

Visualising Proportions



Learning Outcomes

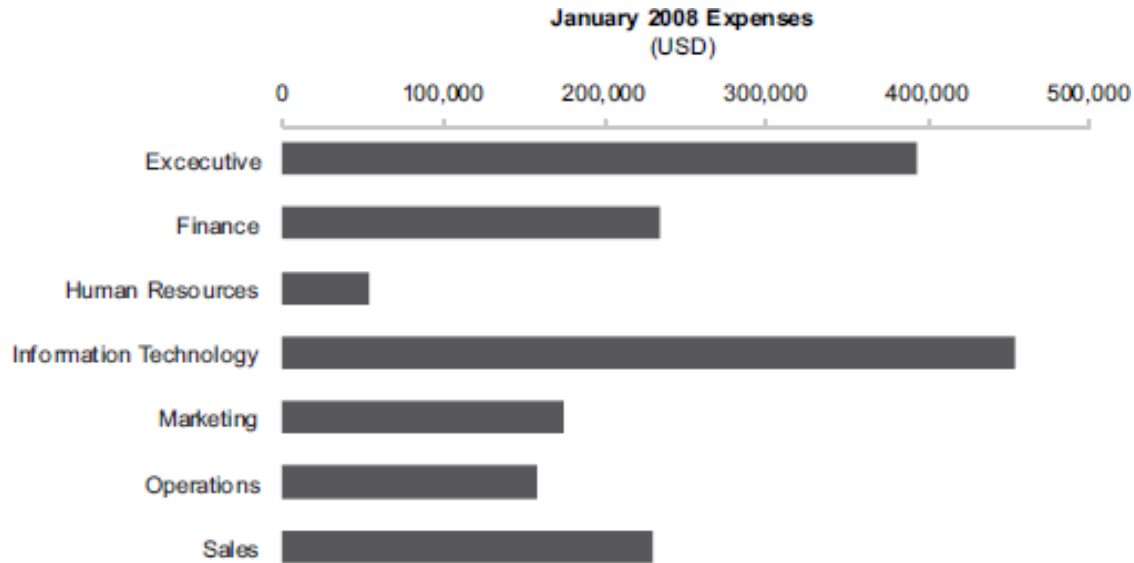
By the end of this class, you should be able to

- Identify the pattern of **part-to-whole** data
- Explain and apply the techniques and best practices used

Introduction

- Proportion data is **grouped by categories, subcategories**. *For example*, total expenses (the whole) is aggregated by expenses of departments (categories or the parts)
- For proportions, the three important things : **maximum, minimum** and the **overall** distribution.

Arranged in
alphabetical
order

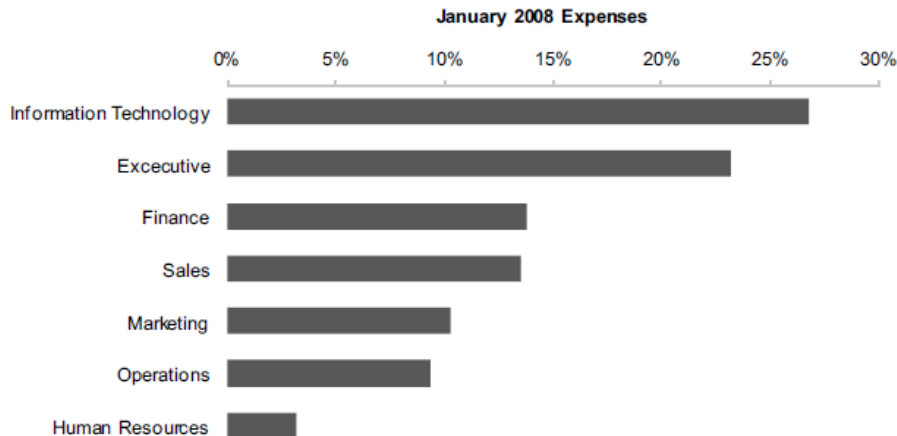


- Which department has the highest or lowest expenses?
- How much expenses has Executive occurred for the whole company?




Introduction

- Sort the expenses (ranking analysis)
- Express expenses as percentage (part-to-whole analysis)

- Which department has the highest or lowest expenses?
- How much expenses has Executive occurred for the whole company?



