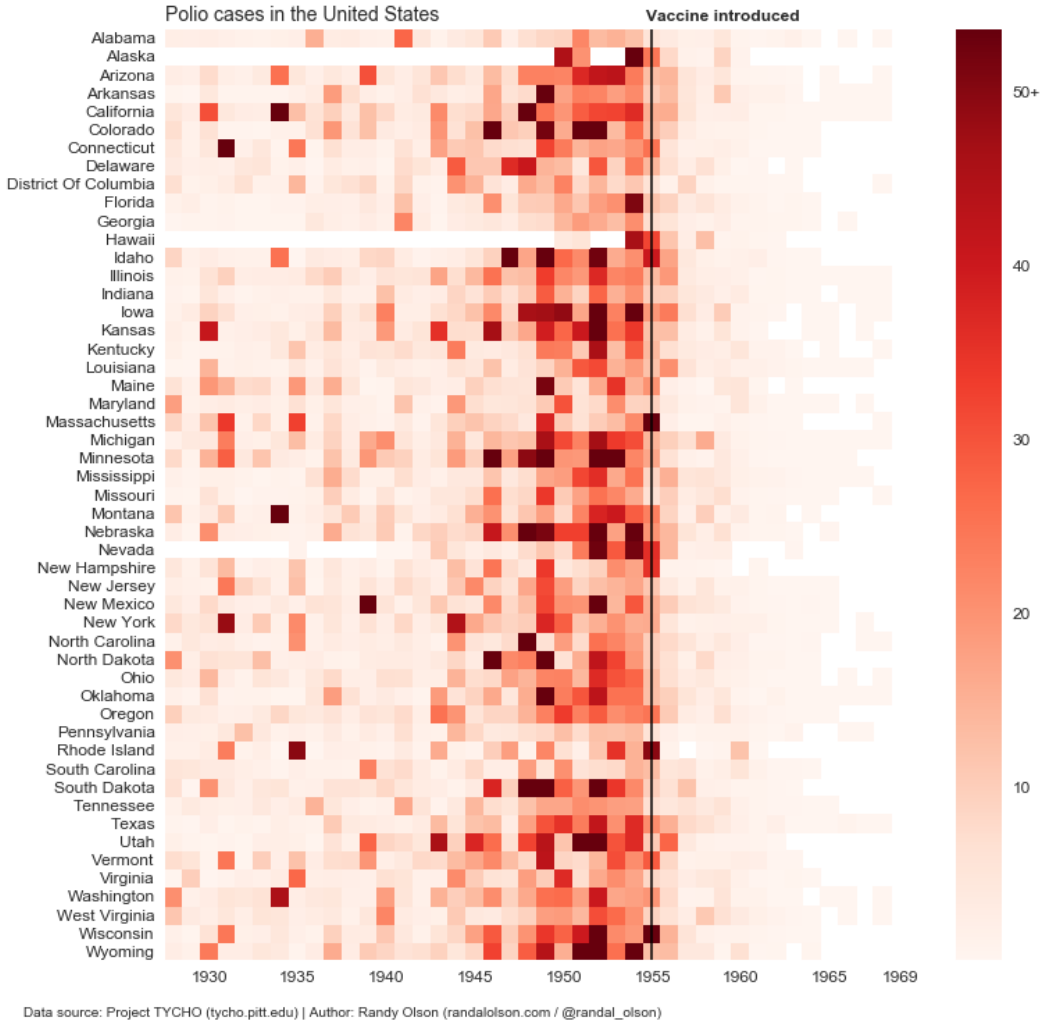
# Data Visualisation
# Lecture 4 – Visualising Comparisons

Dr. Cathy Ennis

# Learning Outcomes Week 4

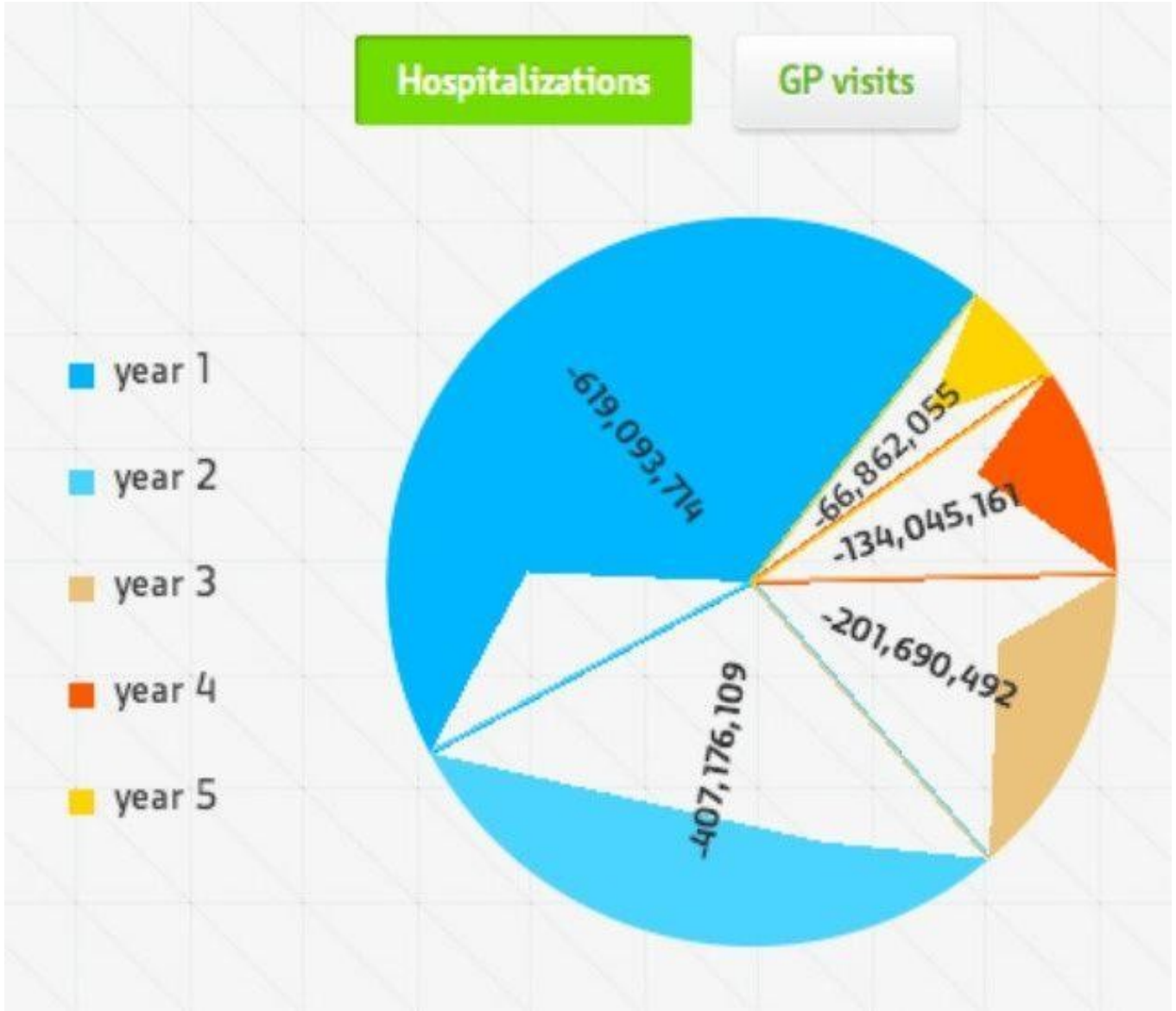
- Design effective visualizations based on principles from perceptual psychology, cognitive science, graphic design and visual art
- Create and deploy successful data visualisations using leading software tools
- Demonstrate an understanding how visualisation is used in date journalism to communicate complex ideas and stories
- Demonstrate understanding how visualisation is used in story telling

# Visualisation of the Week

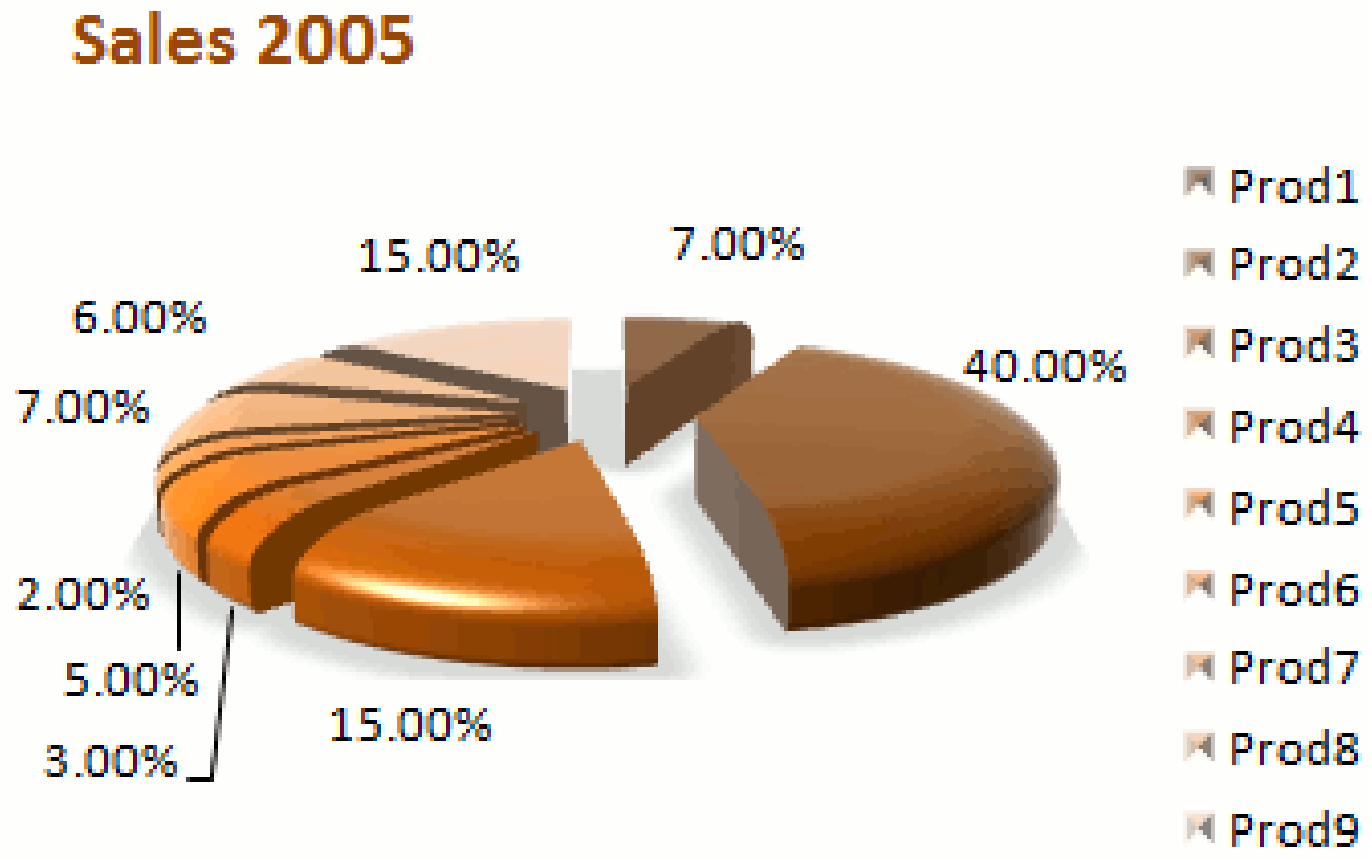



http://www.randalolson.com/2016/03/04/revisiting-the-vaccine-visualizations/

# (Un)Visualization of the Week



# (Un)Visualization of the Week



# Overview

The concept of storytelling and how can visualise comparisons between variable values

- Single variable exploration
- Simple comparisons
- Multivariate distributions

# STORY TELLING

# Storytelling

- Data visualisations are not really about the data, they are about the meaning of the data

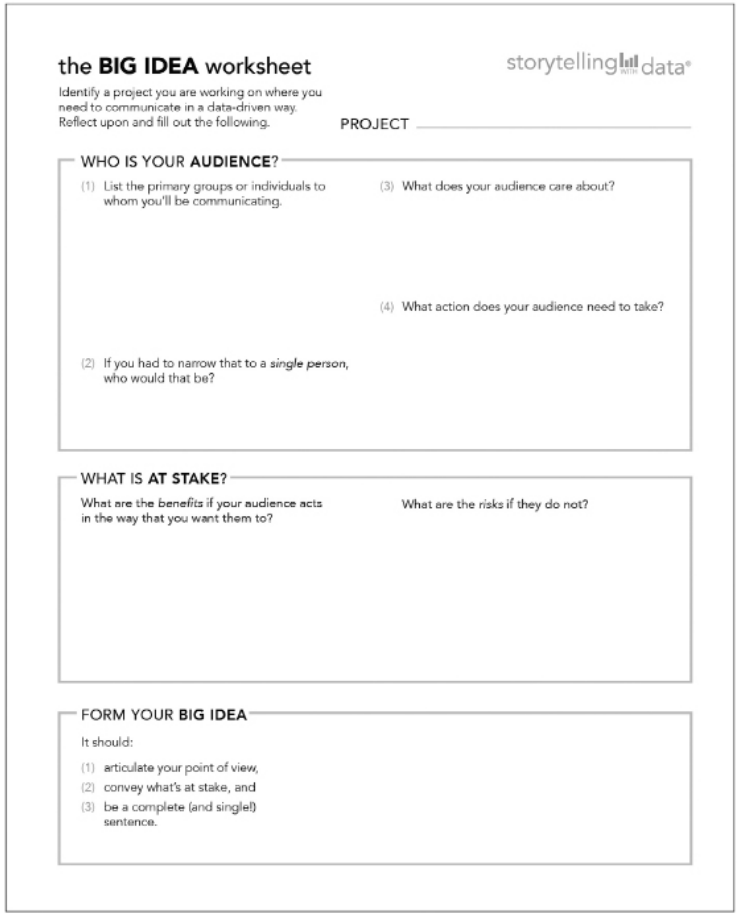
- Storytelling guides your audience from one point or argument to the next. For example:
  - Chronologically
  - When it involves a combination of a broad overview and granular detail
  - When the message has multiple separate components

# Storytelling - Understand the Context

- How to capture audience's attention without losing the most important parts of the data?
- How to turn data into information that can be consumed by an audience?

- **Who** am I communicating to?
- **What** do I want my audience to know or do?
- **How** can I use data to help make my point?

# Storytelling - Understand the Context



## the **BIG IDEA** worksheet

storytelling with data

Identify a project you are working on where you need to communicate in a data-driven way. Reflect upon and fill out the following.

PROJECT _____

### WHO IS YOUR **AUDIENCE**?

(1) List the primary groups or individuals to whom you'll be communicating.

(2) If you had to narrow that to a *single person*, who would that be?

(3) What does your audience care about?

(4) What action does your audience need to take?

### WHAT IS **AT STAKE**?

What are the *benefits* if your audience acts in the way that you want them to?

What are the *risks* if they do not?

### FORM YOUR **BIG IDEA**

It should:

(1) articulate your point of view,
(2) convey what's at stake, and
(3) be a complete (and single!) sentence.

Storytelling with Data, let's practice. Cole Nussbaumer Knaflic

# Storytelling - Explanatory Analysis

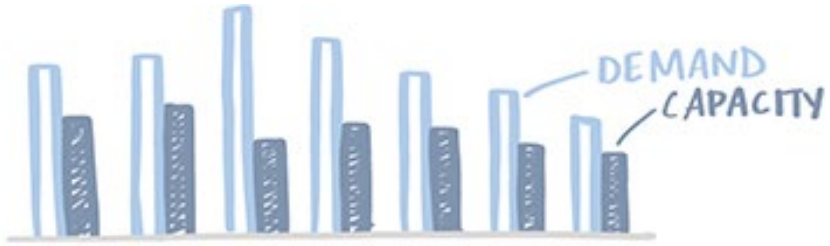
- Focus on explanatory analysis (**NOT** exploratory analysis) and communication

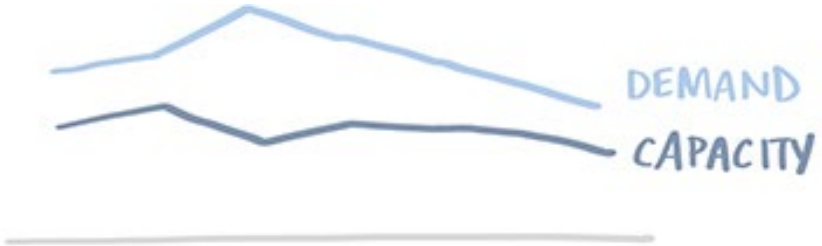
- **Choose an Effective Visual**

- **Eliminate Clutter**

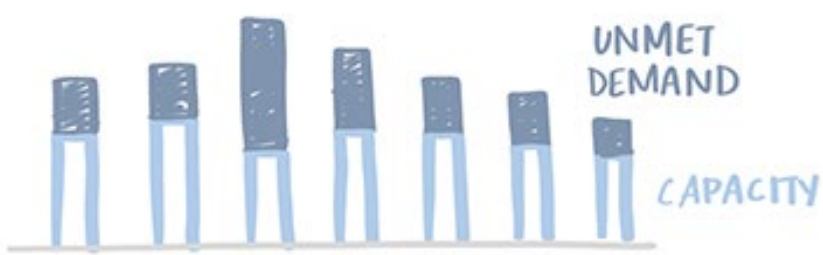
- **Draw Attention Where You Want It**

# Choosing the Appropriate Visual



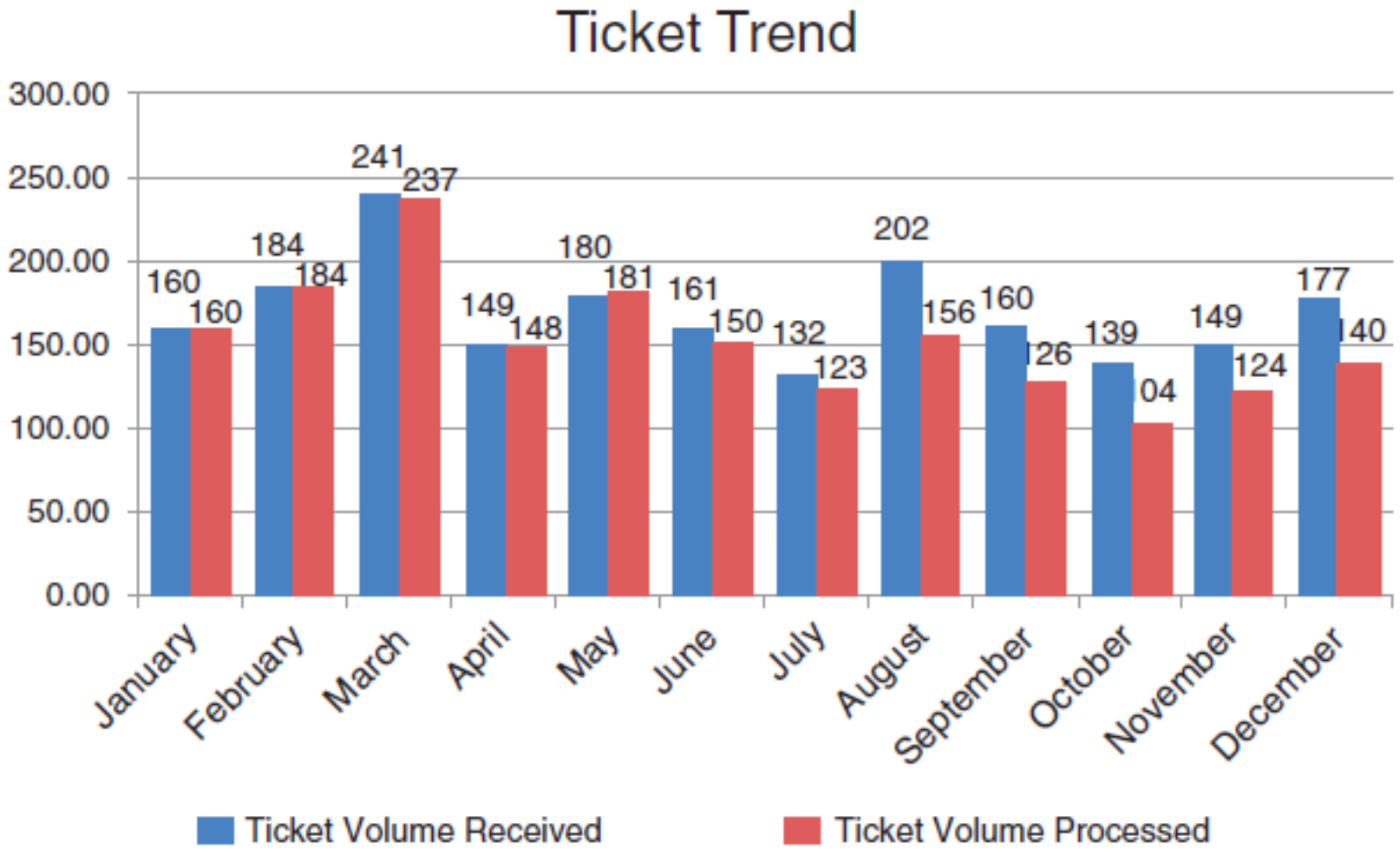







Storytelling with Data, let's practice. Cole Nussbaumer Knaflic

# Example


Ticket Trend
300.00
250.00
200.00
150.00
100.00
50.00
0.00
160
160
184
184
241
237
149
148
180
181
161
150
132
123
202
156
160
126
139
04
149
124
177
140
January
February
March
April
May
June
July
August
September
October
November
December
Ticket Volume Received
Ticket Volume Processed

