practice.
- Argumentum ad populum / Three men make a tiger.

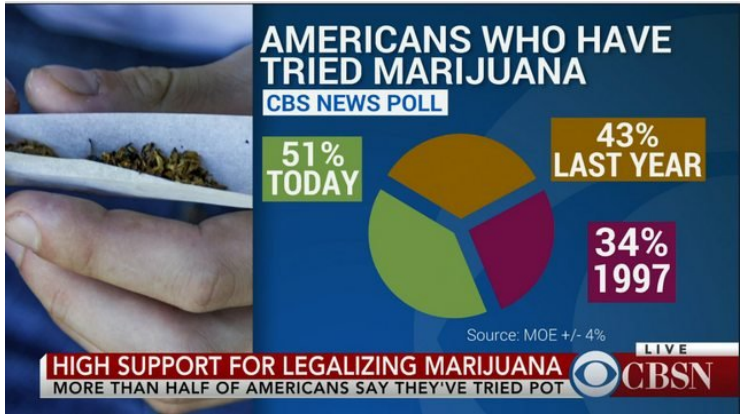
# Lie factor

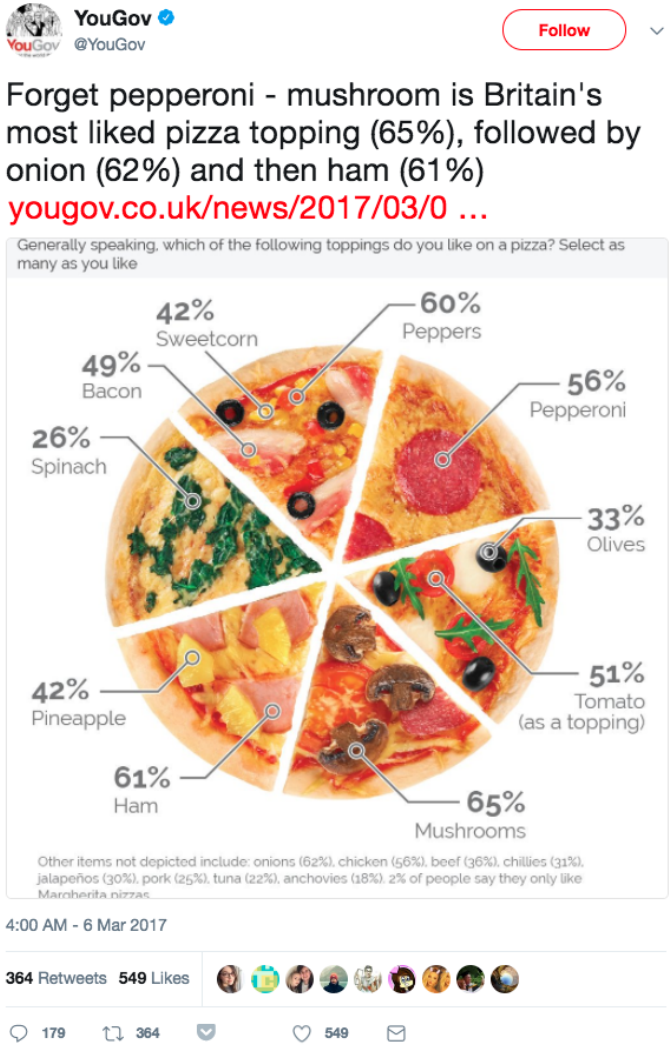

This line, representing 18 miles per gallon in 1978, is 0.6 inches long.
1978
'79
'80
'81
'82
'83
'84
'85
18
19
20
22
24
26
27
27½
Fuel Economy Standards for Autos
Set by Congress and supplemented by the Transportation Department. In miles per gallon.
This line, representing 27.5 miles per gallon in 1985, is 5.3 inches long.

# Lie factor

- The data indicates that mileage rose from 18 to 27.5 which is a 53% increase.
- The line on the graph increases from 0.6 inches to 5.3 inches which is a 783% increase!
- Tufte formalises this into a lie factor of 783/53≈14.
- It should be 1!
- Note that in contrast to previous examples where it was difficult to *decode* insights from the visualisation, here there is an error in how the data are *encoded* into a visualisation.