Proportions data Patterns

Pattern	Description	Visual Example
Uniform	All values are roughly the same.	
Uniformly different	Differences from one value to the next decrease by roughly the same amount.	
Non-uniformly different	Differences from one value to the next vary significantly.	

Proportions data Patterns

Increasingly
different

Differences from one value to the
next increase.



Decreasingly
different

Differences from one value to the
next decrease.



Alternating
differences

Differences from one value to the
next begin small and then shift to
large and finally shift back again to
small.



Exceptional

One or more values are
extraordinarily different from the
rest.



Proportions data Patterns

Which pattern is interesting?

- Unusual differences from one value to the next
- Sets of values that appear to be grouped
- Significant breaks in a pattern
- Obvious exceptions to the norm

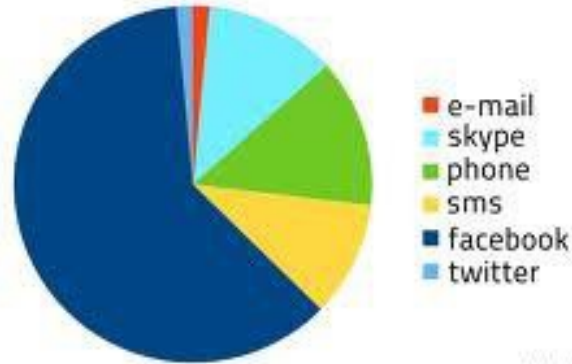
Proportions data display

- Pie chart
- Donut chart
- Stacked area chart
- Stacked bar chart
- Treemap
- Pareto chart

Pie Chart

- The circle represents the whole, and the size of wedge represents the part or percentage of that whole. *Together, those represented values, add up to 100%.*
- Use this only if you're comparing a few values (like three or less)

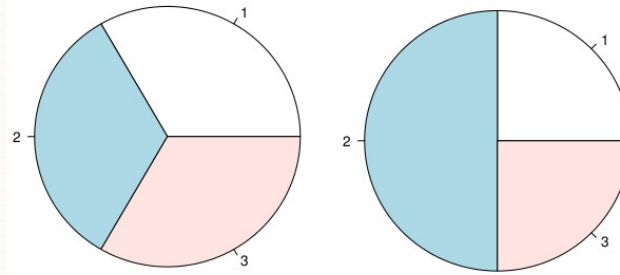
Birthday wishes by channel



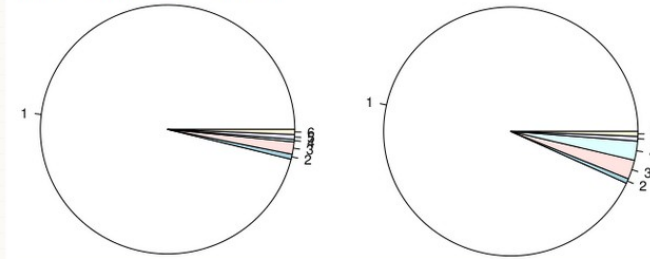
©blog.fia.cx

Pie Chart

My personal problem with pie charts is while they may be useful to show differences like this:



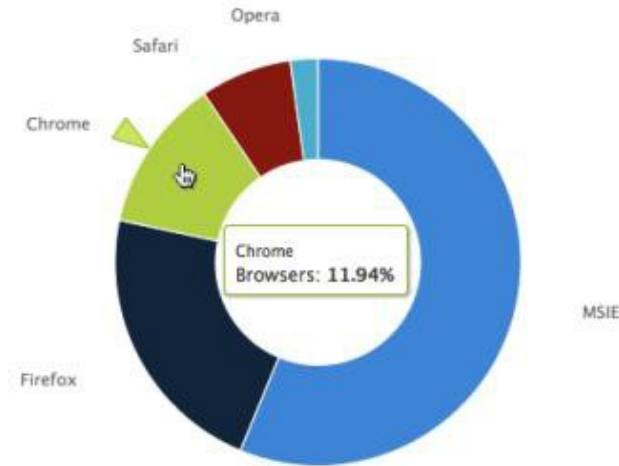
way too many people use it to show that:



<http://stats.stackexchange.com/questions/8974/problems-with-pie-charts>

Donut Chart

- Pie's lesser-used cousin.
- Same idea as the pie, but with a hole cut out in the middle. The same arguments of angles and human perception still apply

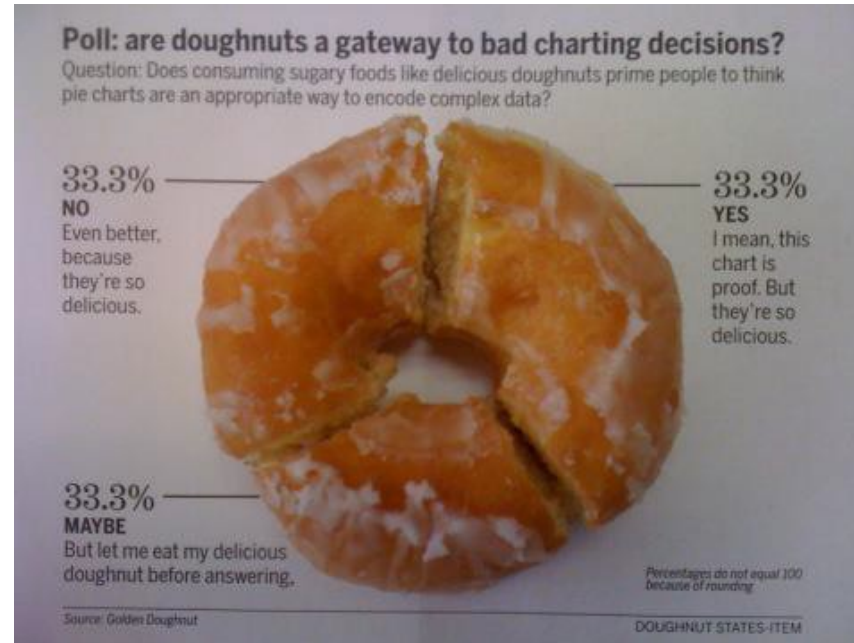


<http://joekuan.wordpress.com/2013/06/09/highcharts-enhancing-user-interaction-on-piedonut-charts-dynamic-connector/>

When to Use Pie Charts

- Do the parts make up a meaningful whole?
- Are the parts mutually exclusive?
- Do you want to compare the parts to each other or the parts to the whole?
- How many parts do you have?