# Example

## Please approve the hire of 2 FTEs

to backfill those who quit in the past year

Ticket volume over time

# Craft a Narrative

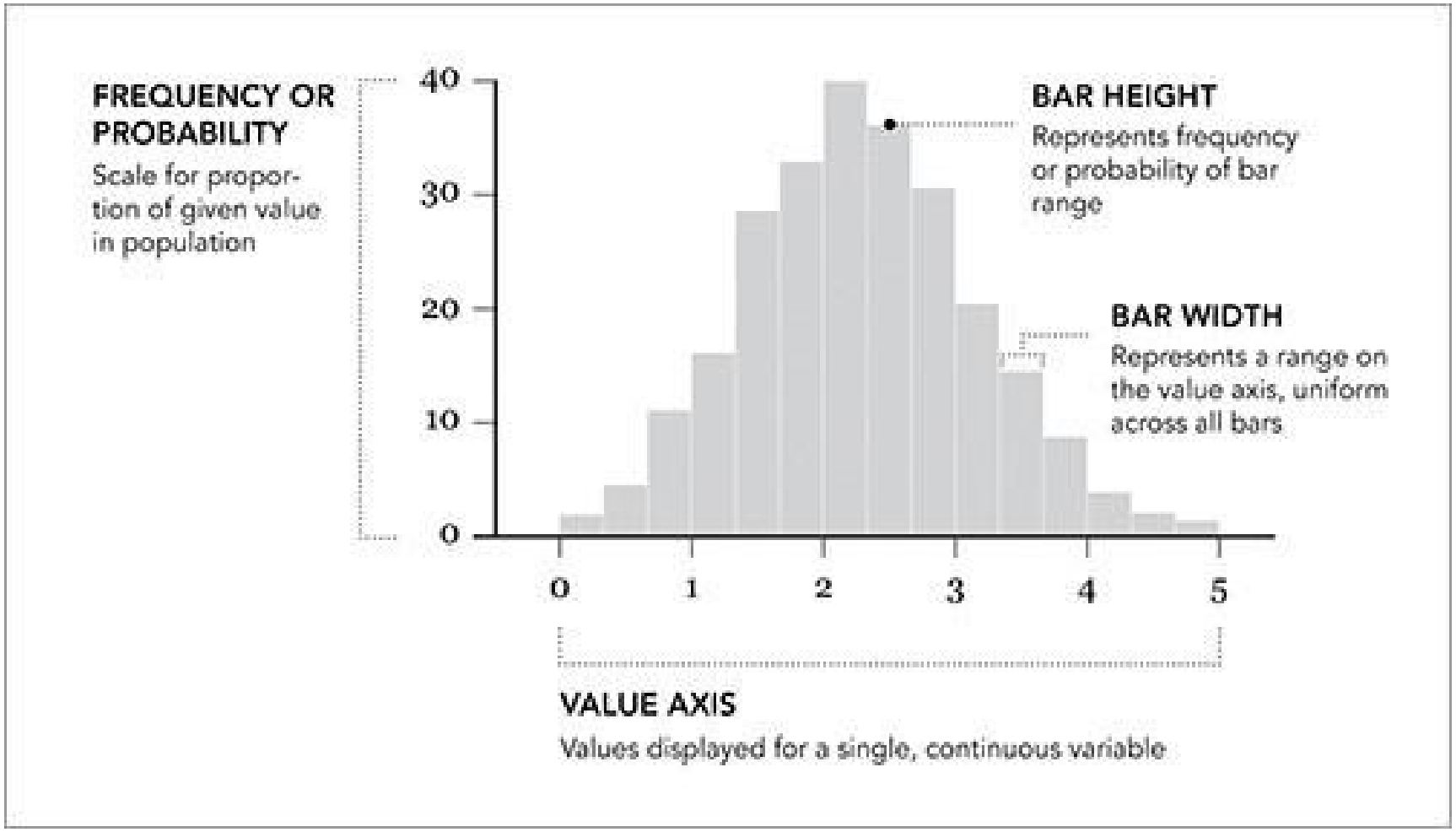
- Telling a story that the reader will remember, and possibly re-tell later
- Plot – what is the essential context?
- Twist – what are the most interesting findings in the data?
- Ending – what do you want your audience to do? What is the take home message?

# Storytelling References

- https://www.storytellingwithdata.com/books
- https://youtu.be/8EMW7io4rSI
- https://www.forbes.com/sites/evamurray/2019/02/06/how-do-you-tell-a-story-with-data-visualization
- https://towardsdatascience.com/storytelling-with-data-a-data-visualization-guide-for-business-professionals-97d50512b407

# SINGLE VARIABLE EXPLORATION

# Histogram

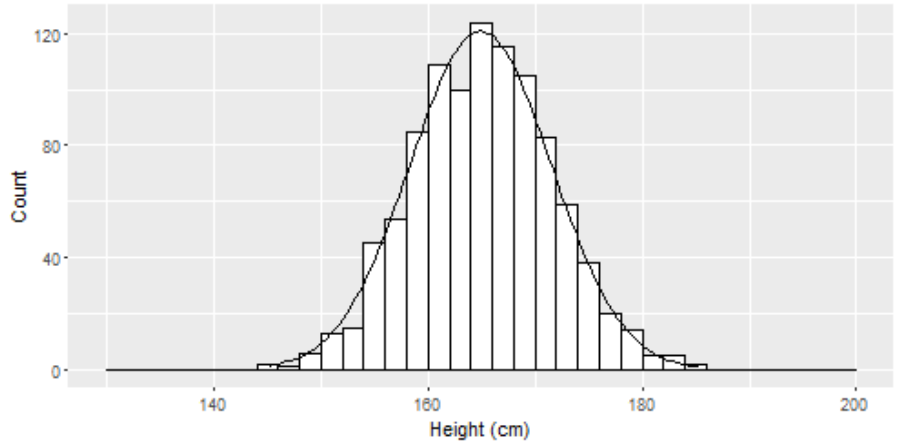



Visualize This: The Flowing Data Guide to Design, Visualization, and Statistics. Nathan Yau. 2011

# Histogram

- A histogram is an estimate of the probability distribution of a continuous variable (quantitative variable)

- It differs from a bar graph. A bar graph relates two variables (categorical and temporal data), but a histogram relates only one




Visualize This: The Flowing Data Guide to Design, Visualization, and Statistics. Nathan Yau. 2011

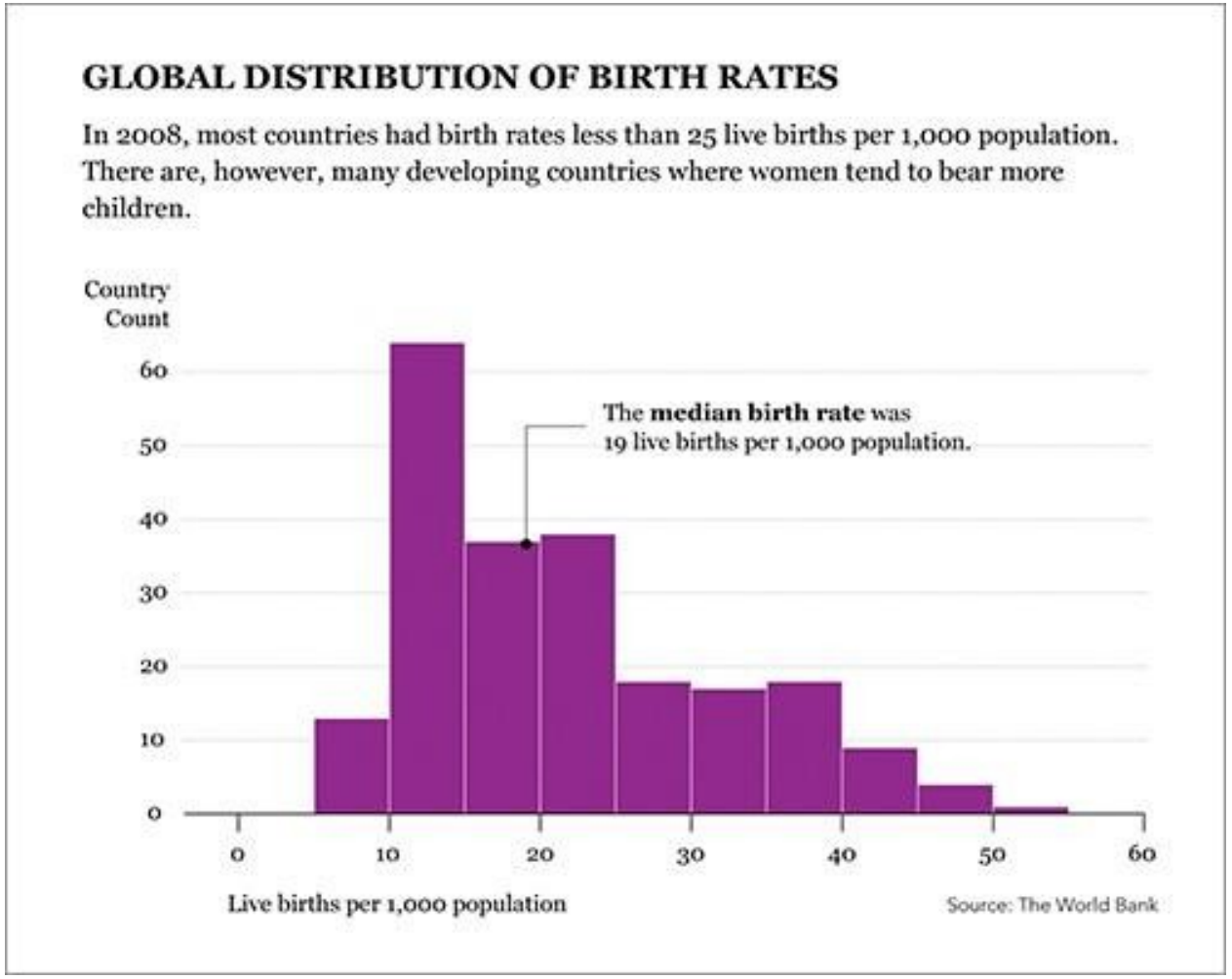
# Histogram

- A histogram gives us an in-depth view of a single numeric variable. To construct a histogram:

1. Divide the data range into bins
2. Count the occurrence frequency of each bin within the data
3. Normalize the frequency counts
   - Display "relative" frequencies. It shows the proportion of cases that fall into each of several categories, with the sum of the heights being equal to 1
4. Plot a bar graph to show the normalised count for each bin

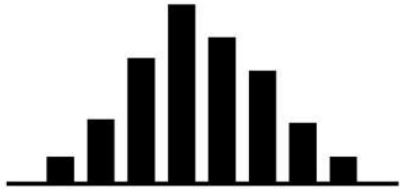
Visualize This: The Flowing Data Guide to Design, Visualization, and Statistics. Nathan Yau. 2011

# Histogram

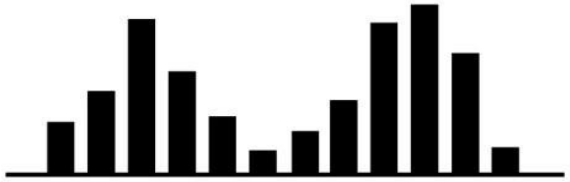



Visualize This: The Flowing Data Guide to Design, Visualization, and Statistics. Nathan Yau. 2011

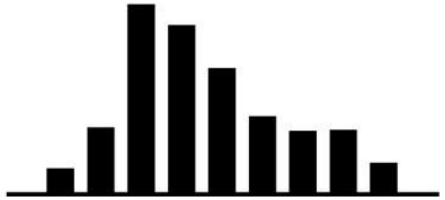
# Histogram Shapes



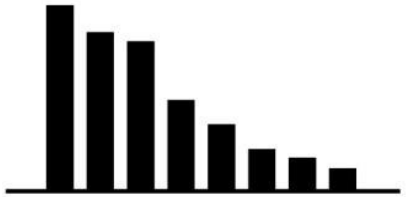
*Bell Shaped:*
The normal pattern



*Double Peaked:* Suggests two distributions



*Skewed:* Look for other processes in the tail



*Truncated:* Look for reasons for sharp end of distribution or pattern

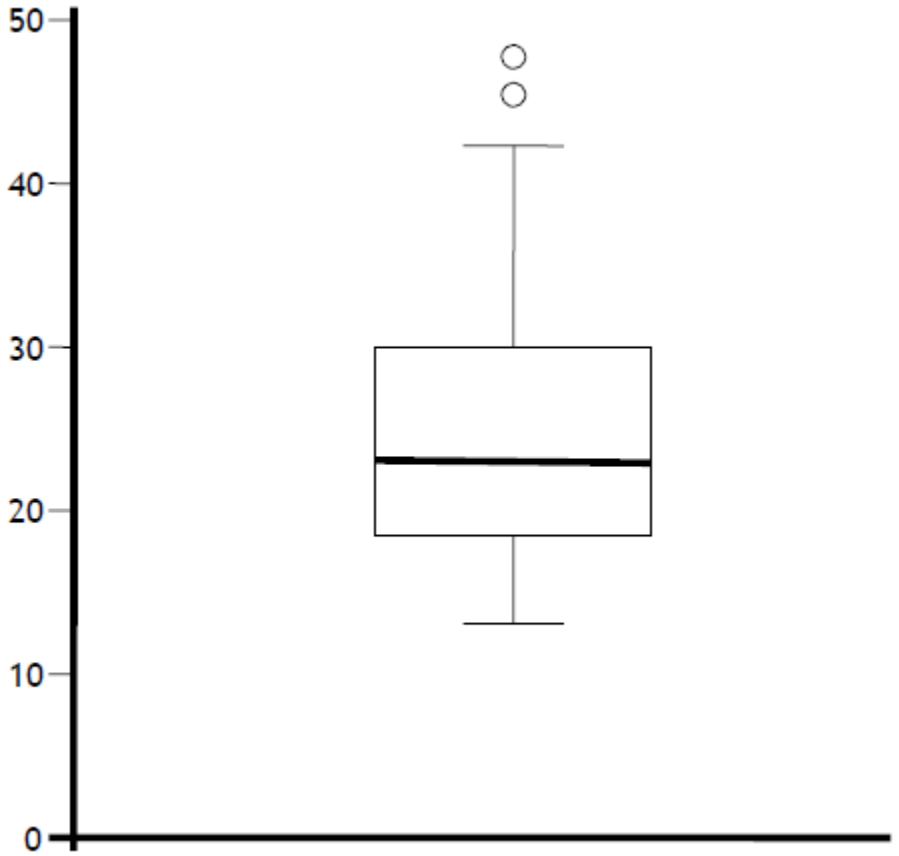
*Ragged Plateau:* No single clear process or pattern

# Histogram

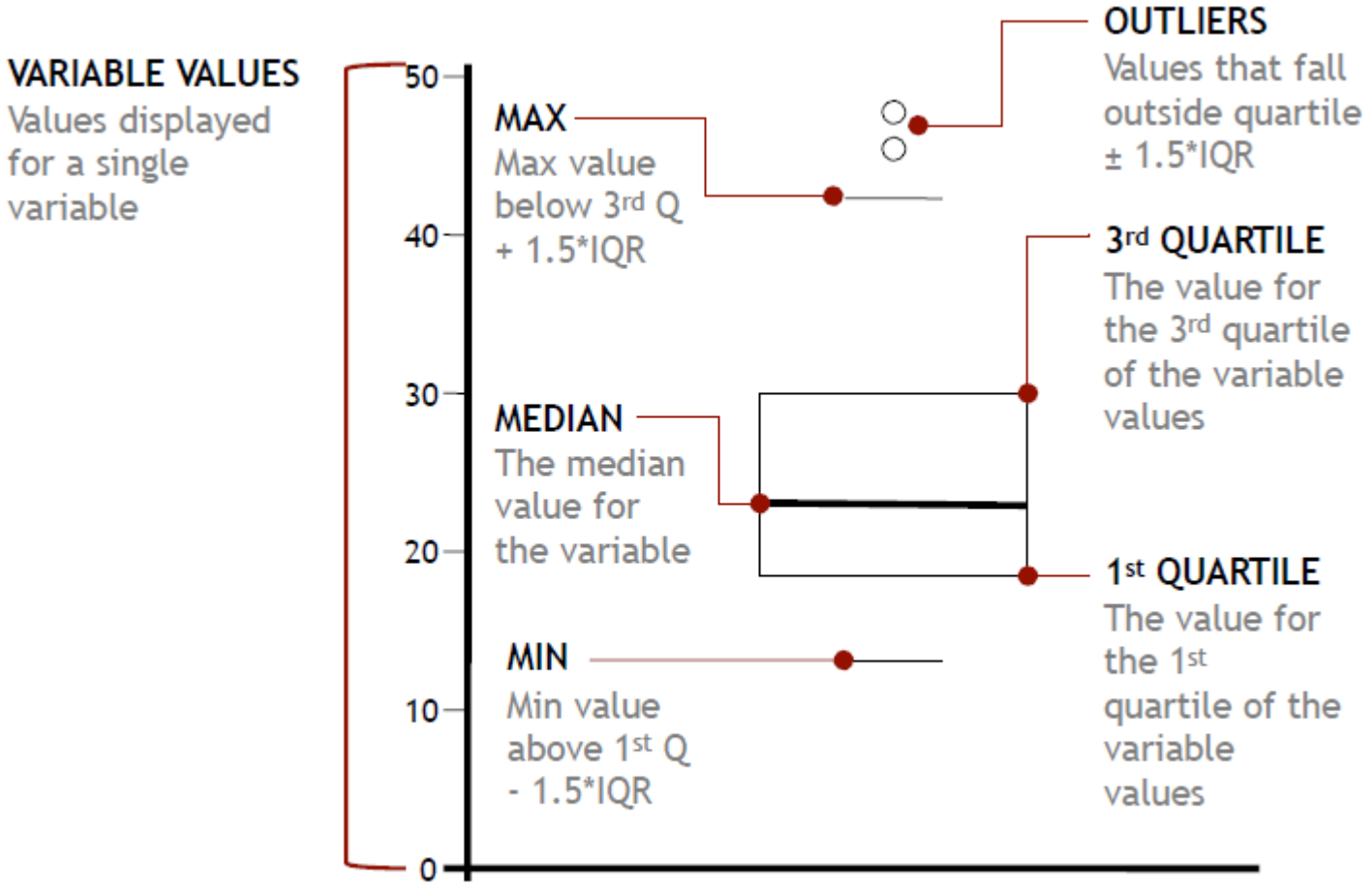
- It is a good idea to plot the data using several different bin widths to learn more about it

- The histogram is quite possibly your most important visual data **exploration** tool!!!