# Wrong data or plot

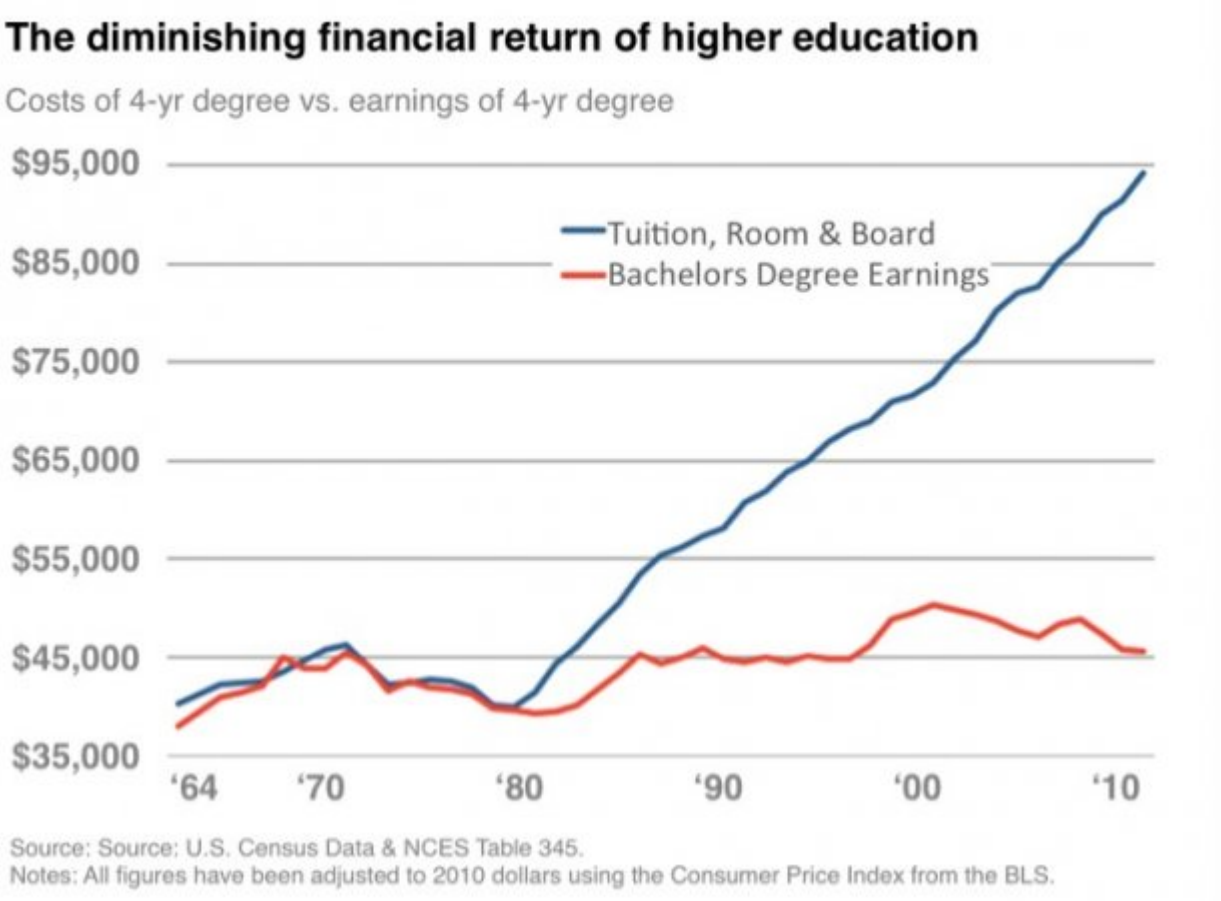


# Issues

- The percentages do not add up to 100.
- In the plot on the left, the data are a time series.
- Dates not given ('today' is 2016).
- In the plot on the right, respondents can like more than one topic
- In both cases the pie chart is a poor choice of visualisation for the data at hand.
- Also pineapple should never be on a pizza!

# Bad (but not wrong) data



# Problems

- There is nothing incorrect about this graph.
- However the message is misleading.
- The income is a yearly income while the cost of college is over four years (and only paid once).
- Also it does not show the income of people who are not college graduates.
- Think carefully about comparisons on a plot.
- Make sure conclusions align with what is in the plot.