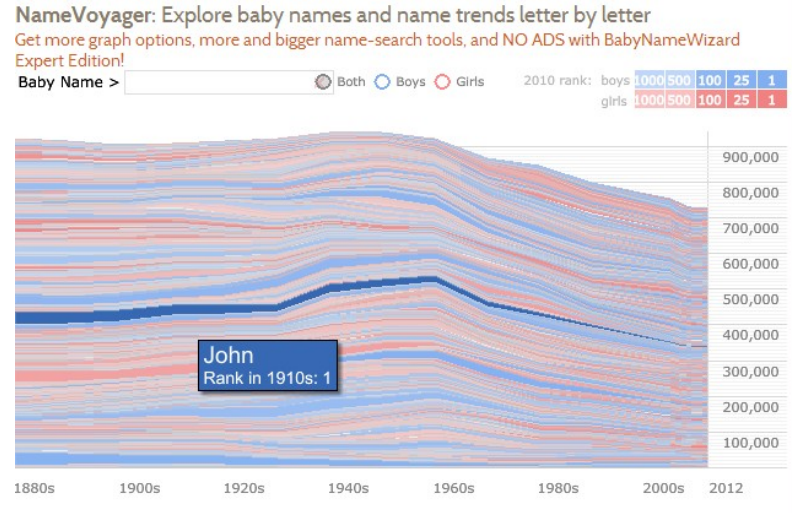
Donut Chart



<http://missom.wordpress.com/>

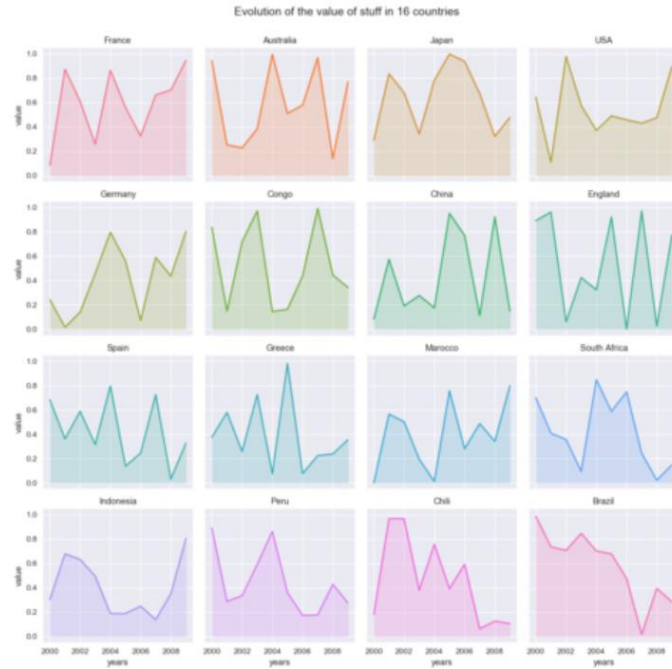
Stacked Area Chart

- Use the stacked area chart if you want to show **changes over time** for several variables.
- Percentages -> vertical always adds up to 100%
- raw counts → peaks and valleys.



<http://www.babynamewizard.com/voyager#ms=false&exact=false>

Facetting

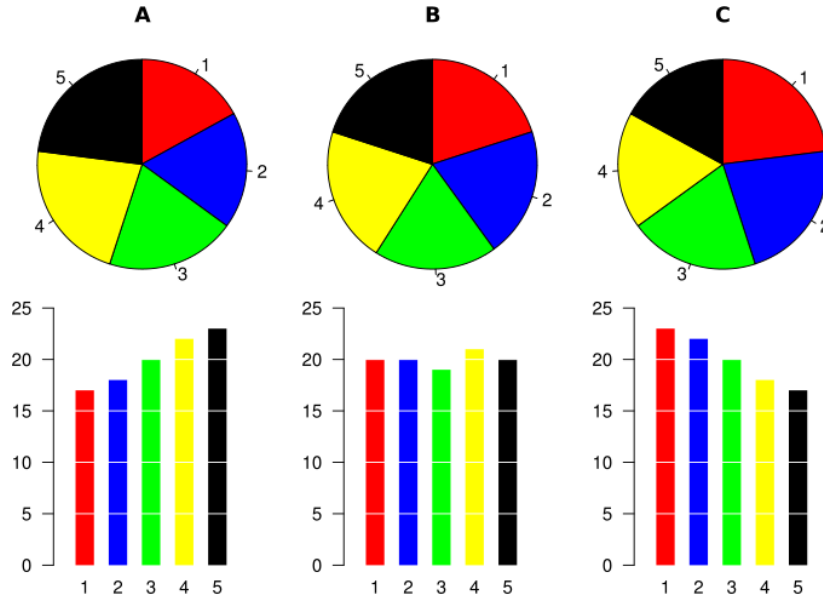


Stacked Bar Chart

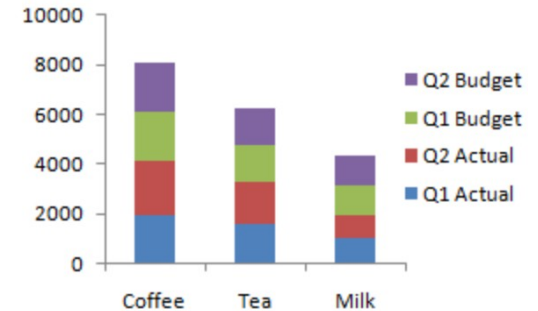
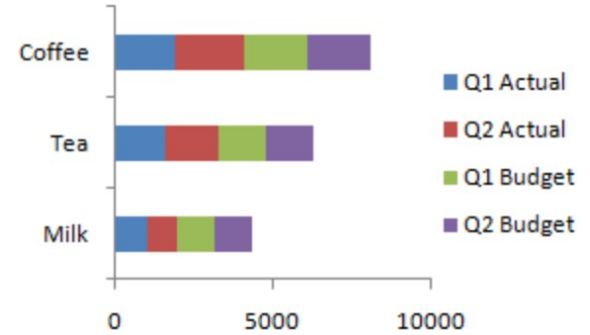
- If you have only a few distinct points in time, you can use the stacked bar chart in the same way you use the stacked area (just set the bars vertical).
- This is better than pie chart as it's sans angle perception problem.