# Box Plot



# Box Plot


VARIABLE VALUES
Values displayed for a single variable
50
40
30
20
10
0
MAX
Max value below 3rd Q + 1.5*IQR
MEDIAN
The median value for the variable
MIN
Min value above 1st Q - 1.5*IQR
OUTLIERS
Values that fall outside quartile ± 1.5*IQR
3rd QUARTILE
The value for the 3rd quartile of the variable values
1st QUARTILE
The value for the 1st quartile of the variable values

# Box Plot

The components of a box plot are:

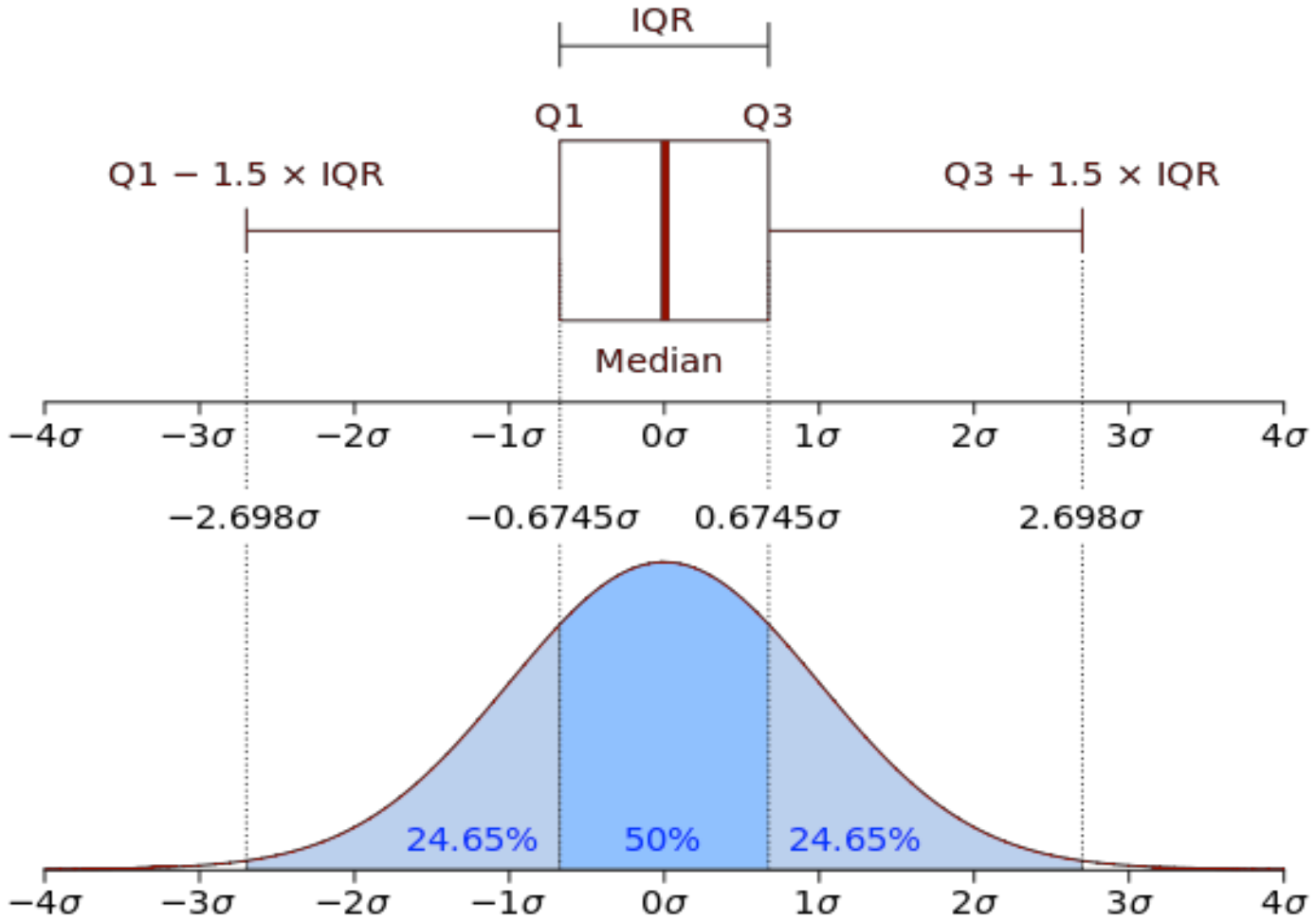
- A thick dark line at the median
- A horizontal lines at the 1st quartiles
- A horizontal lines at the 3rd quartiles
- A whisker down to the *low* value
  - Multiply the IQR by 1.5 to calculate the *step*
  - The low value is the lowest value above the 1st quartile minus the step
- A whisker up to the *high* value
  - The high value is the highest value above the 3rd quartile plus the step
- Any values outside low and high are marked as **outliers**

# Box Plot

Some important points about a box plot:

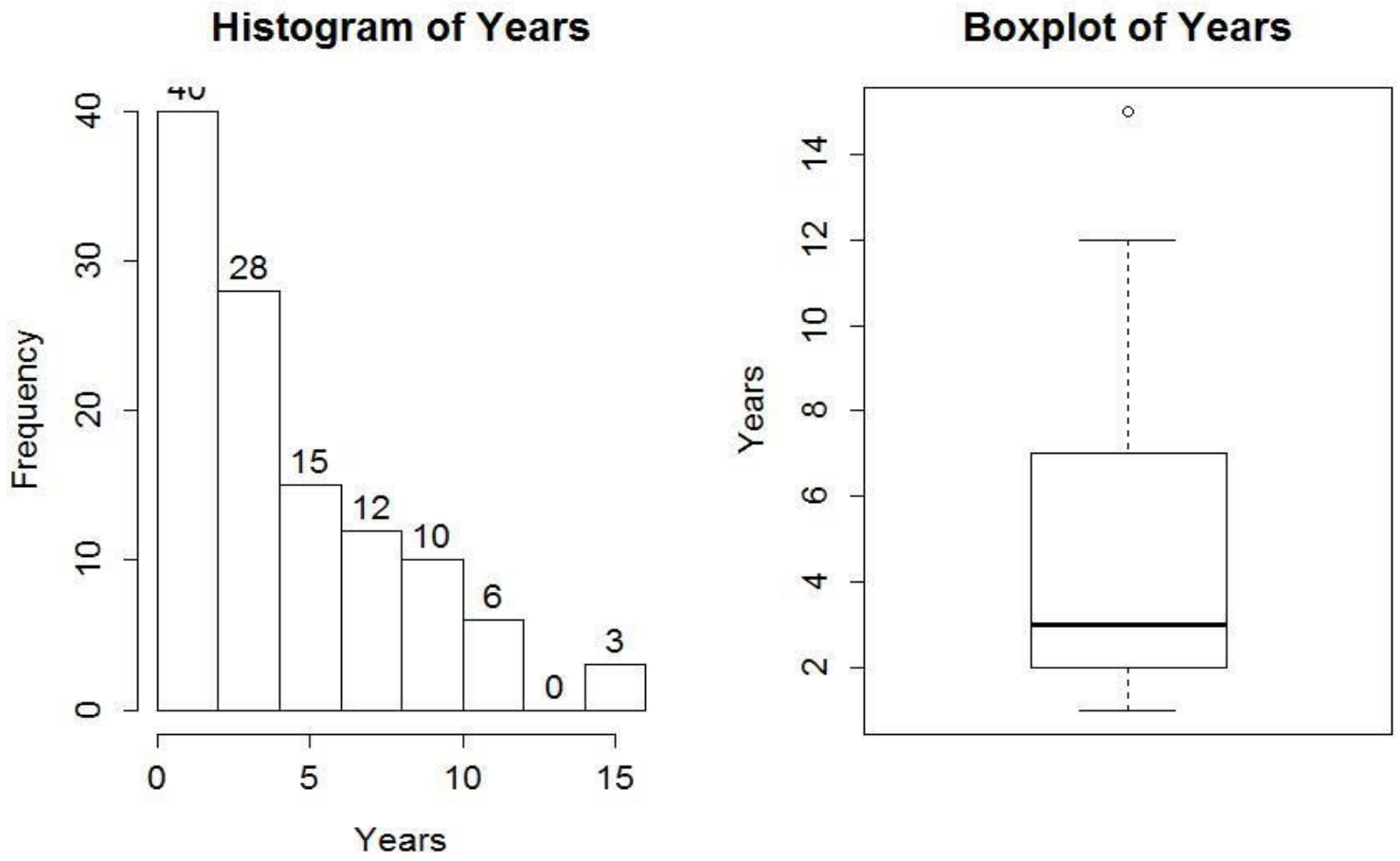
- 50% of the data occurs between the lower and upper edges of the box
- The lower 50% of the data occurs below the median
- The upper 50% of the data occurs above the median line in the box.
- The lower 25% of the data occurs between the bottom edge of the box and the bottom edge of the lower whisker
- The upper 25% of the data occurs above the top edge of the box and the top edge of the upper whisker

# Box Plots & Density Functions



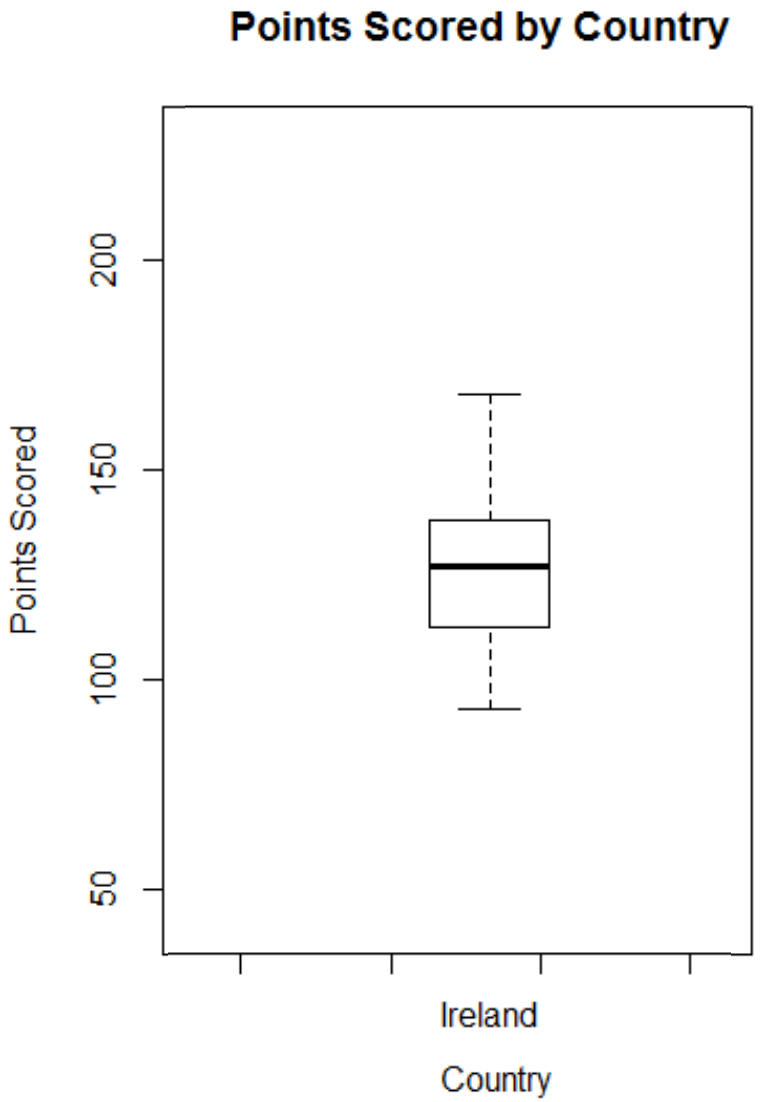

http://en.wikipedia.org/wiki/Probability_density_function

# Box Plots & Density Functions

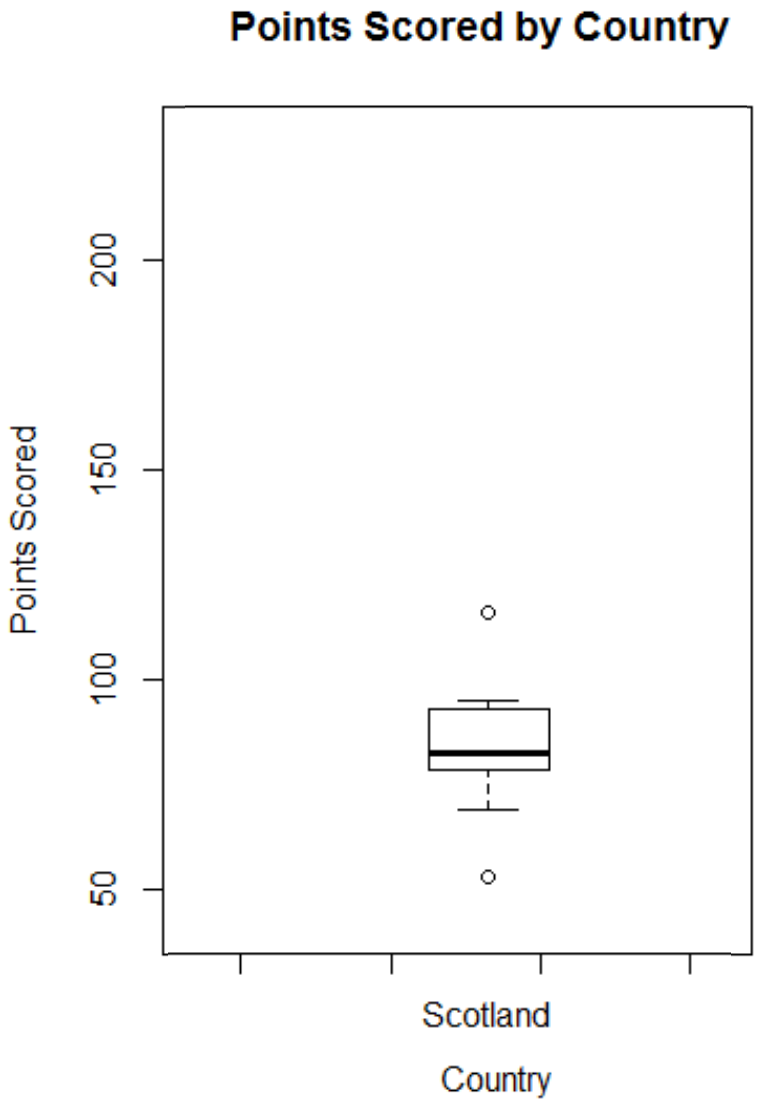



https://arc.lib.montana.edu/book/statistics-with-r-textbook/meta/img/Figure1.1.jpg

# Box Plot


Points Scored by Country
Points Scored
50
100
150
200
Ireland
Country

# Box Plot

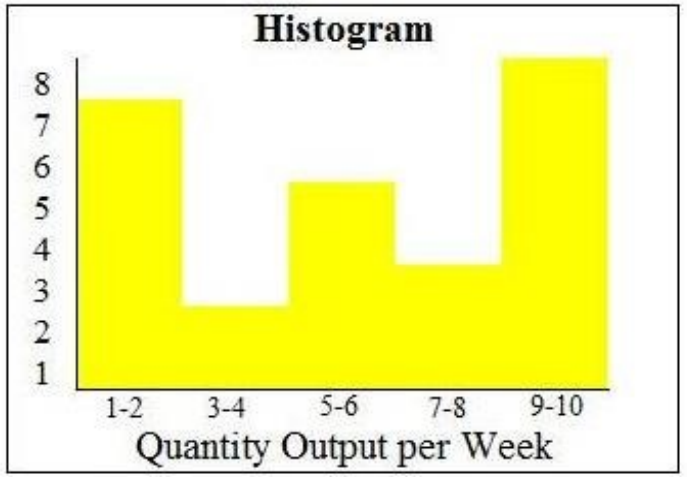

Points Scored by Country
Points Scored
50
100
150
200
Scotland
Country

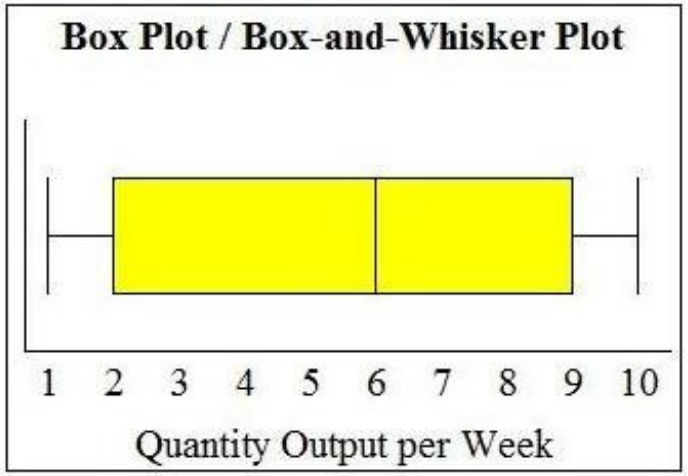
# Box Plot vs Histogram

Histogram useful when

- a wide variances exist among the observed frequencies for a particular dataset

- Very little variance amongst the observed frequencies



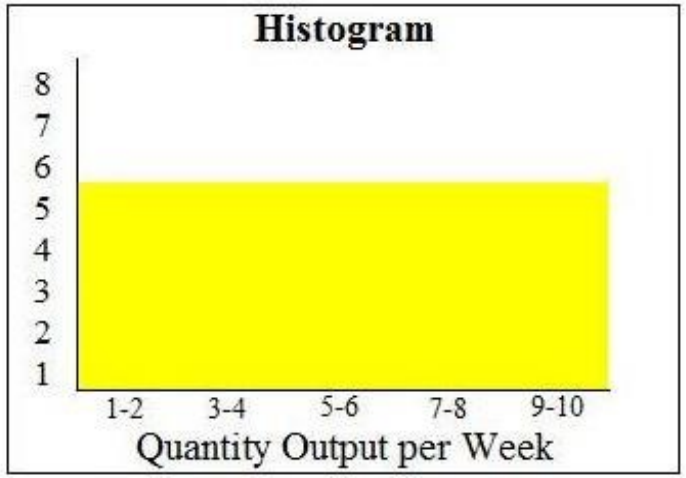

https://www.brighthubpm.com/six-sigma/58254-box-plots-vs-histograms-in-project-management/#imgn_1

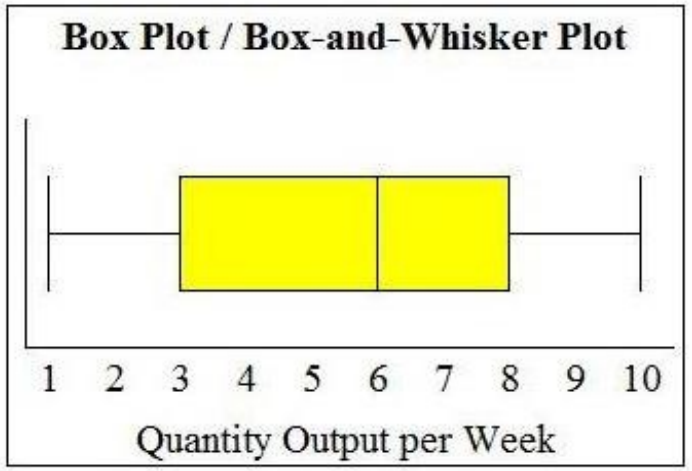
# Box Plot vs Histogram

Histogram useful when

- a wide variances exist among the observed frequencies for a particular dataset.

- Very little variance amongst the observed frequencies



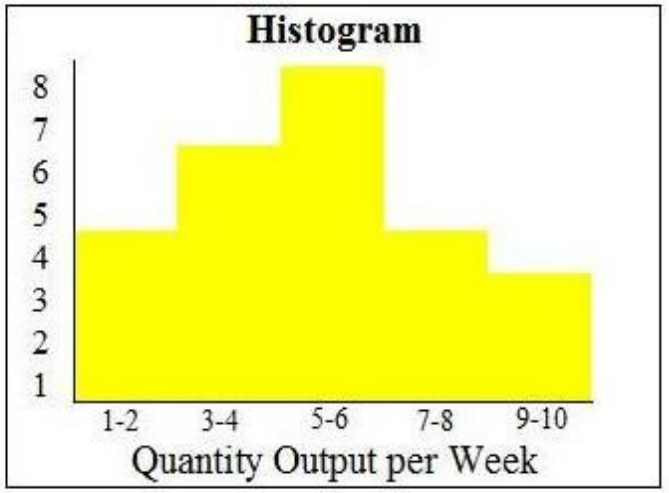



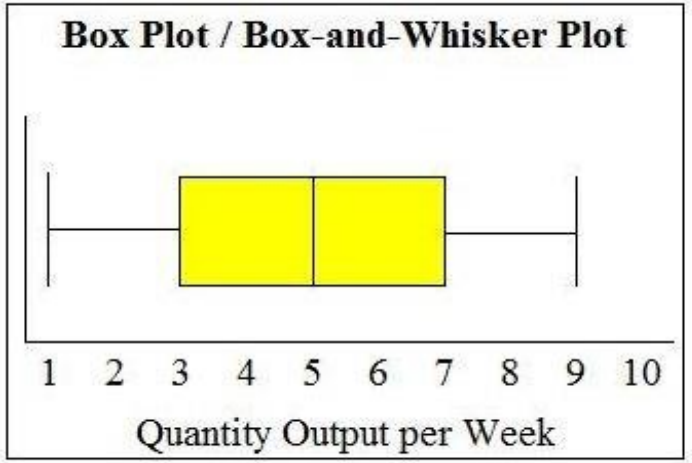

https://www.brighthubpm.com/six-sigma/58254-box-plots-vs-histograms-in-project-management/#imgn_1

# Box Plot vs Histogram

Box Plot useful when

- A moderate variance exist among the observed frequencies, which causes the histogram to look ragged and non-symmetrical due to the way the data is grouped.
- This may lead into the assumption that data is slightly skewed.
- A box plot for the same data shows a perfect normal distribution.