# Storytelling with data

# Data storytelling

- Data storytelling is not about generating pretty charts and data presentations.
- It is about communicating insights that deliver real value.
- Good data stories have data, visualisations and narrative

# Guide to building narrative

Following discussion based on some Harvard Business Review articles

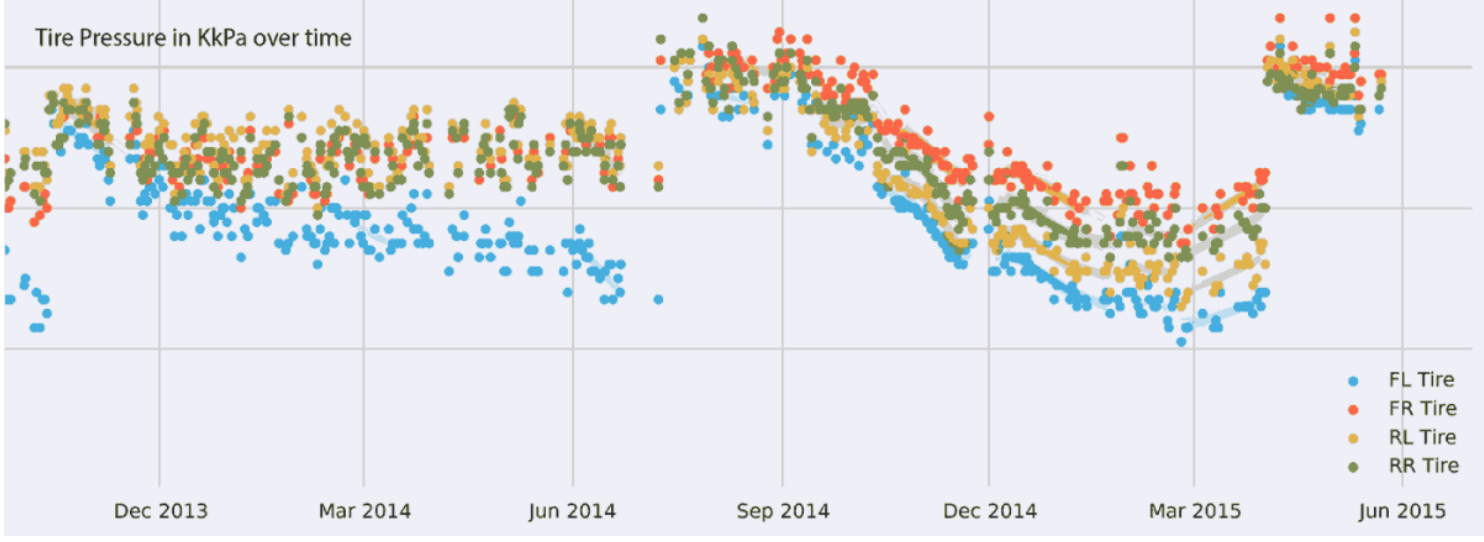
- How to make sure you're not using data just to justify decisions you've already made (link)
- Use data to answer you key business questions(link)
- Visualisations that really work (link)

# Key Business Questions

- What problem am I trying to solve?
  - Focus on something *actionable*!
- Immerse yourself in data
  - Including visualisation
- Generate KBQs
  - Make purpose specific
- Prioritise KBQs
  - Focus on easily activated, high impact KBQs
- Iterate!

# Example: Tesla

- Purpose: Improve customer satisfaction and operations with tyres?
- Visualisation:



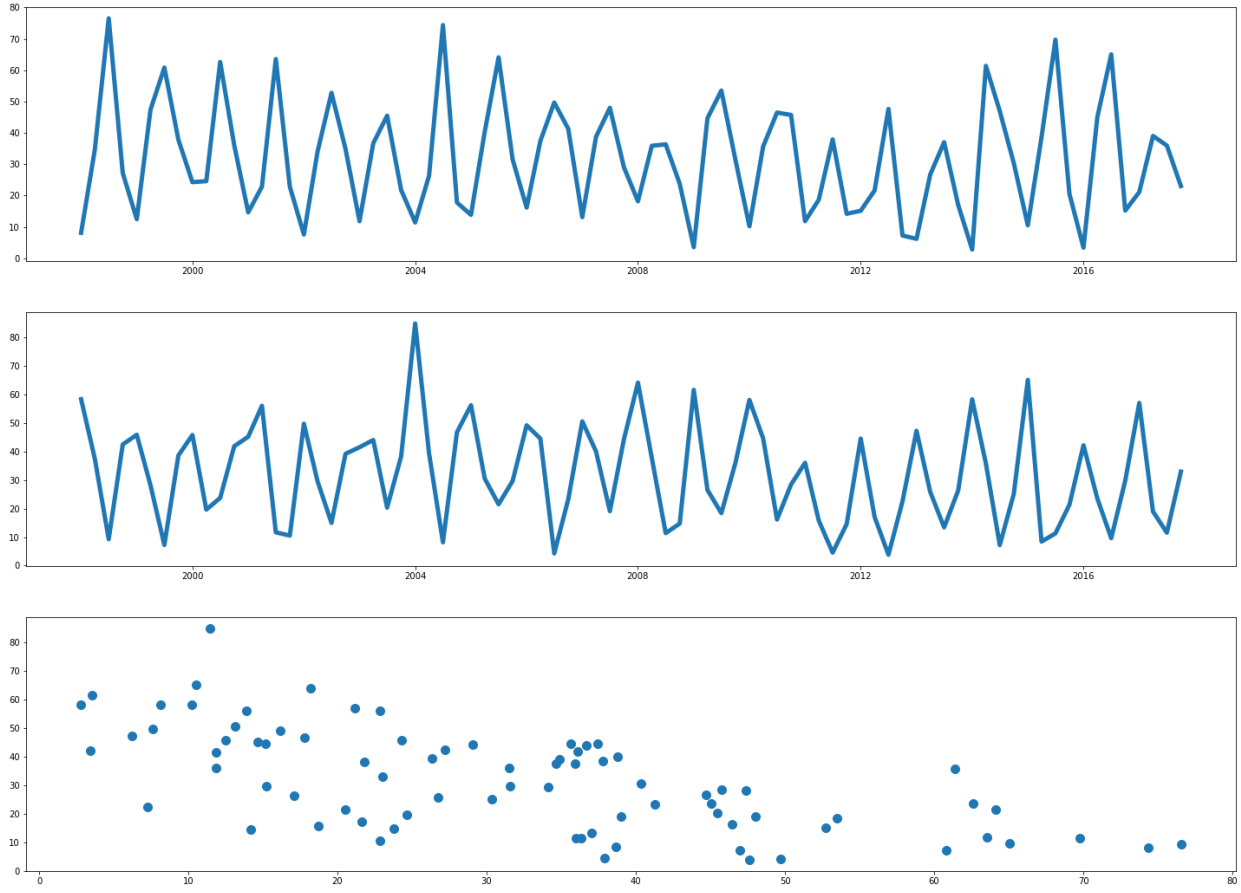
# Tesla: KBQs

- Identify KBQs
  - Is there sufficient quality control on tyres leaving factory?
  - How long do customers take to respond to a low pressure alert?
  - Can we predict when when tyres go flat?
- Prioritise KBQs
  - Will depend on context
  - Predictive model may not be easily activated.

# Narrative

- All stories consist of
  - Setup (current reality)
  - Conflict (change)
  - Resolution (new reality)
- For an example (with house prices) see here.
- We will work through an example now due with the Tourism data

# Tourism narrative



# The tools

# Python

- General purpose programming language.
- Language of choice for data science.
- Open source libraries for visualisation.

# Matplotlib

- Most popular Python package for visualisation.
- Highly customisable.
- Works with with other packages.

# Seaborn

- Builds on top of Matplotlib.
- Easier integration with Pandas dataframes
- Nice themes

# Plotly

- Good for interactive plots.
- Suited to web-based interface.
- Also implemented in other languages.

# Bokeh

- Alternative for interactive plots.
- Good for interactive dashboards.

# Why not Tableau?

- Tableau is a popular commercial tool for visualisation.
- It arguably has an easier interface (no coding).
- Python is more customisable.
- Python can be used everywhere and anywhere.
- If you know Python, easier to learn Tableau (compared to the other way around).
- You will need to learn coding, but this is not a coding course.

Questions?