<https://eagereyes.org/techniques/stacked-bars-are-the-worst>



<https://peltiertech.com/clustered-stacked-column-bar-charts/>



Treemap

- Treemap uses the areas of rectangles to show **relative** proportions. It works especially well if your data has a **hierarchical structure** with parent nodes, children, etc.
- Display quantities for each category via area size
- Colour/gradient can be a 2nd measure

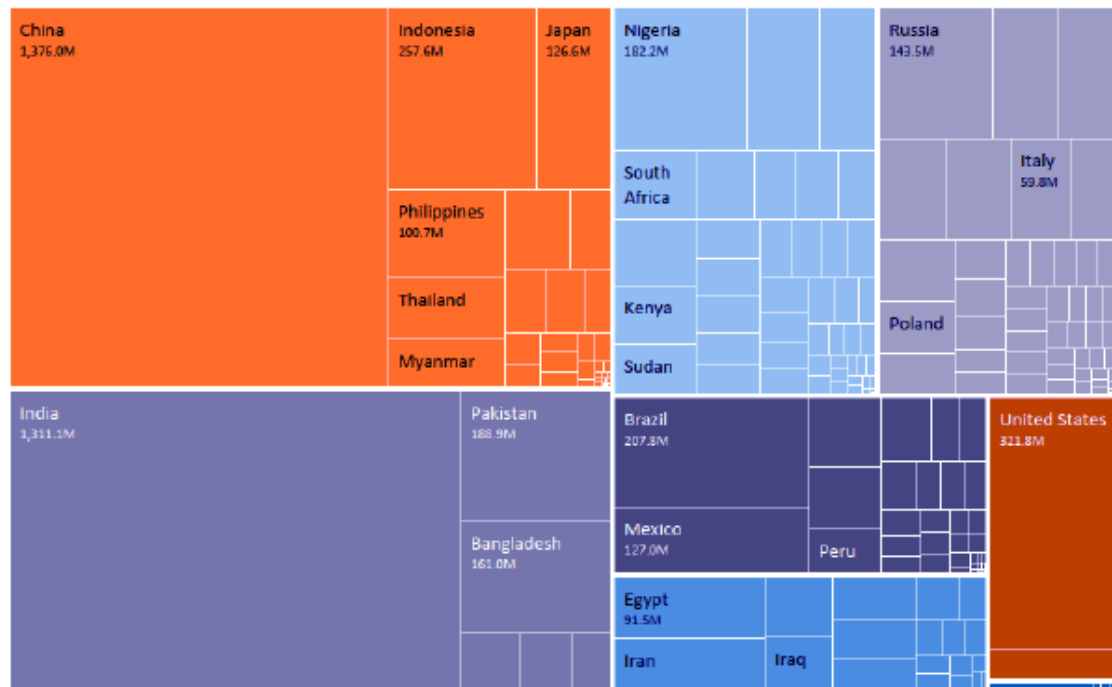


Treemap of Benin's exports by product category

Treemap of Benin's exports by product category, 2009. The Product Exports Developed by the Harvard-
[MIT Observatory of Economic Complexity](https://atlas.media.mit.edu/en/visualize/tree_map/hs92/export/ben/all/show/2009/)