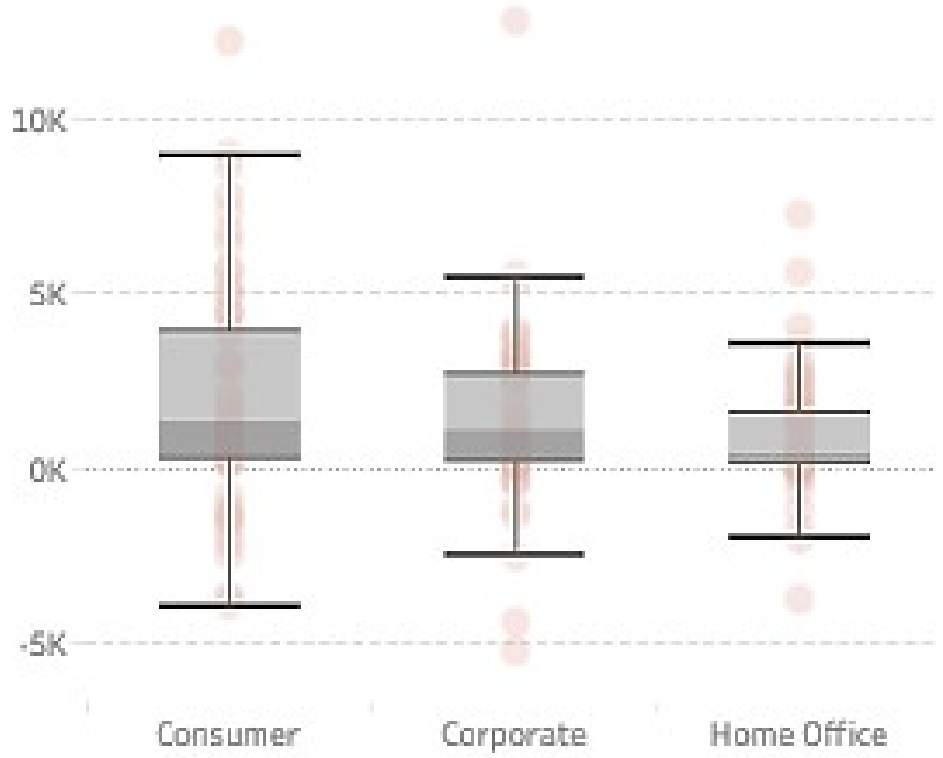




https://www.brighthubpm.com/six-sigma/58254-box-plots-vs-histograms-in-project-management/#imgn_1

# Box Plot vs Histogram

Box Plot useful when

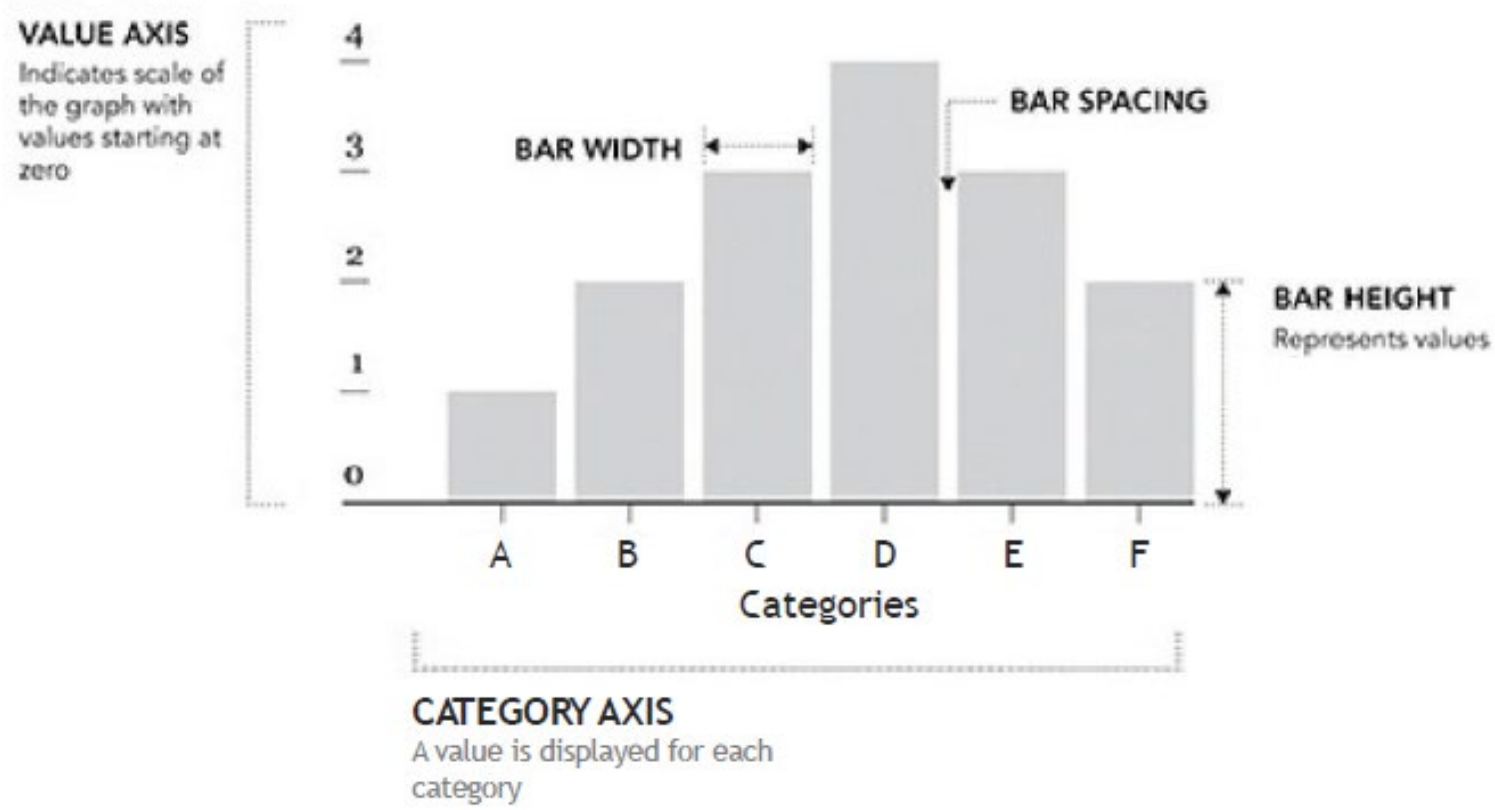
- Outliers are present in the data
- We are interested in values such as the median or the quartiles




https://www.brighthubpm.com/six-sigma/58254-box-plots-vs-histograms-in-project-management/#imgn_1

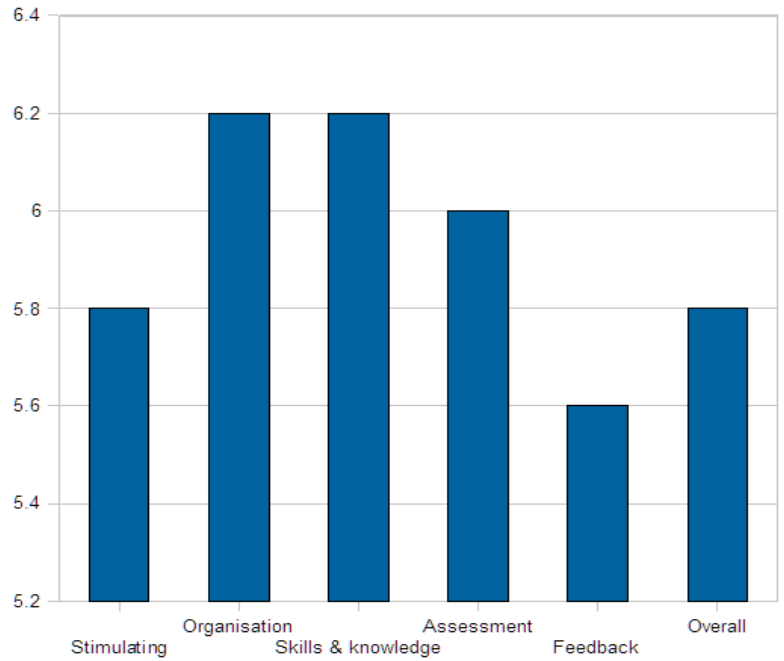
# SIMPLE COMPARISONS

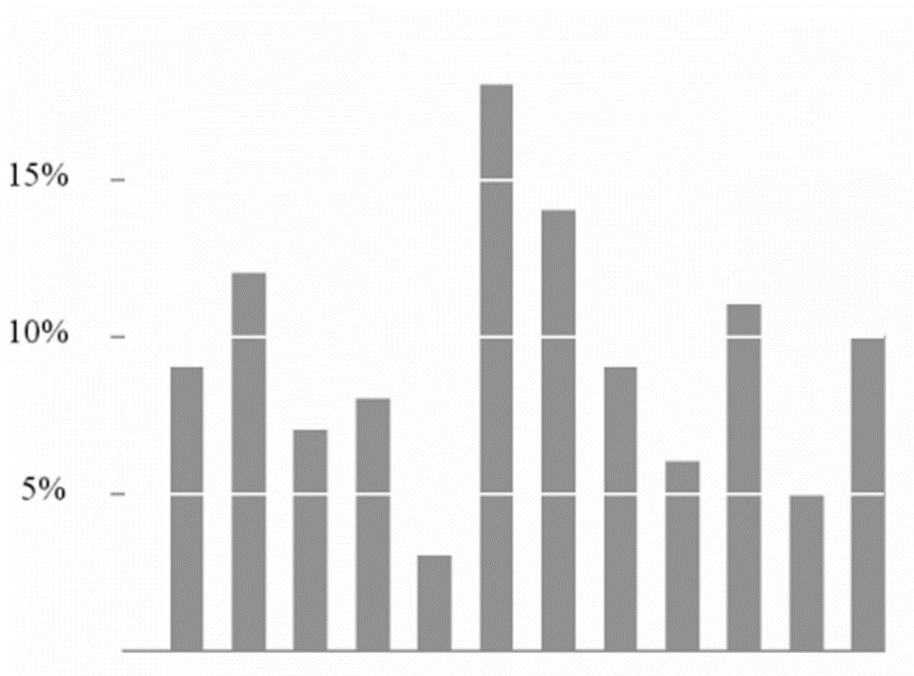
# Simple Bar Graph




Visualize This: The Flowing Data Guide to Design, Visualization, and Statistics. Nathan Yau. 2011

# Simple Bar Graph

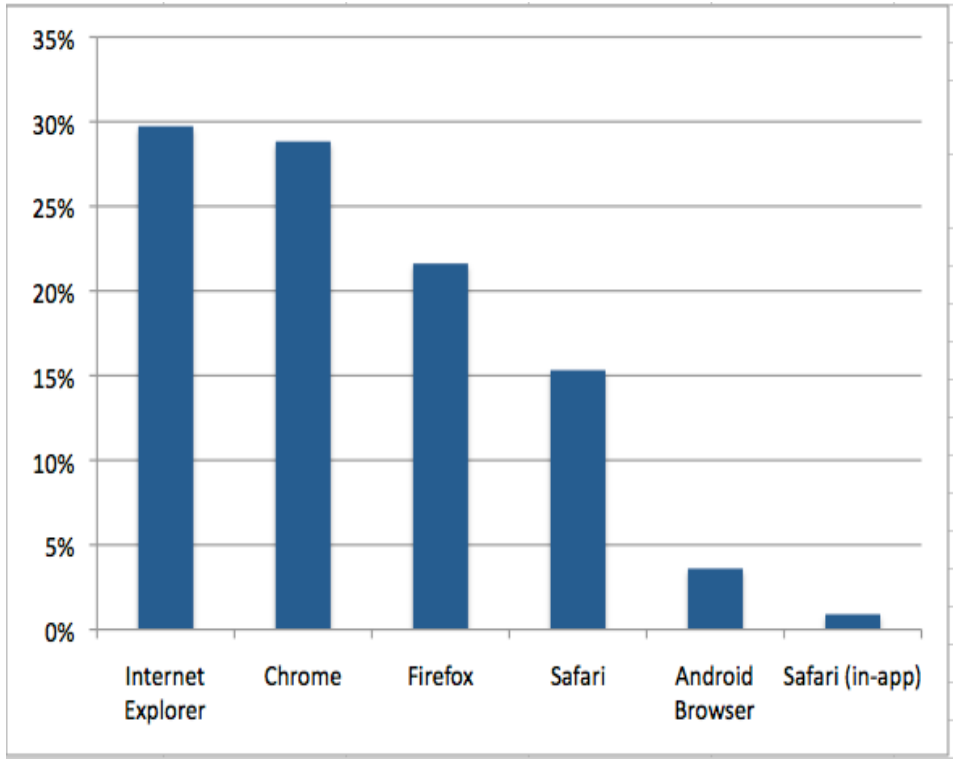


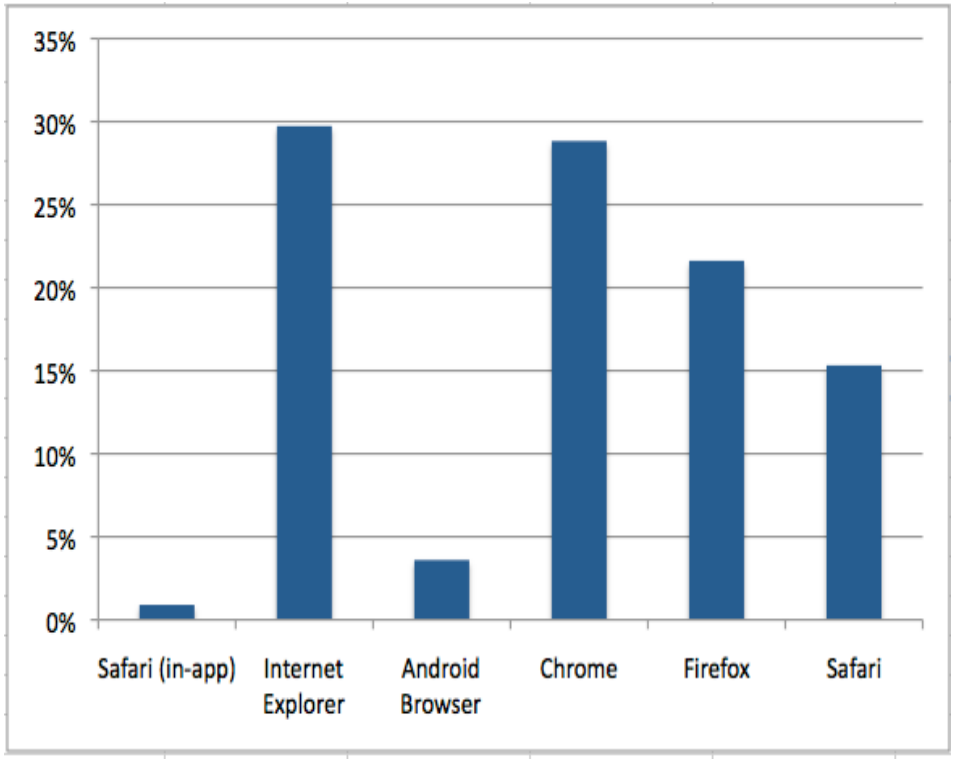




- Tufte removes the grid and uses simple labeling on the y- axis, and a line to illustrate quantity.

Edward Tufte, “The Quantitative Display of Information”, 2009

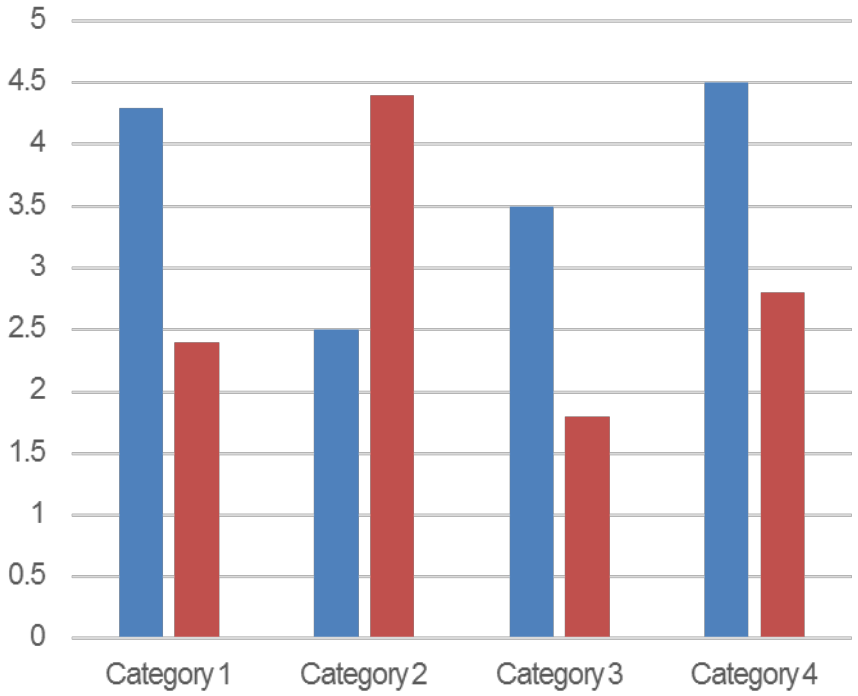
# Simple Bar Graph

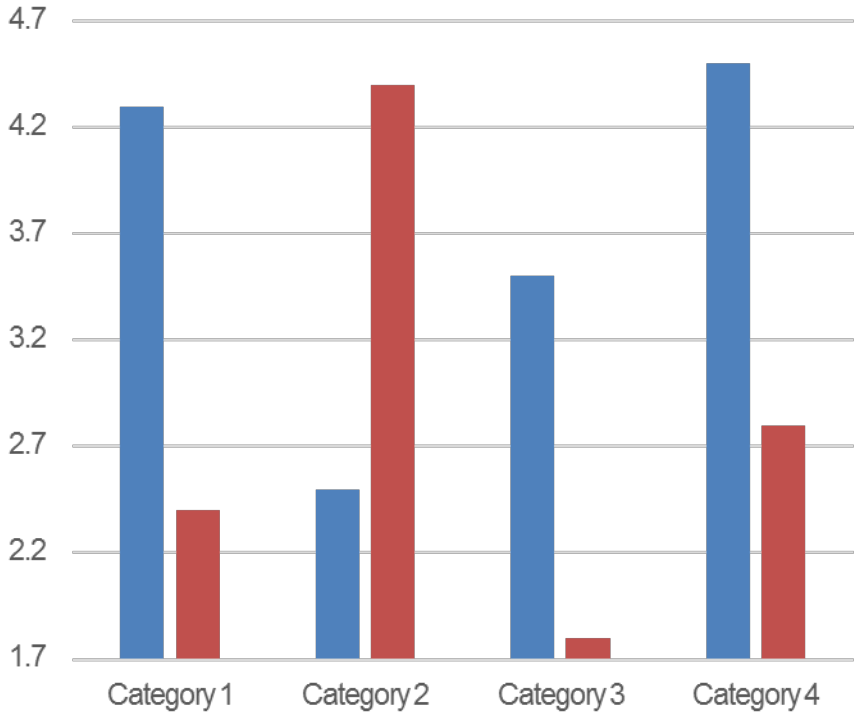



- Always show increase or decrease – do not have different categories without some illustration.

# Simple Bar Graph

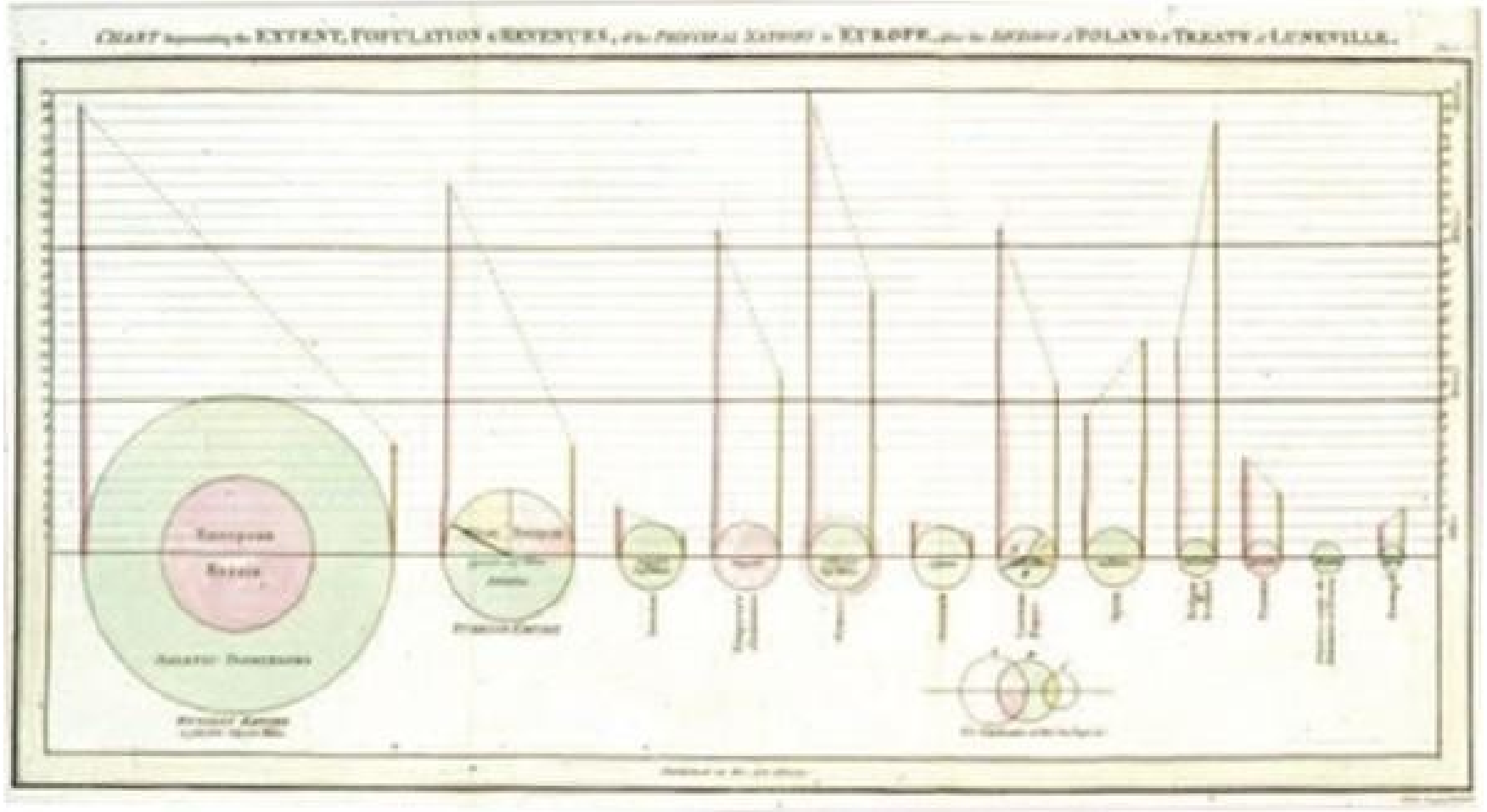



- If you have multiple bar charts, ensure that they have the same scale.
- Start from 0.

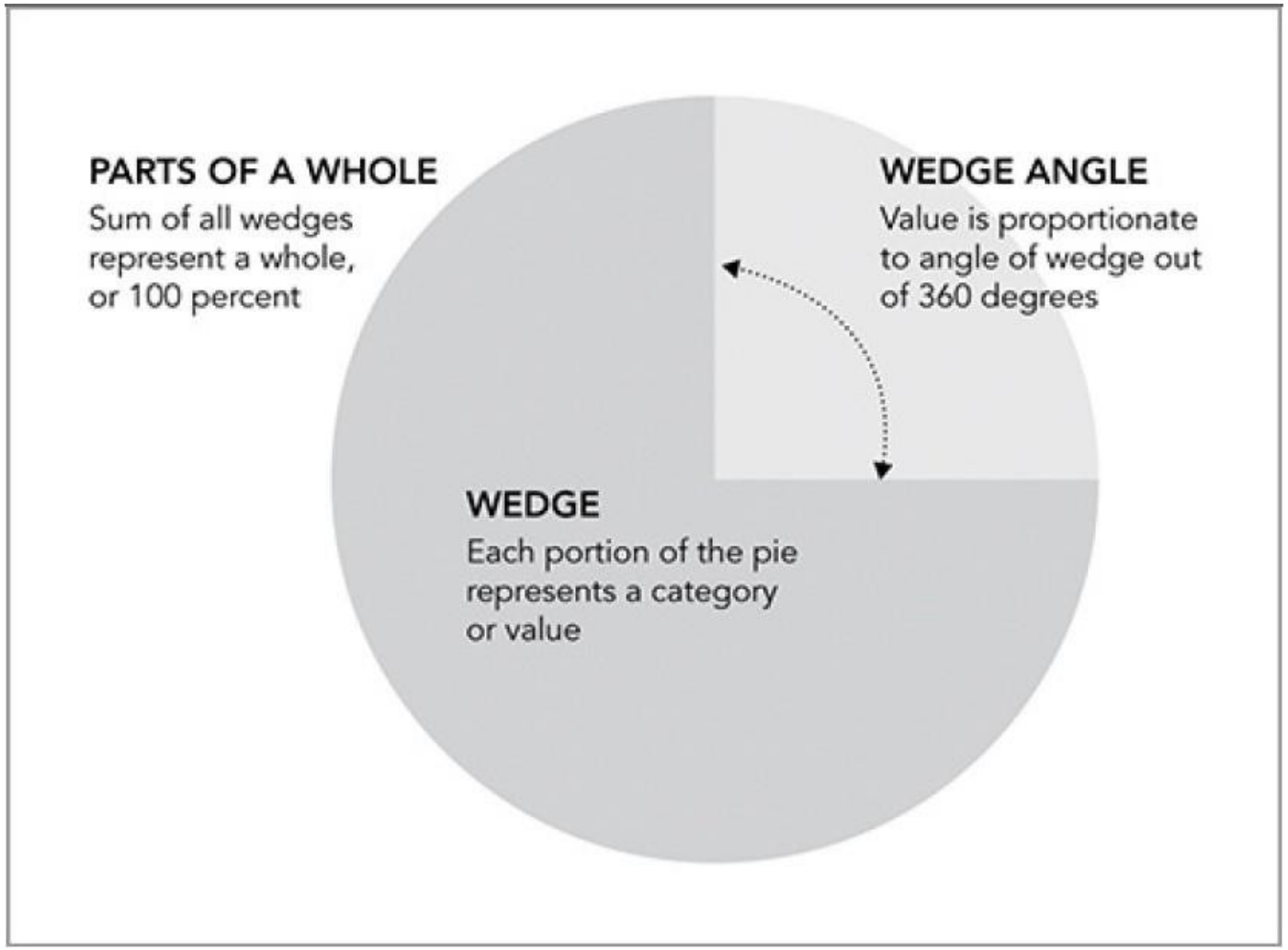
# PARTS OF A WHOLE CHARTS

# Pie Charts



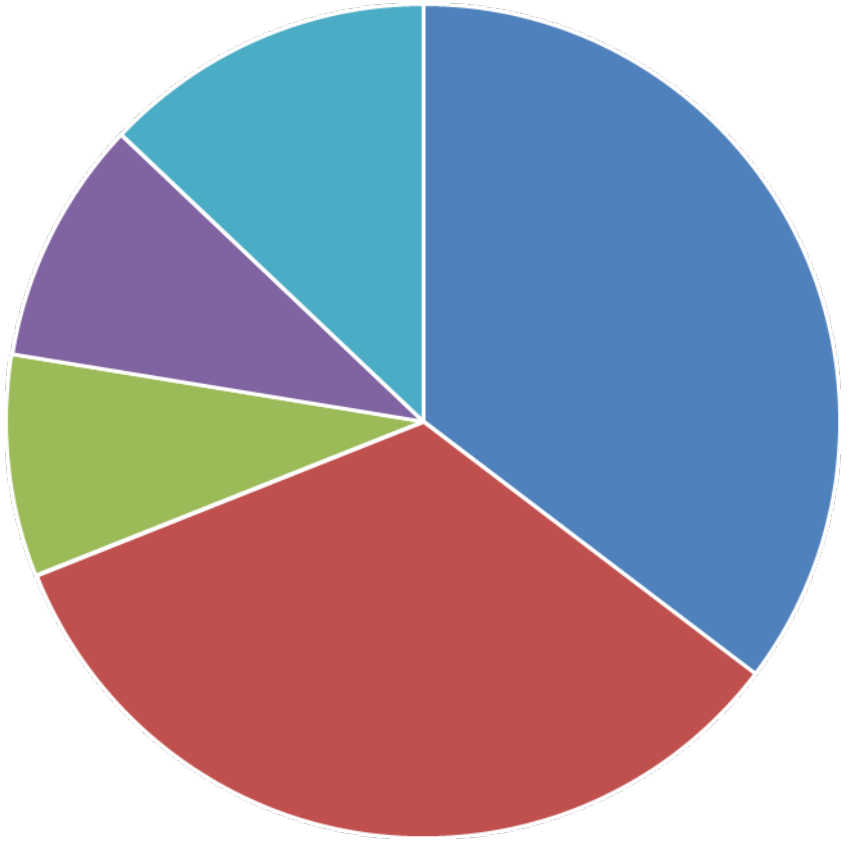
William Playfair's "Statistical Breviary," 1801 via The New York Times http://www.nytimes.com/2012/04/22/magazine/who-made-that-pie-chart.html?_r=0

# Pie Charts

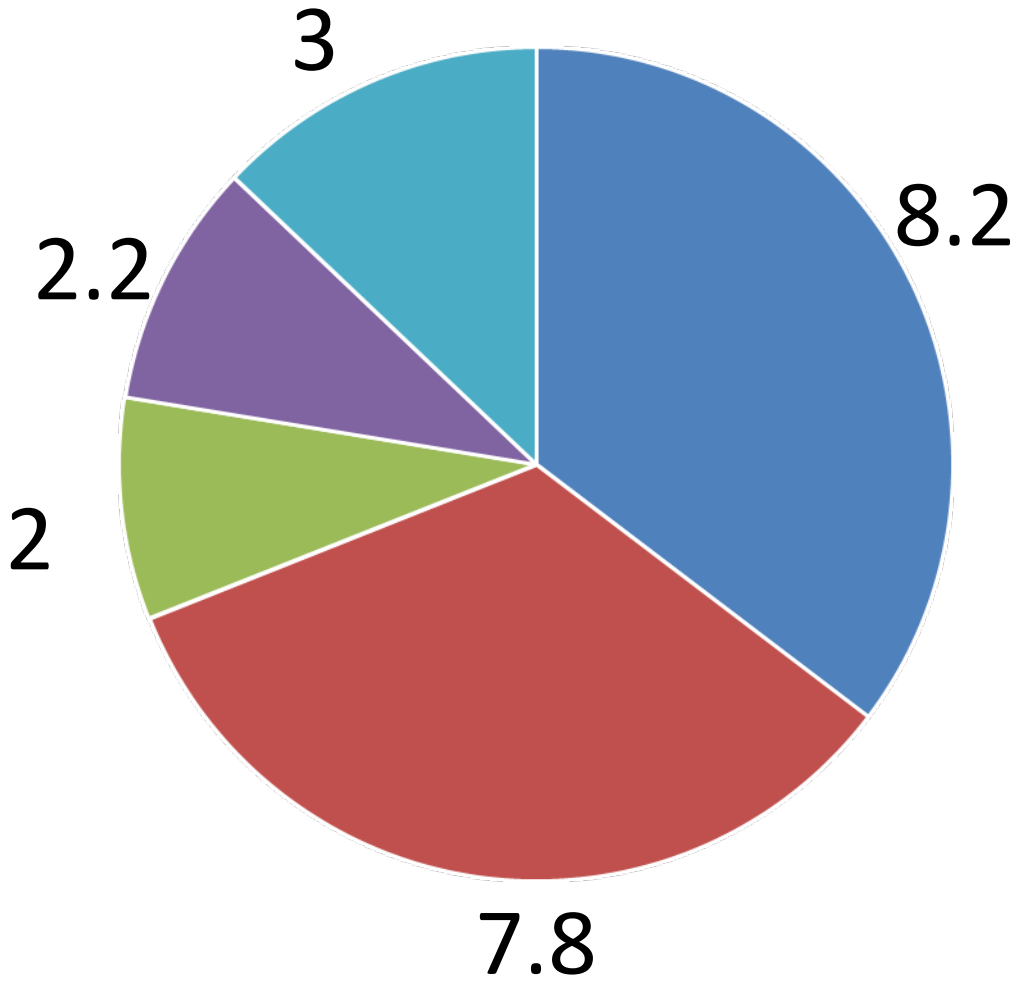



Visualize This: The Flowing Data Guide to Design, Visualization, and Statistics. Nathan Yau. 2011

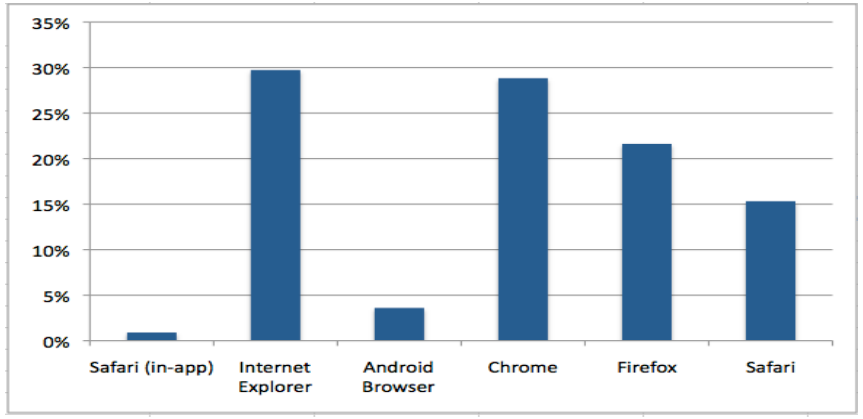
# Pie Charts

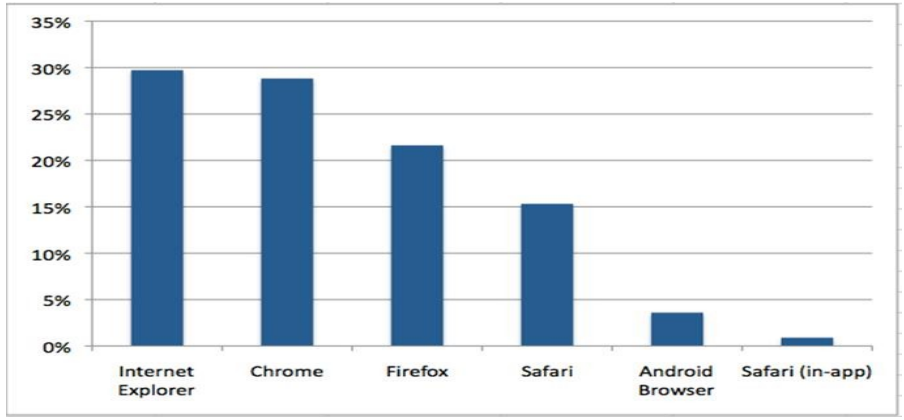


# Pie Charts

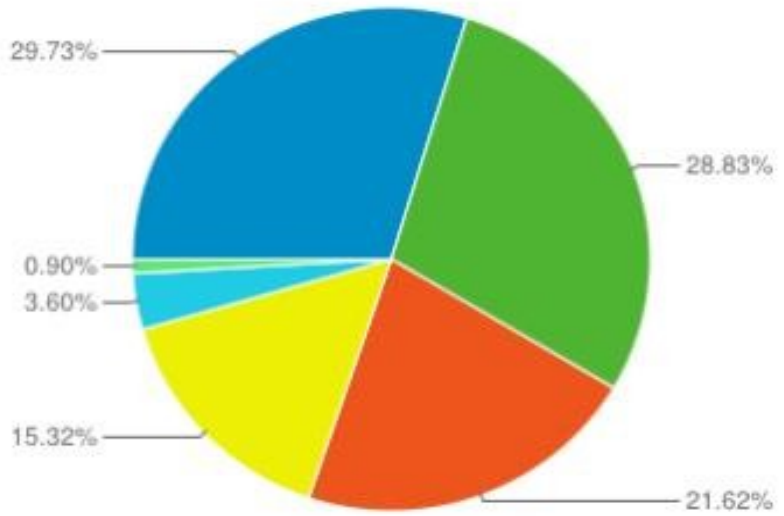


# Pie Charts




| | | | |
|---|---|---|---|
| 1. | Internet Explorer | 33 | 29.73% |
| 2. | Chrome | 32 | 28.83% |
| 3. | Firefox | 24 | 21.62% |
| 4. | Safari | 17 | 15.32% |
| 5. | Android Browser | 4 | 3.60% |
| 6. | Safari (in-app) | 1 | 0.90% |




http://www.uh.edu/engines/epi1712.htm

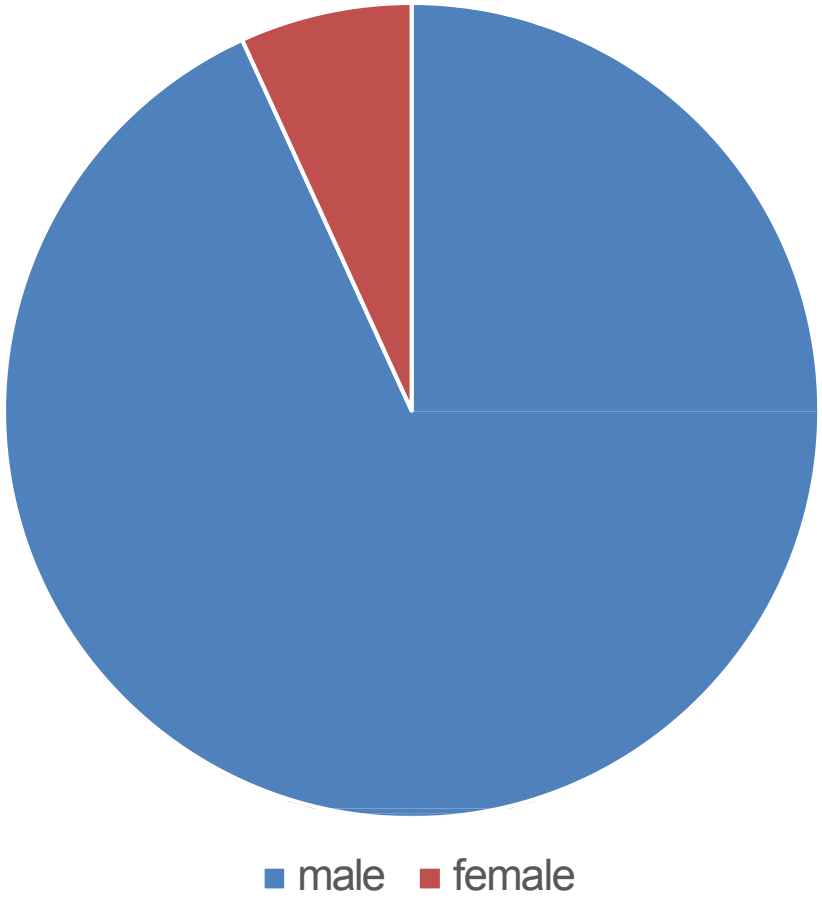
# Pie Charts

- Pie charts are the subject of a lot of negative comment
  - The main reason is that their descriptive power is based on our ability to interpret differences in angle

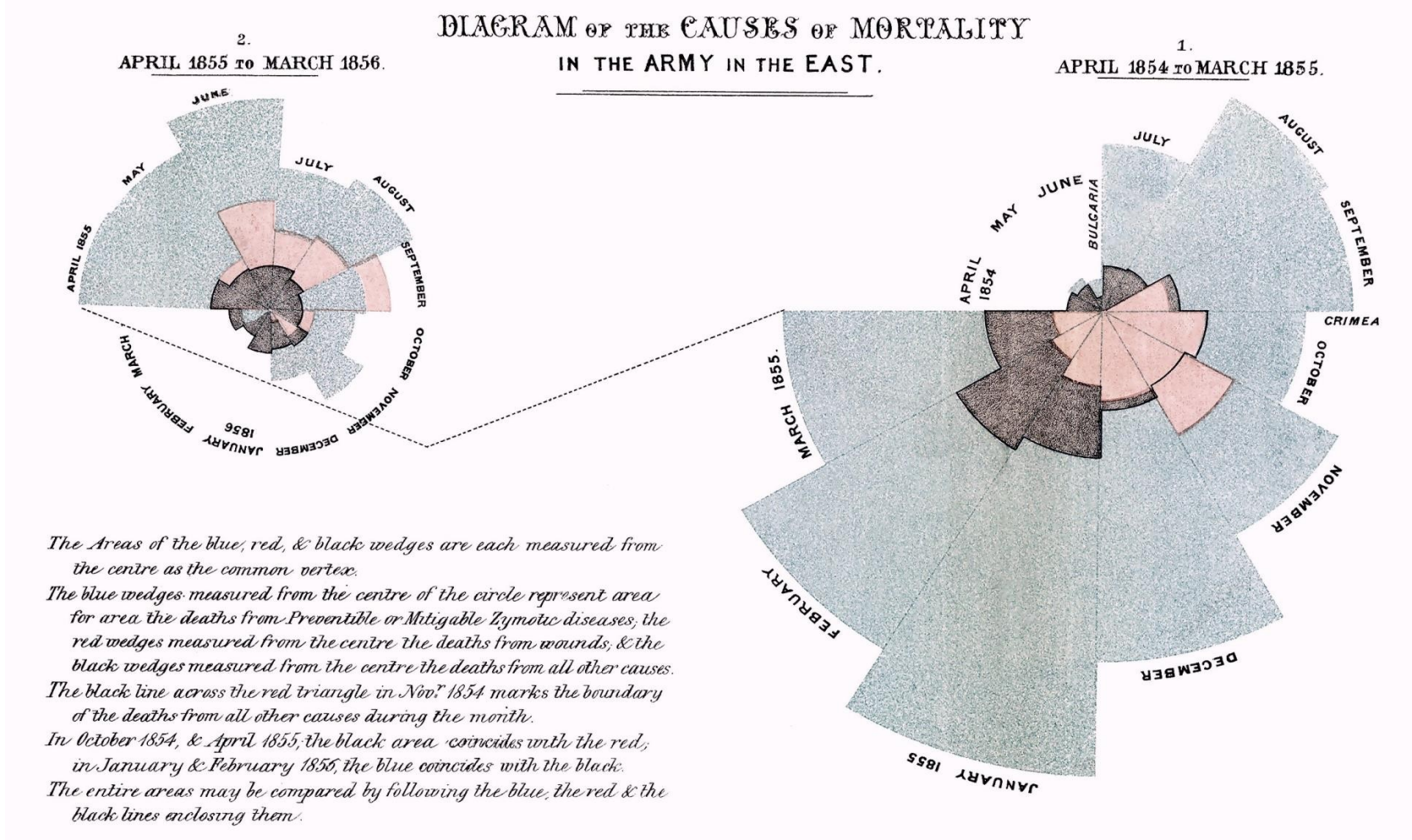
- Pie charts are useful when:
  - We have a small number of categories (< 8)
  - The values sum to a meaningful whole
  - The differences are coarse