https://atlas.media.mit.edu/en/visualize/tree_map/hs92/export/ben/all/show/2009/

World Population in 2015

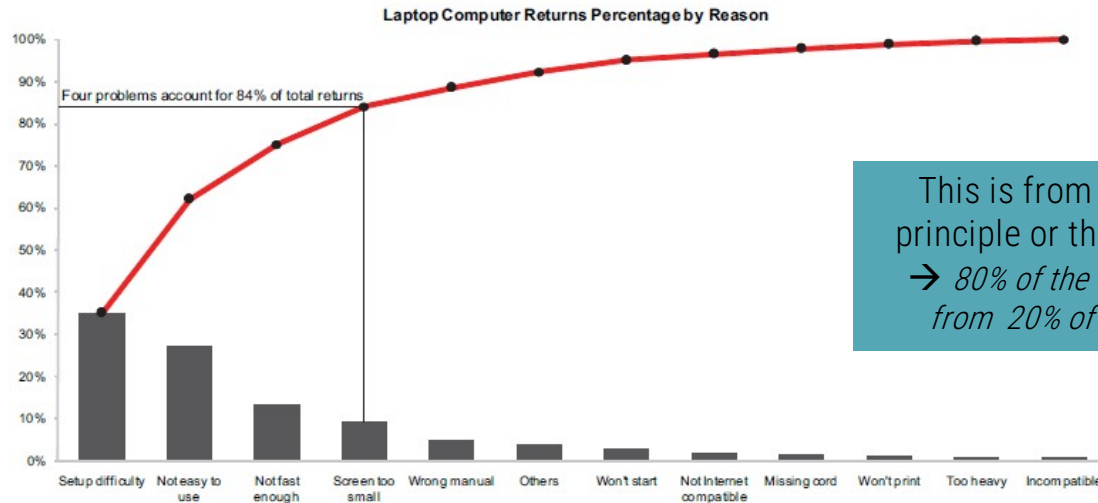


Region

- Unknown
- East Asia & Pacific
- Europe & Central Asia
- Latin America & Caribbean
- Middle East & North Africa
- North America
- South Asia
- Sub-Saharan Africa

Pareto chart

- Besides ranking values, important to examine the cumulative contribution of parts to the whole
- Can be constructed using bar chart and line chart together



This is from the Pareto principle or the 80/20 rule
→ 80% of the efforts come from 20% of the causes

Caution on Pareto chart

- A wildly **fluctuating system** will produce inconsistent Pareto rankings that can lead to misjudgement.

For example, if the retail manager failed to note that customer furniture returns varied greatly from month to month, the ranking of categories may be entirely different in a month with high returns from those of a month in which returns were unusually low.

- Repeated Pareto analyses can help to confirm rankings, but the **most effective protection** against being misled is to first use a **control chart** to tell if the system is stable and predictable.

http://www.pqsystems.com/qualityadvisor/DataAnalysisTools/pareto_diagram.php

When to use Pareto Chart?

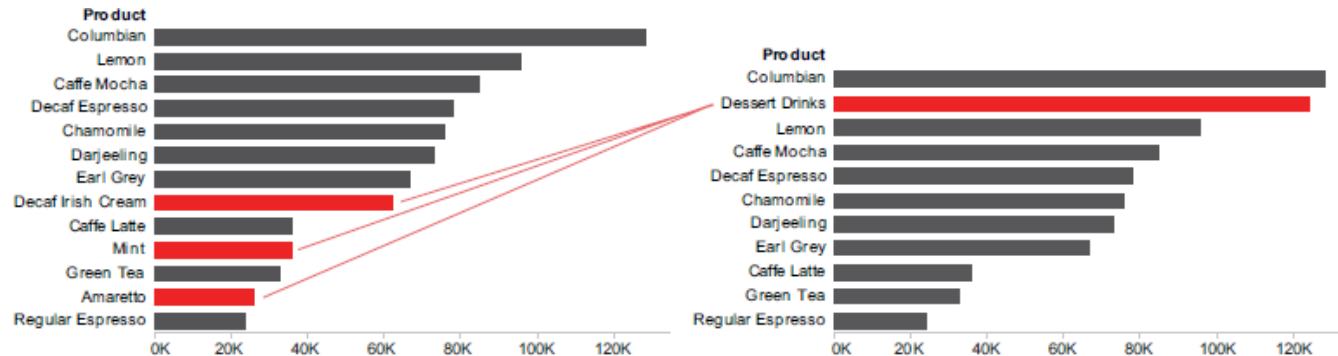
- Can data be arranged into categories?
- Is the rank of each category important?
- When analysing data about frequency of problems or causes in a process

Techniques and Best Practices

- Grouping categorical items in an ad hoc manner
- Using Pareto charts with percentile scales
- Re-expressing values to solve quantitative scaling problems
- Using line graphs to view ranking changes through time

Grouping

- Segment data into meaningful grouping for analysis.
- Ability to group categorical items in an ad-hoc manner is critical to uncover interesting insights.