- “We cannot easily rank categories when using a pie chart “ Stephen Few

# Pie Charts



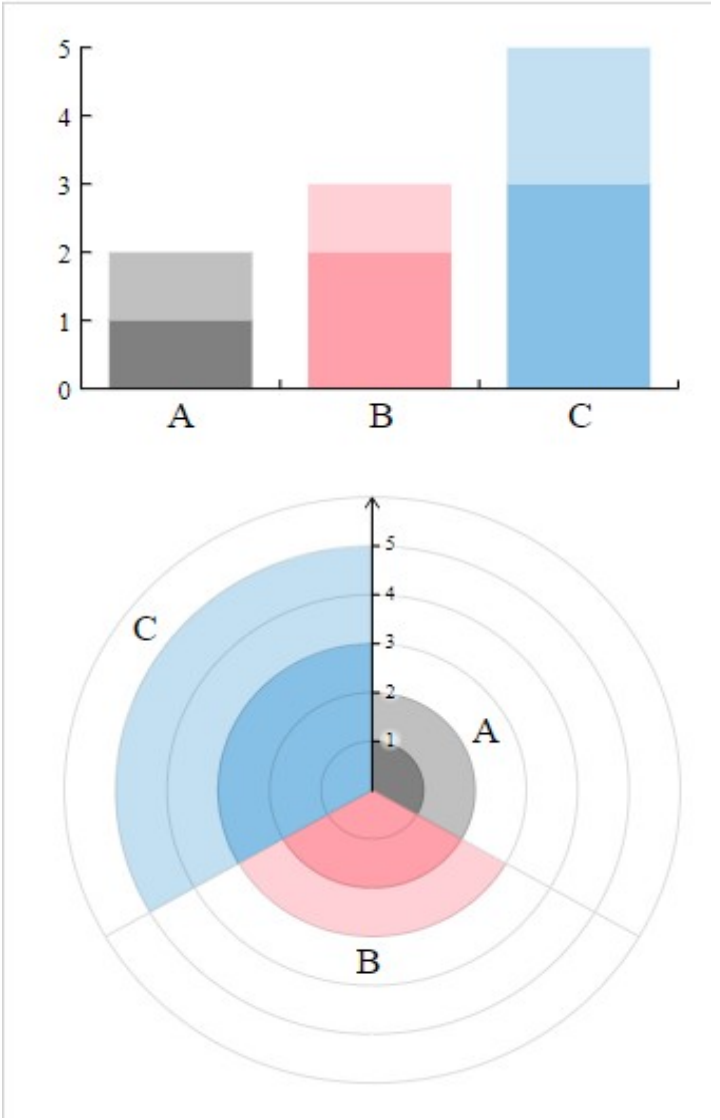
# Nightingale Rose Charts




Florence Nightingale's Crimean War Death Charts via: http://www.uh.edu/engines/epi1712.htm

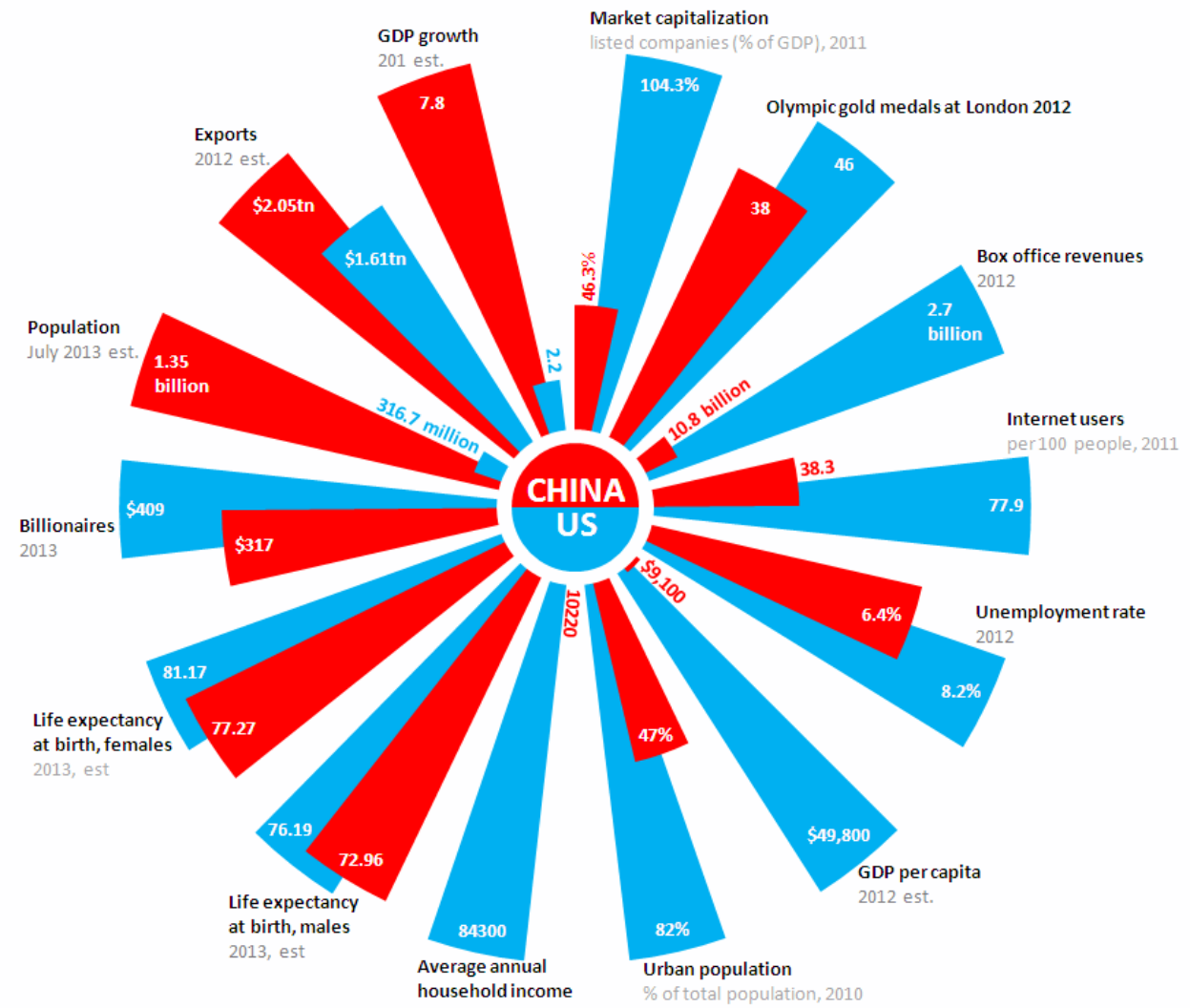
# Nightingale Rose Charts

- Nightingale Rose Charts are drawn on a polar coordinate grid.
- Each category or interval in the data is divided into **equal segments** on this radial chart.
- How far each segment extends from the centre of the polar axis depends on the value it represents.
- So each ring from the centre of the polar grid can be used as a scale to plot the segment size and represent a higher value.
- Therefore, it's important to notice with Nightingale Rose Charts that it's the area, rather than the radius of a segment that represents its value.



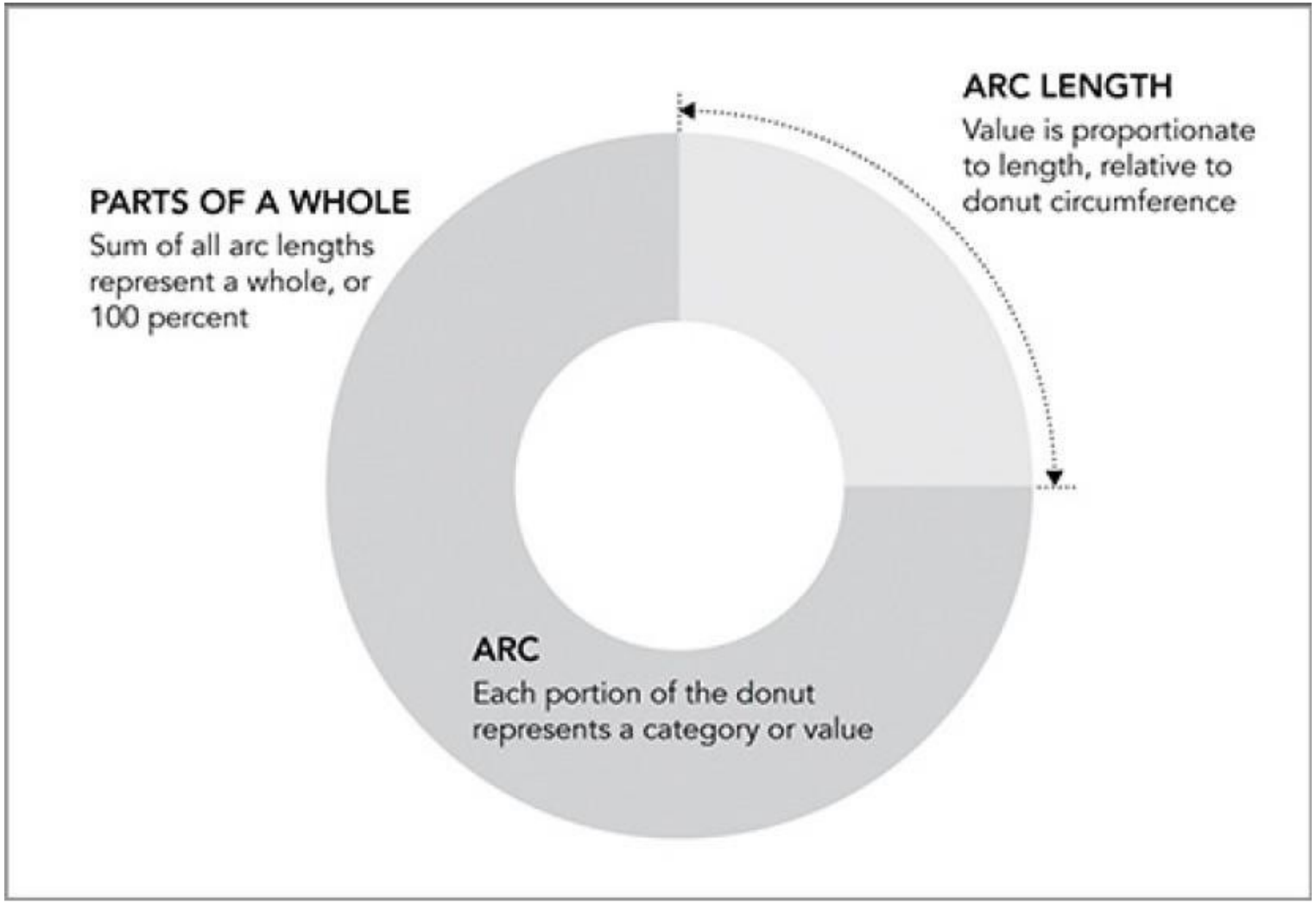

Florence Nightingale's Crimean War Death Chart https://datavizcatalogue.com/methods/nightingale_rose_chart.html
https://boraberan.wordpress.com/2014/03/30/creating-coxcomb-charts-in-tableau/

# Rose Charts




https://www.excelhowto.com/nightingale-rose-diagram-excel-template/

# Doughnut Chart

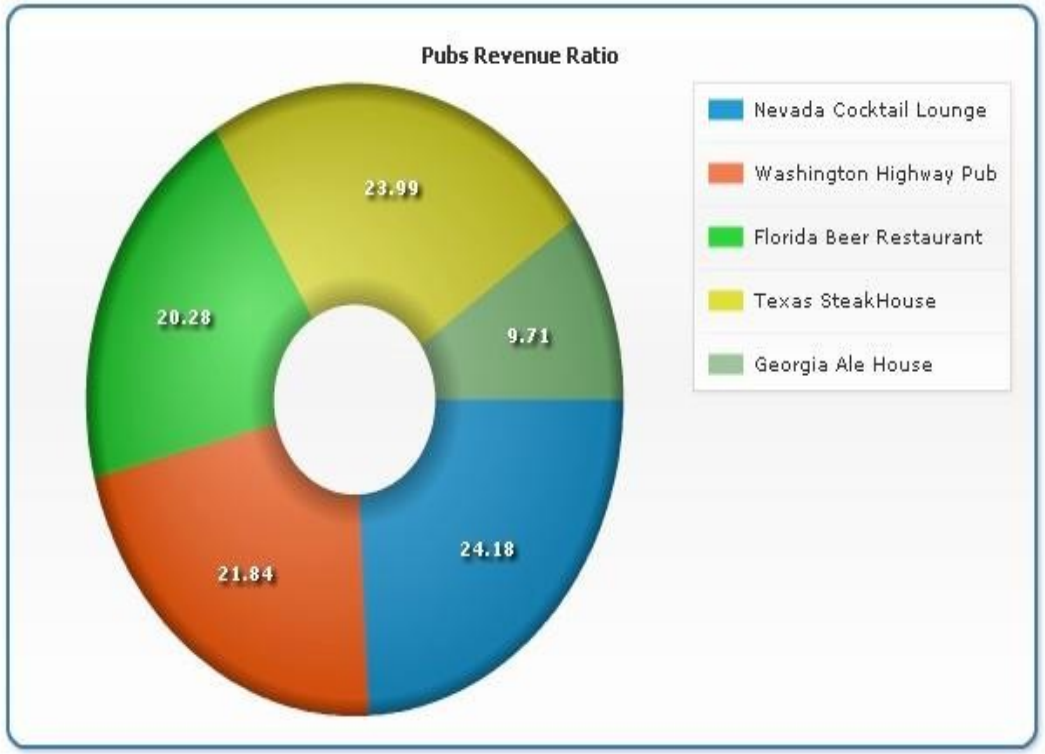



Visualize This: The Flowing Data Guide to Design, Visualization, and Statistics. Nathan Yau. 2011

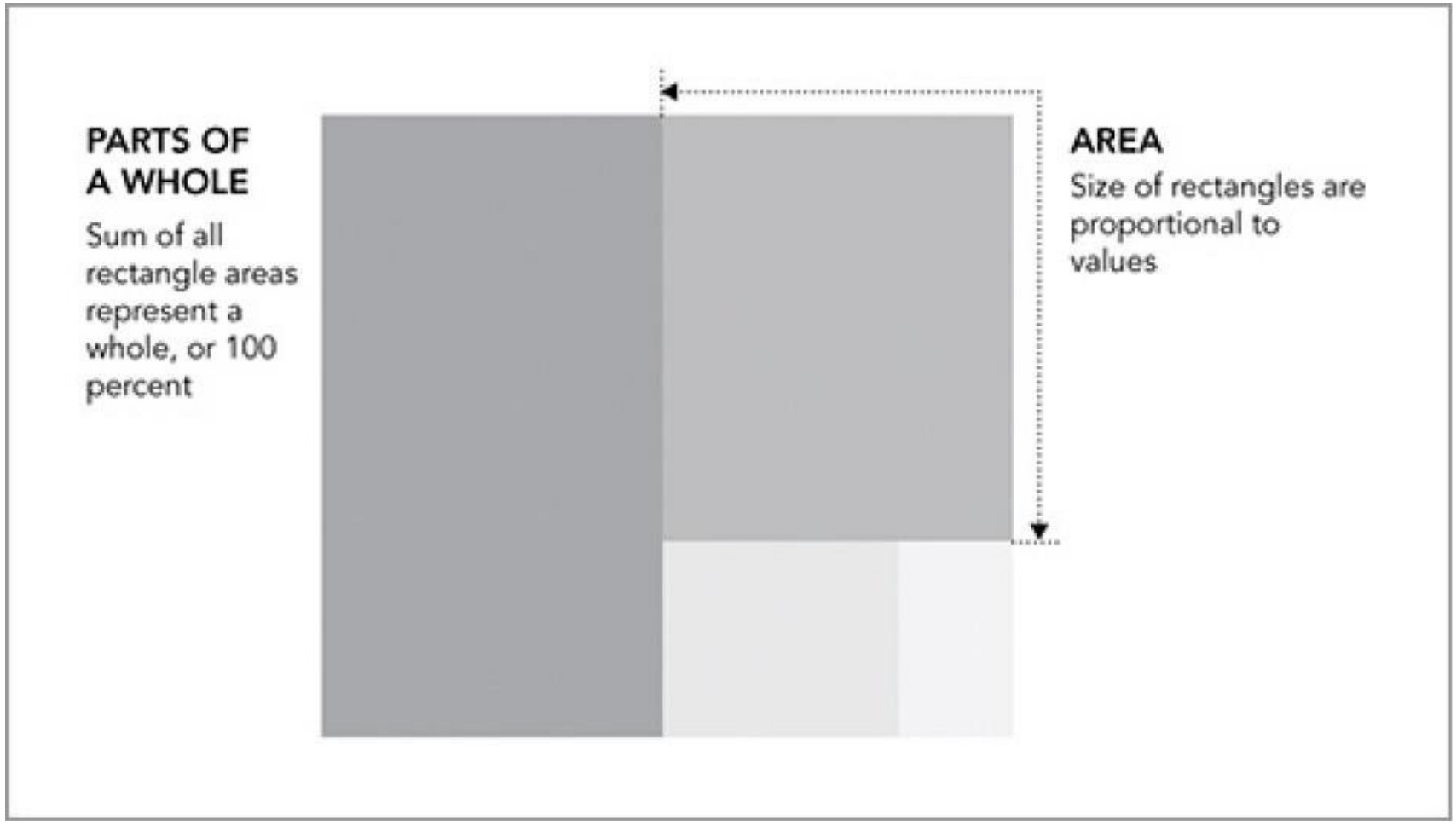
# Doughnut Chart

- Sections within a Donut Chart may be hard to compare to each other

- Use of Donut Chart
  - Compare an individual section to the whole Part of a whole -> must add up to 100% No more than six categories.

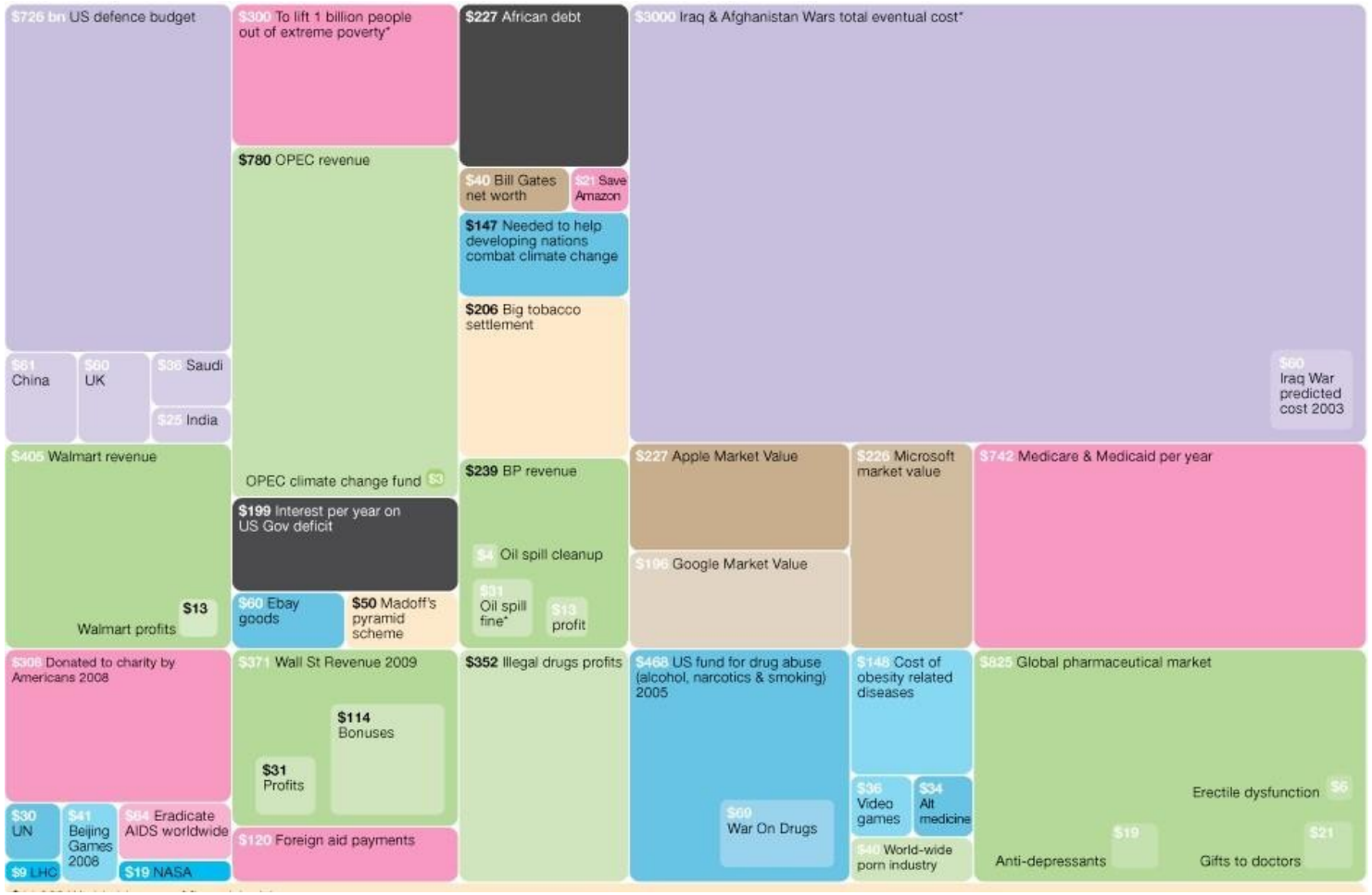


# Tree Map




Visualize This: The Flowing Data Guide to Design, Visualization, and Statistics. Nathan Yau. 2011

# Billion-Dollar-O-Gram




www.informationisbeautiful.net/2009/the-billion-dollar-gram/

# Treemaps

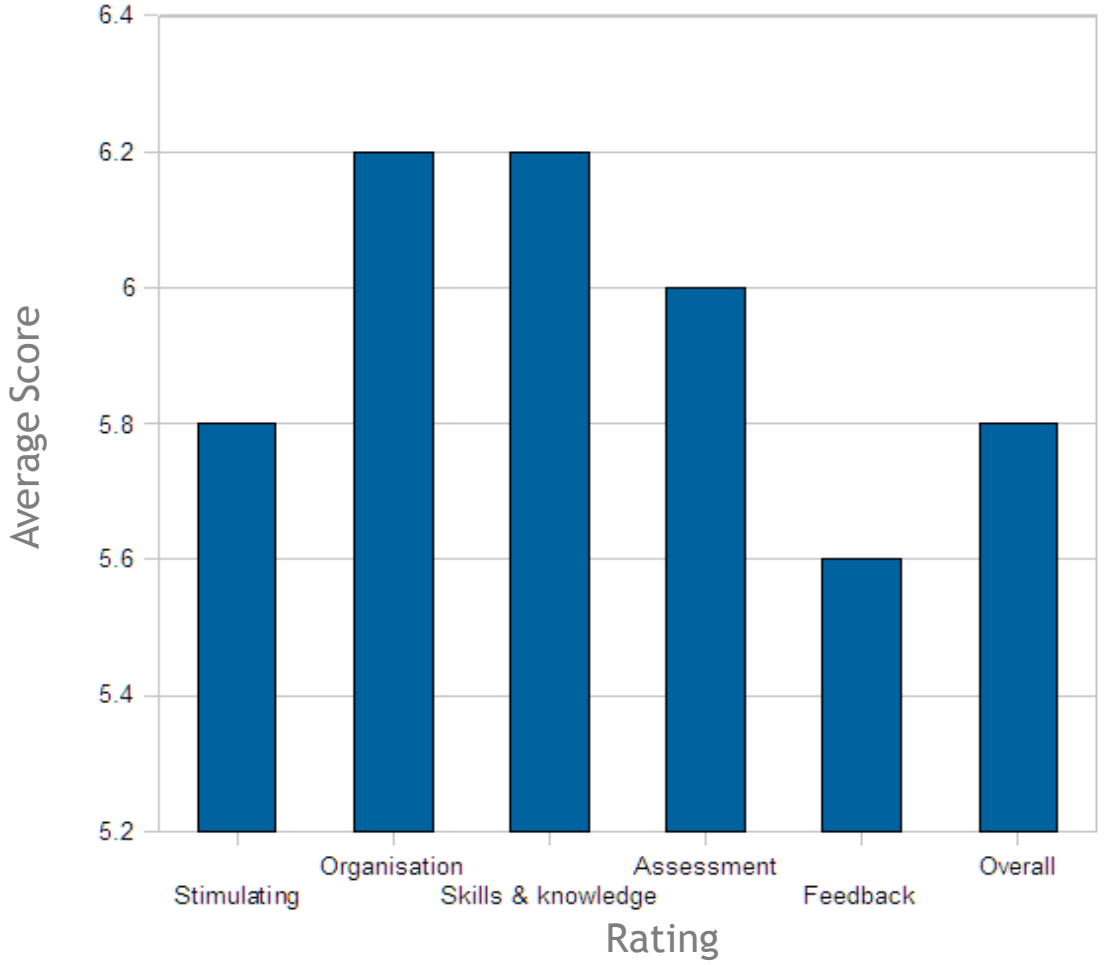
- Treemaps were originally designed to handle hierarchical structures – such as disk drives – but can be used for non-hierarchical data

- Treemaps rely on a tiling algorithm to figure out how to position the rectangles

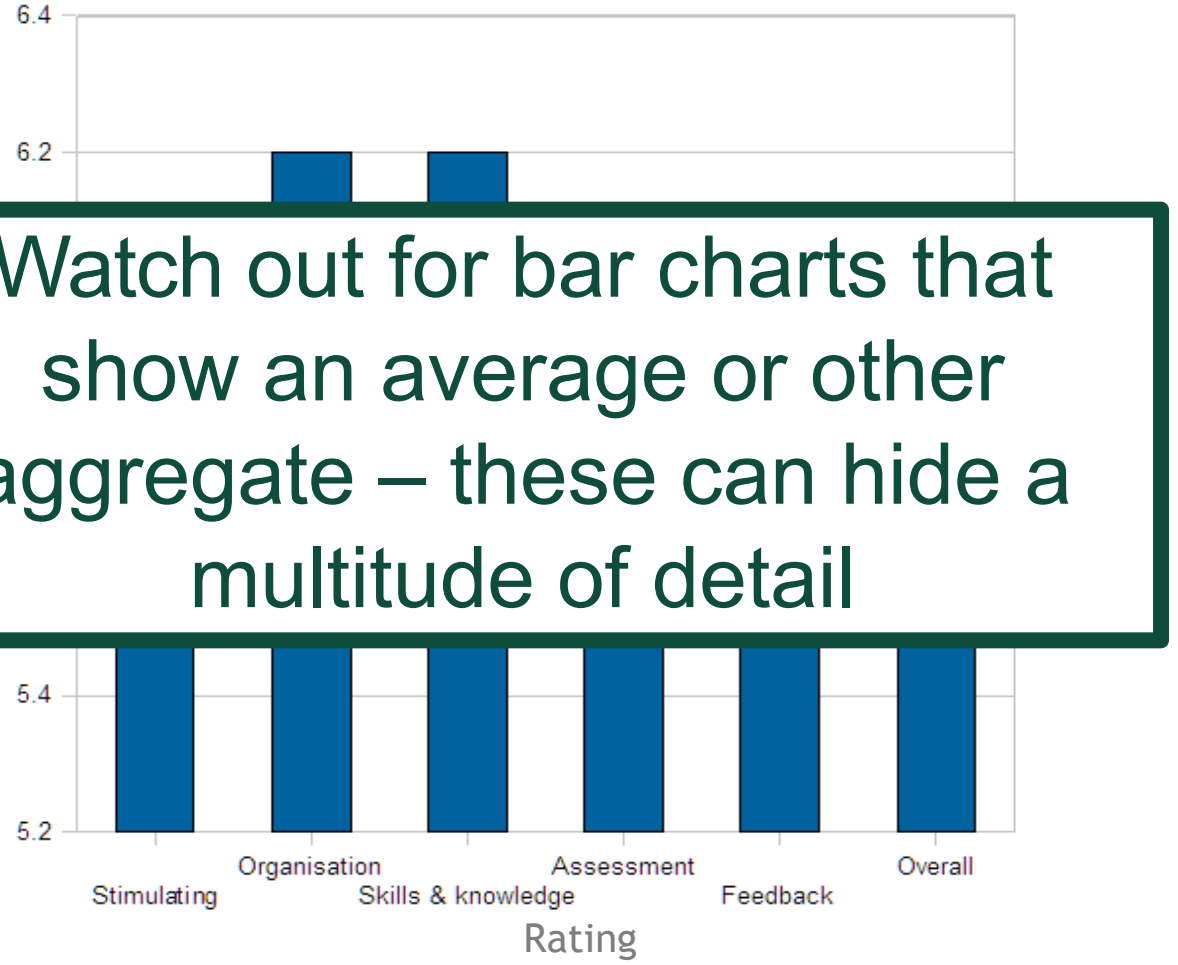
- For more:
  - TreeMap page by Ben Schneiderman (TreeMap Pioneer): http://www.cs.umd.edu/hcil/treemap-history/index.shtml
  - Early paper on TreeMaps: http://ieeexplore.ieee.org/xpl/freeabs_all.jsp?isNumber=4467&arNumber=175815&isnumber=4467&arnumber=175815

# MULTI DISTRIBUTION COMPARISONS

# Aggregate Values

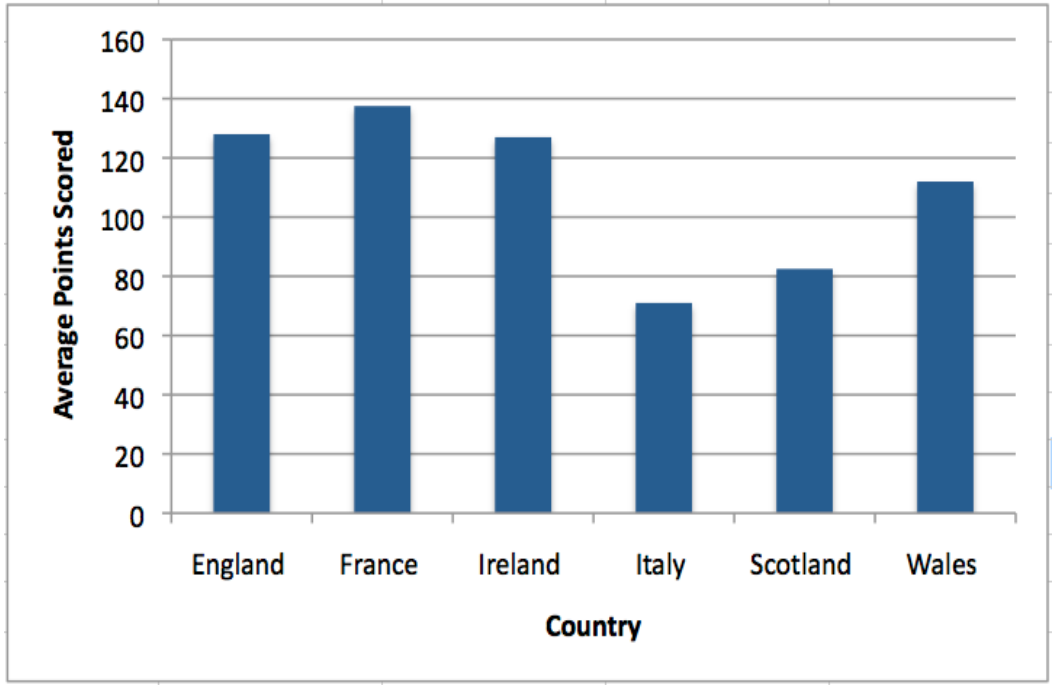


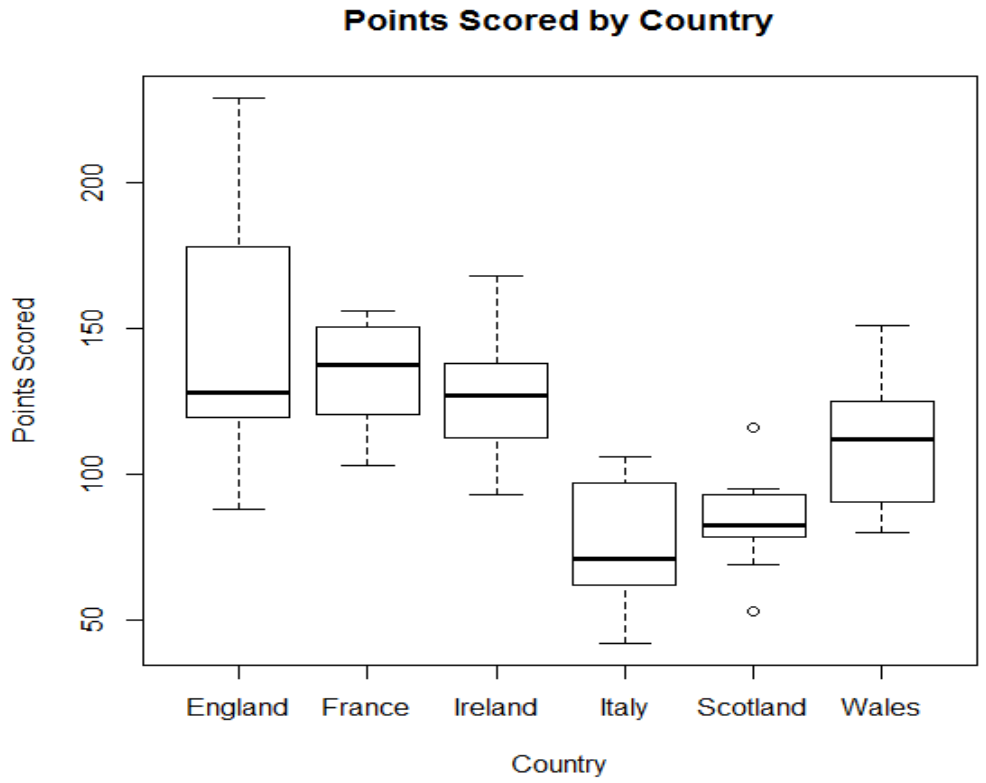
# Aggregate Values




Watch out for bar charts that show an average or other aggregate – these can hide a multitude of detail

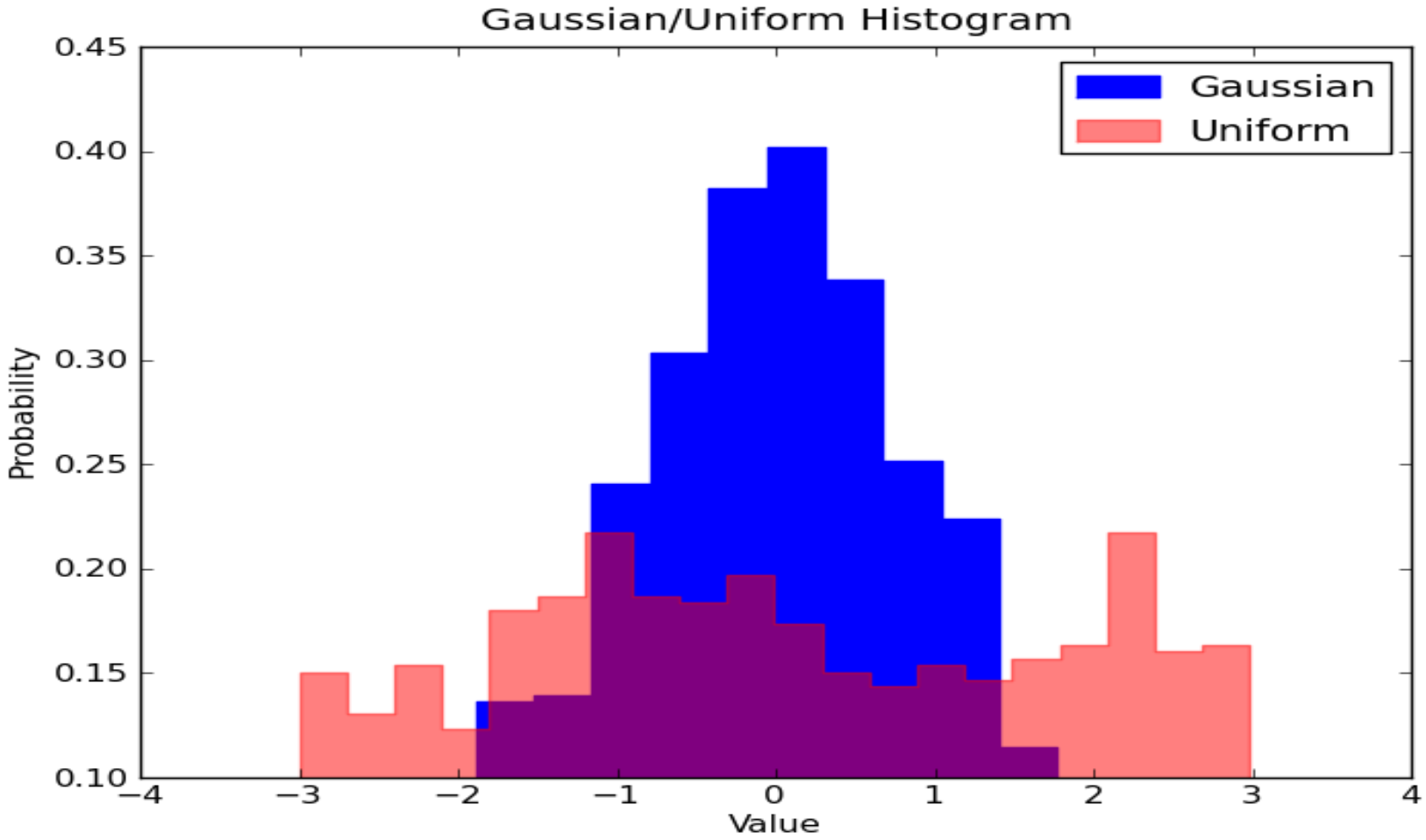
# Six Nation Points



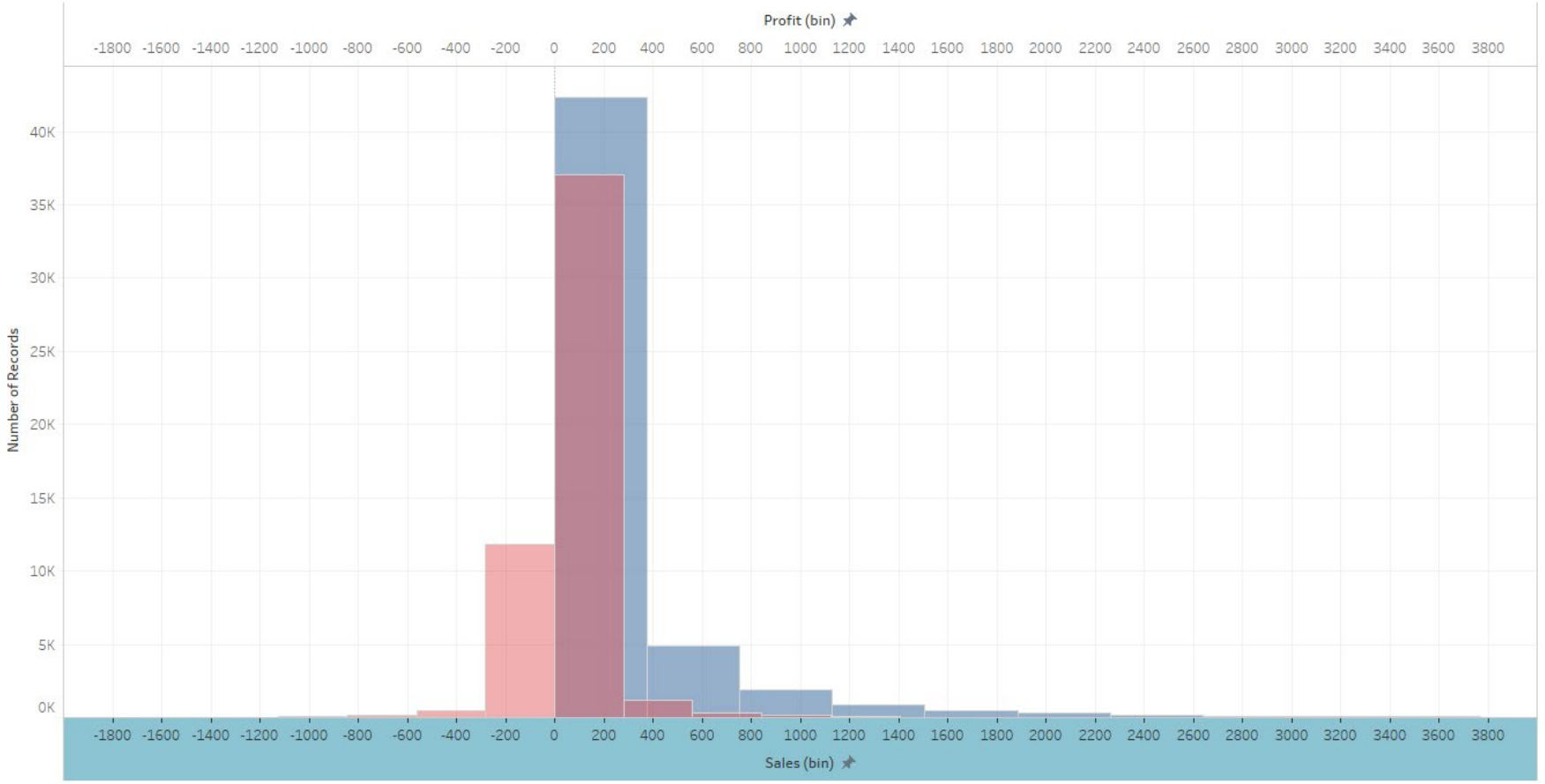

- Multiple box plots are a great way to show multiple distributions

# Overlaid Histograms




https://www.kdnuggets.com/2018/07/5-quick-easy-data-visualizations-python-code.html