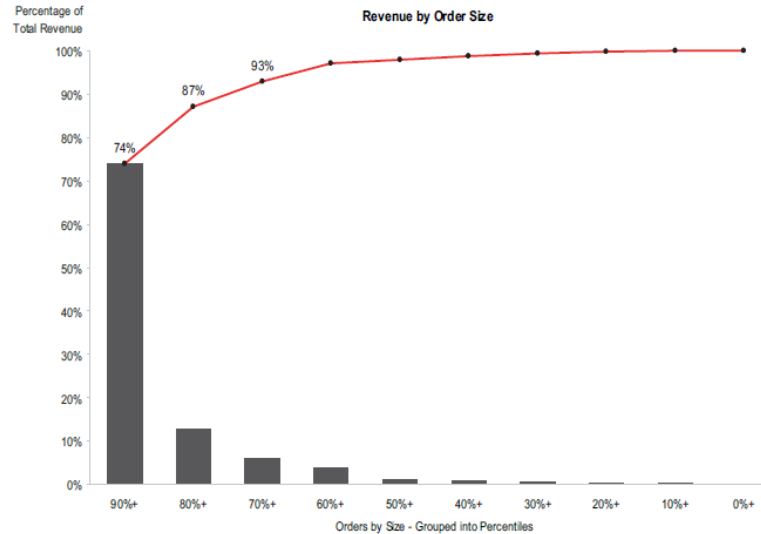


Percentile Scales

Used in interval scale (size of orders from largest to smallest, grouped into percentile intervals)

What does this graph tell you?

70% of their orders (everything to the right of the 70%+ bar) accounted for only 7% of their revenue even though these orders ate up a majority of their sales efforts

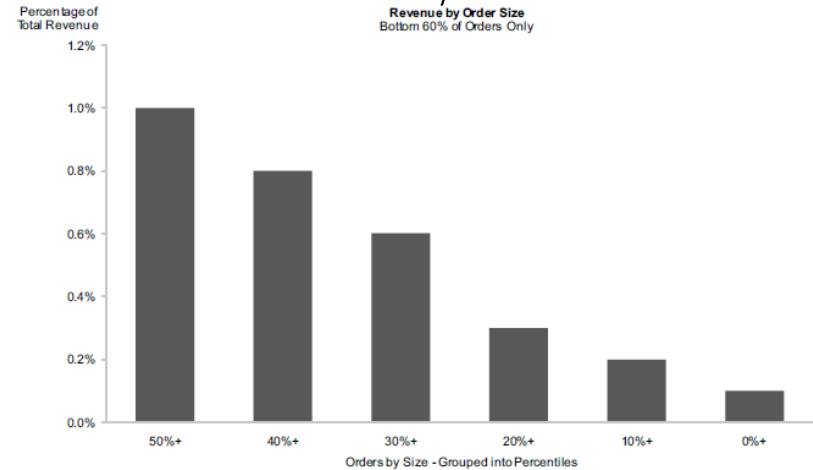
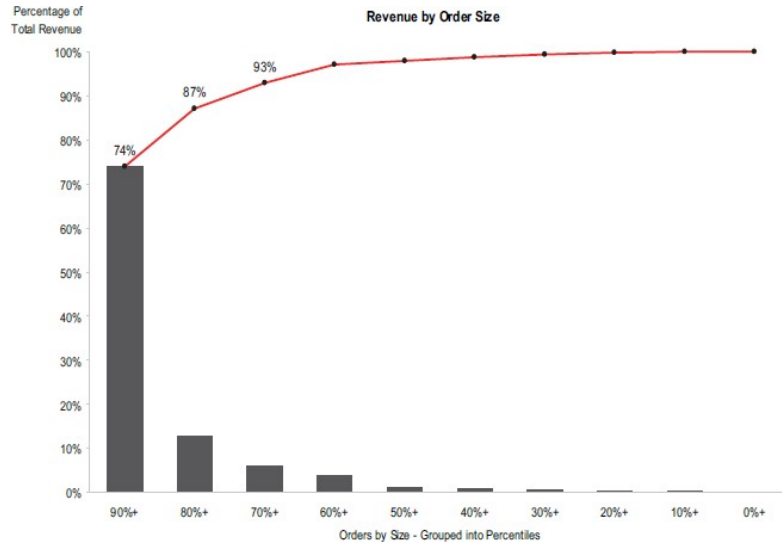