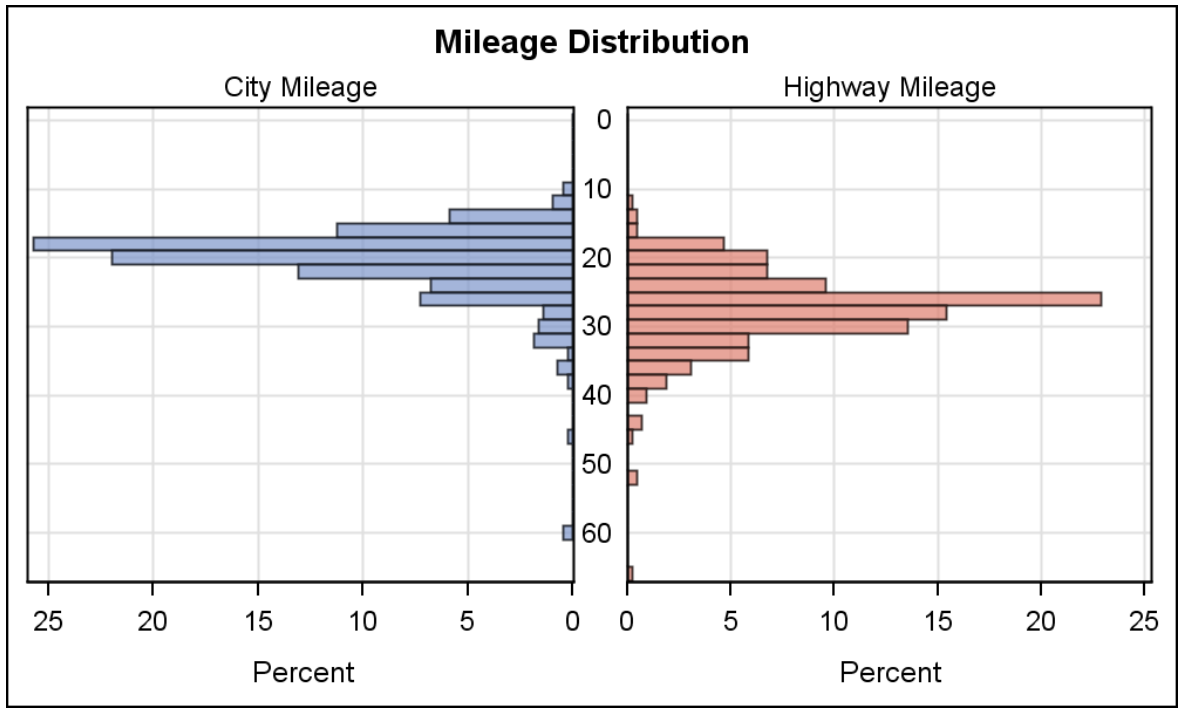
# Overlaid Histograms

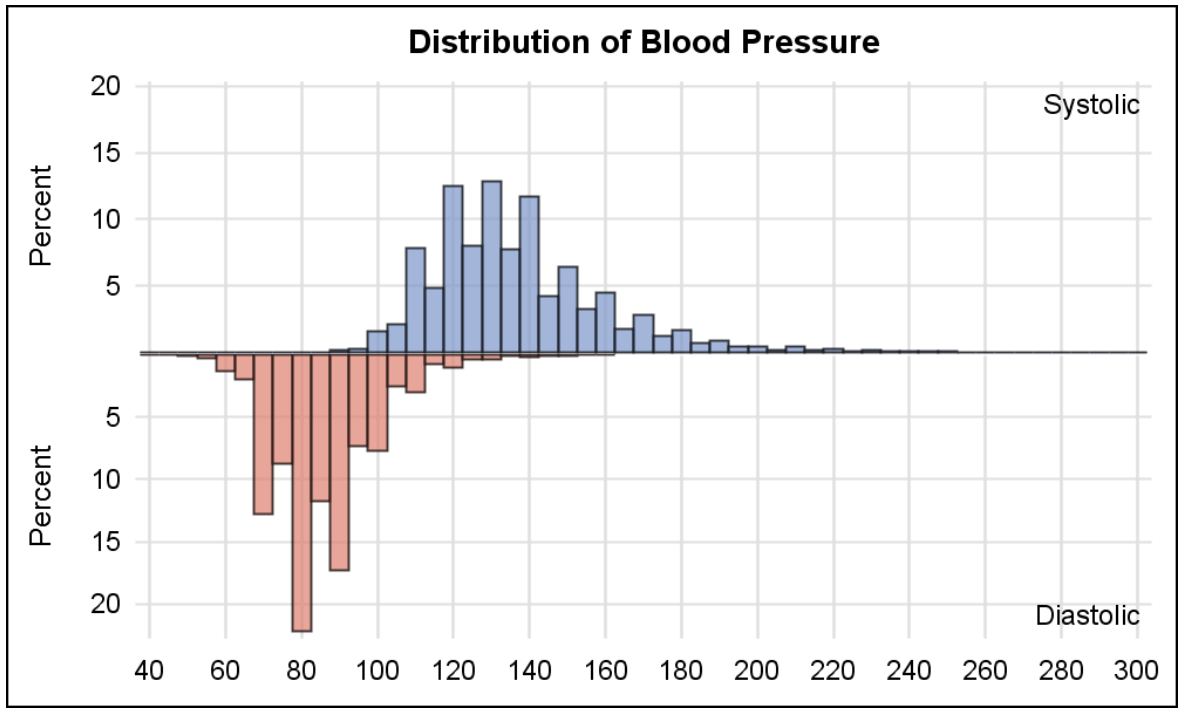



54

https://kb.tableau.com/articles/howto/create-overlapping-histograms

# Back-to-Back Histograms

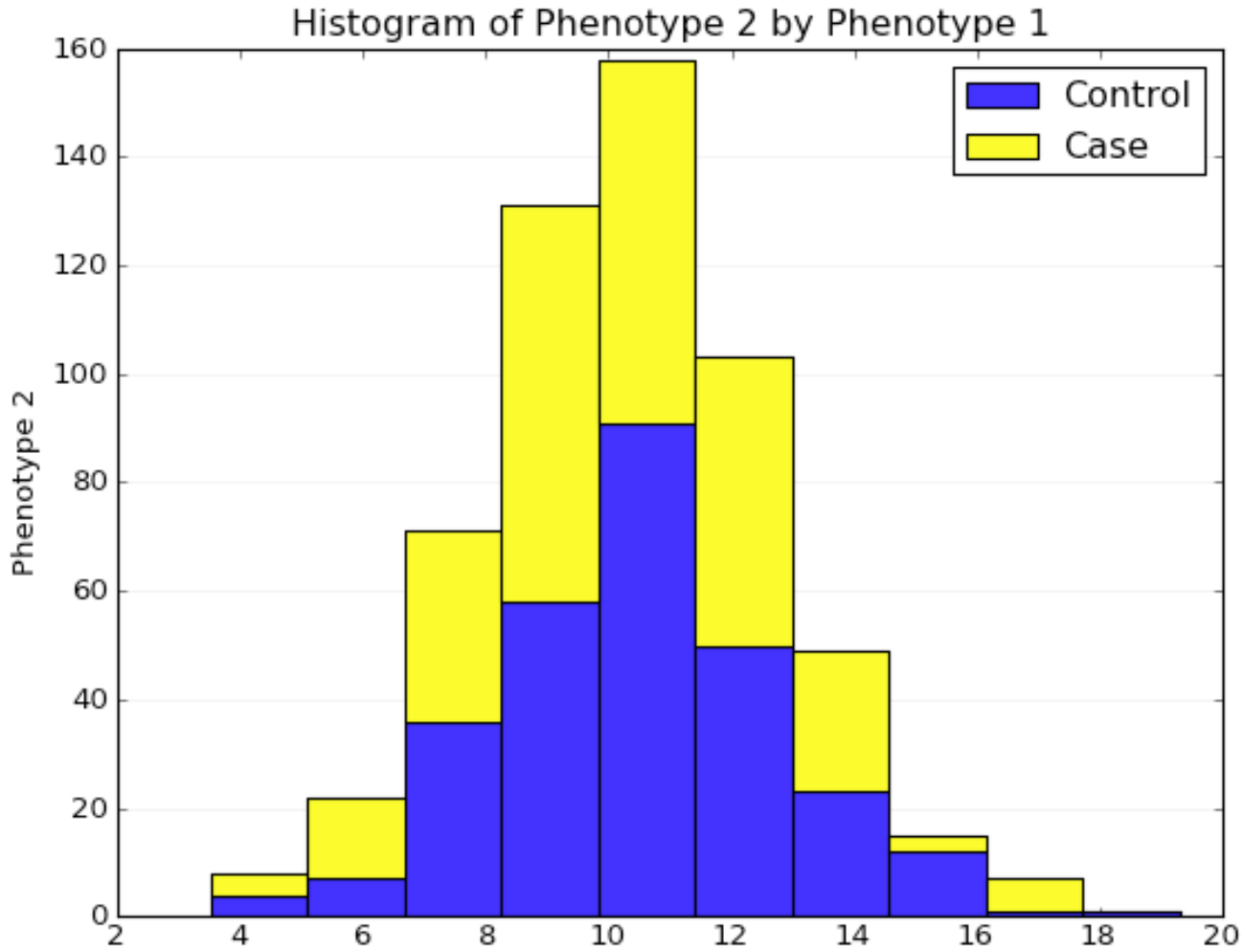


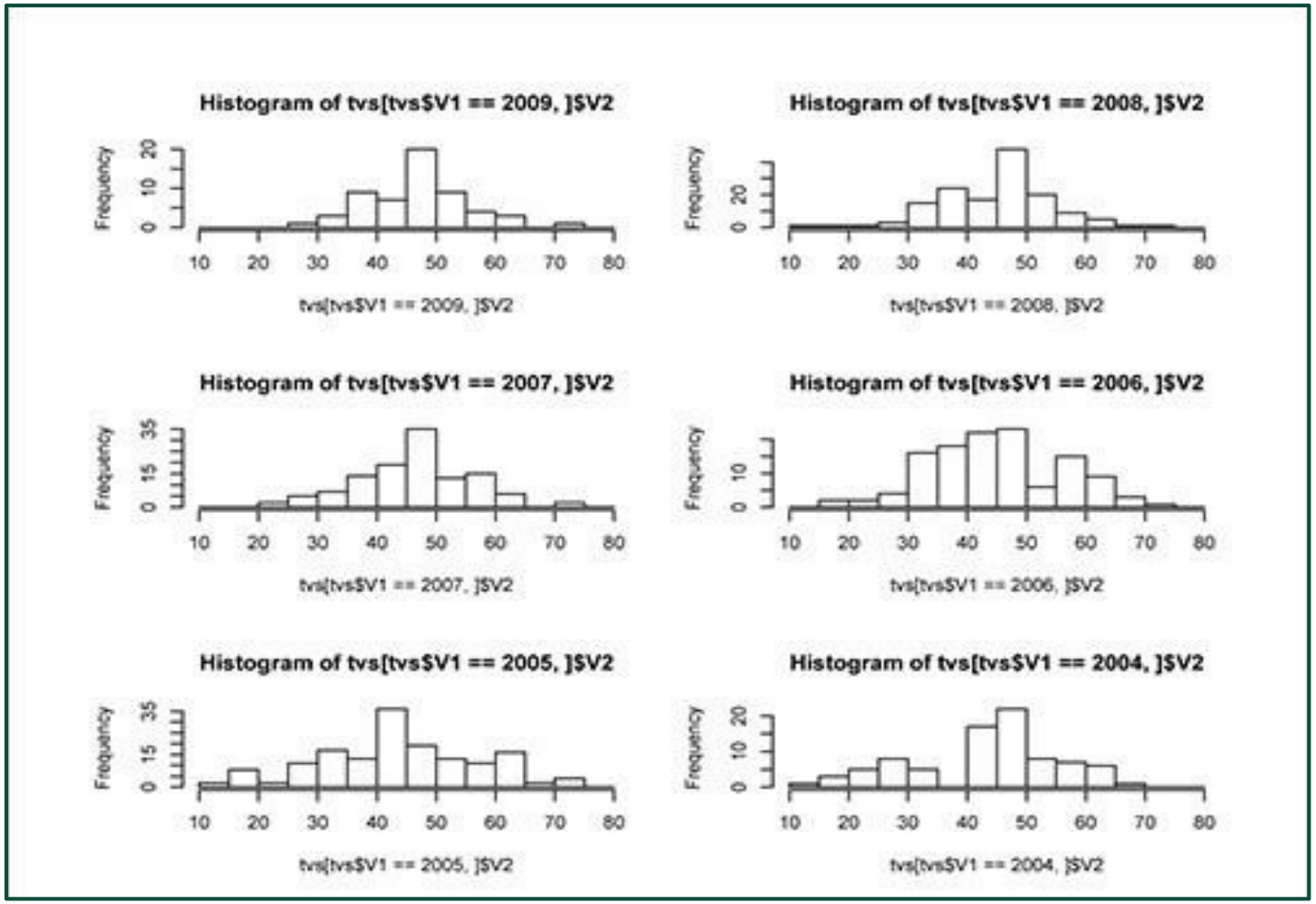

5

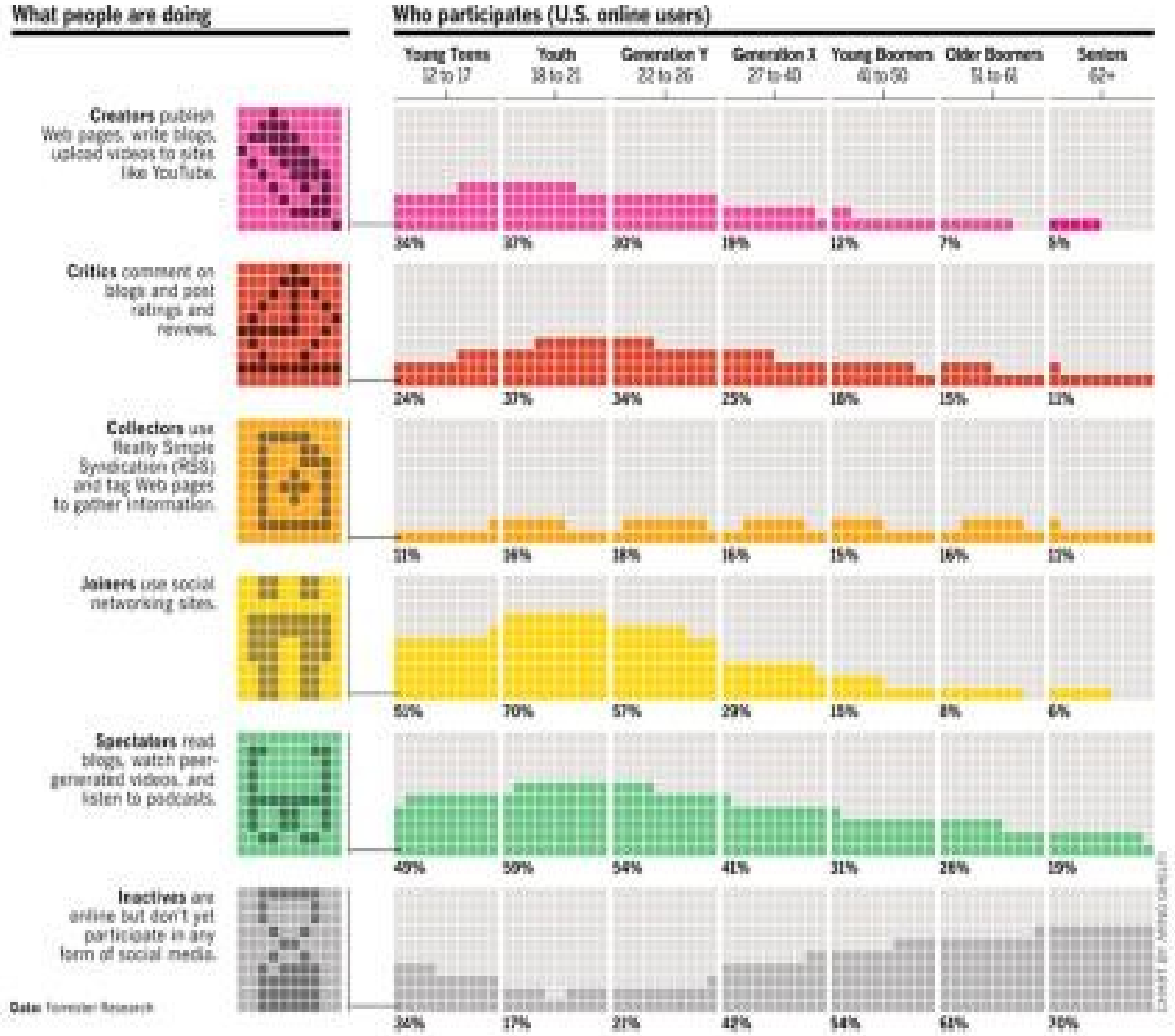
http://blogs.sas.com/content/graphicallyspeaking/2012/02/06/comparative-densities/

# Beware of Stacked Histograms



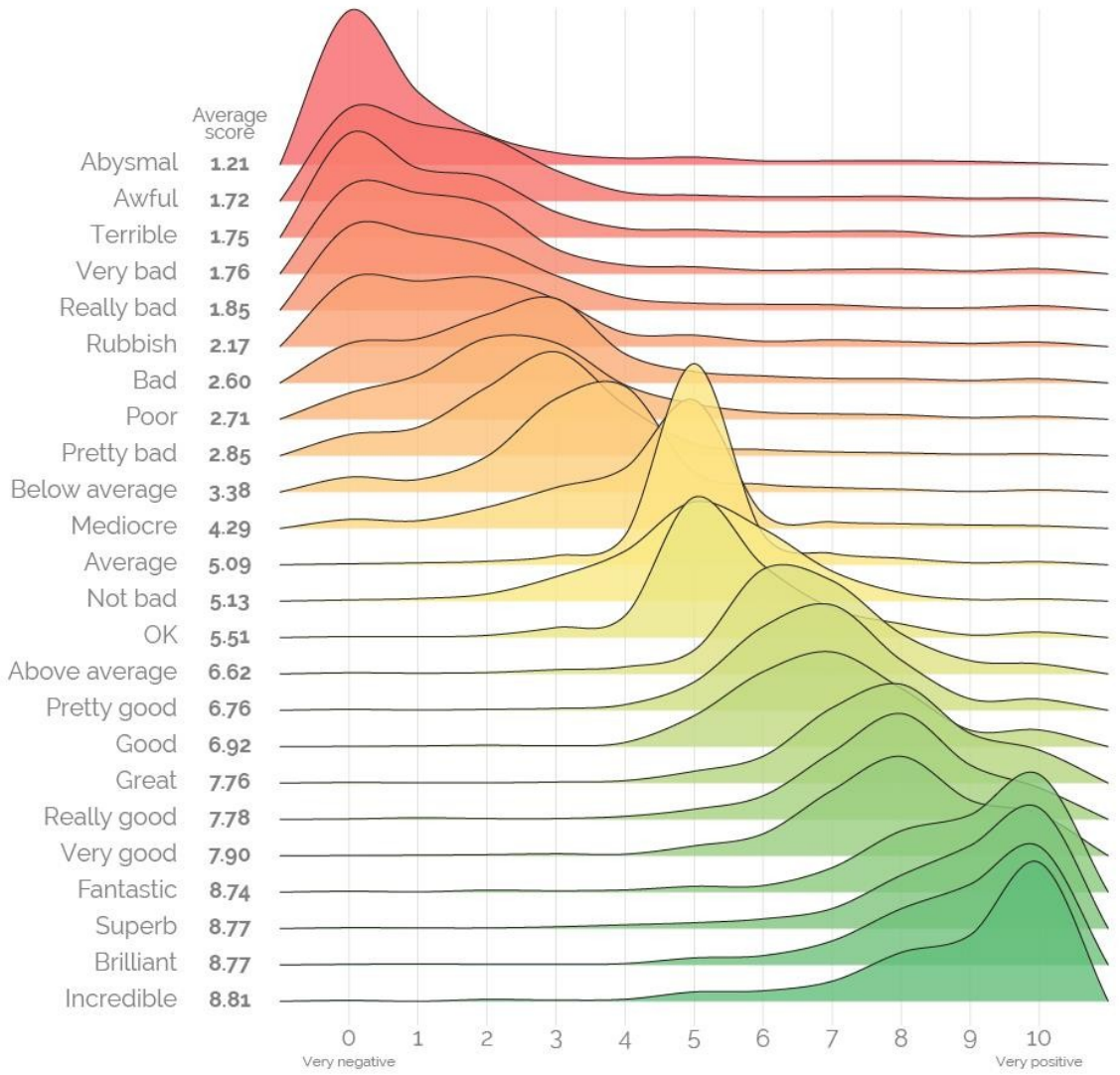
# Don’t Forget Small Multiples




What people are doing
Who participates (U.S. online users)
Young Teens 12 to 17
Youth 18 to 21
Generation Y 22 to 26
Generation X 27 to 40
Young Boomers 41 to 50
Older Boomers 51 to 61
Seniors 62+
Creators publish Web pages, write blogs, upload videos to sites like YouTube.
34% 37% 30% 19% 13% 7% 5%
Critics comment on blogs and post ratings and reviews.
24% 37% 34% 25% 18% 15% 12%
Collectors use Really Simple Syndication (RSS) and tag Web pages to gather information.
31% 36% 18% 16% 15% 10% 12%
Joiners use social networking sites.
52% 70% 57% 29% 15% 8% 6%
Spectators read blogs, watch peer-generated videos, and listen to podcasts.
49% 59% 54% 41% 31% 26% 19%
Inactives are online but don't yet participate in any form of social media.
34% 17% 21% 43% 54% 61% 70%
Data: Forrester Research

## How good is "good"?

On a scale of 0 to 10, where 0 is 'very negative' and 10 is 'very positive', general, how positive or negative would the following word/phrase be to someone when you used it to describe something?




YouGov | yougov.com

August 14-15, 2018

https://yougov.co.uk/news/2018/10/02/how-good-good/?utm_source=website_article&utm_medium=twitter&utm_campaign=word_sentiment

# Conclusions

- We often need to create visualisations to compare values
- There are a range of ways to do this
- Key things to keep in mind are:
  - Are you comparing values or proportions?
  - Are you comparing single values or distributions?
  - Are you comparing across one or many dimensions?

# Thanks To


- Marisa Llorens-Salvador, John McAuley, Colman McMahon and Brian Mac Namee for an earlier version of these lecture notes