Example

- Percentage – absolute; an exact quantitative rating
$$(\text{Your Marks} / \text{Total Marks}) * 100$$
- Percentile – comparative quantitative measurement. The 99th percentile is the top 1%, 98th percentile is the top 2%
$$(\text{no. of ppl behind you} / \text{total num of ppl}) * 100$$
- Example: A student has scored 90 percentile in a given exam. It means 90 percent of test takers are behind him or acquired lesser marks and rank than him.

Re-expressing Values

What are the value below 60 percentile?



One way: Viewing of the low values independently in a separate graph.

When a set of ranked values extends a vast scale, it could be difficult to see and compare with the lowest values at the end graph.

Re-expressing Values

- How to solve this scaling problem so that values are evenly distributed across the quantitative scale in a single graph?

Re-expression

■ BUT

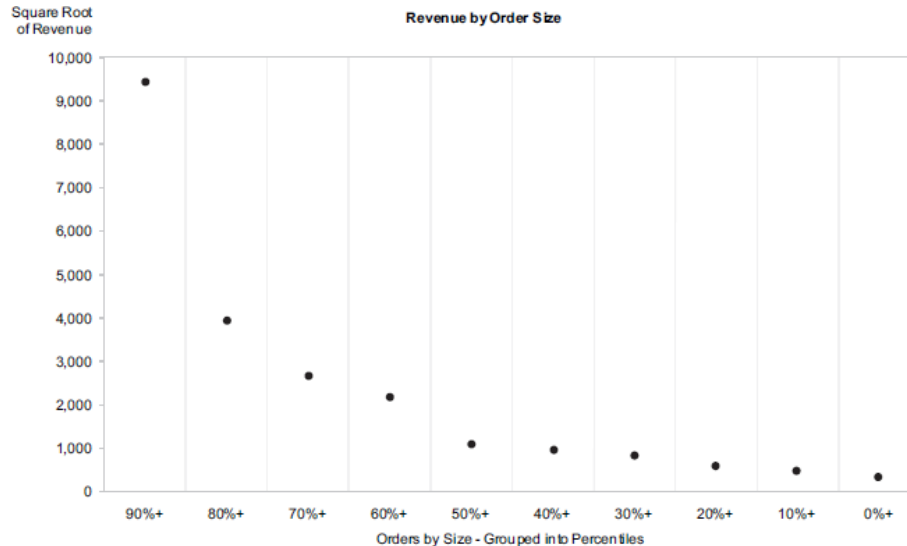
- It alters distances between values and distorts the actual magnitudes of the difference so use it with care!

Re-expressing Values

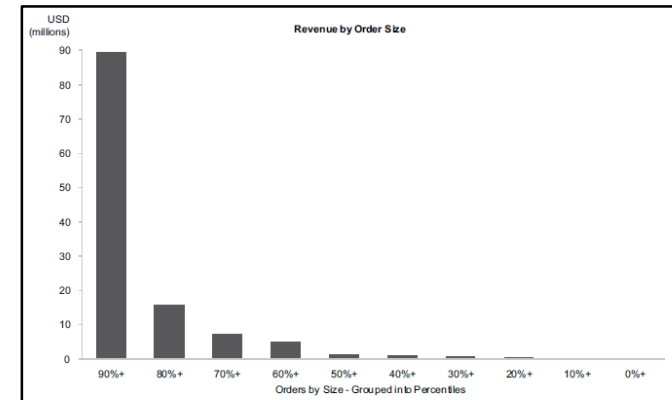
- Re-expression is good in solving scaling problem where low values are hard to read.
- Two methods
 1. Square root re-expression
 2. Logarithmic re-expression
- The above is sequenced by **amount of stretching** and **compression**, from least to greatest.
- Stretching low values out across $>$ space in the graph & compressing high values into $<$ space
- Try from the least first to see if problem is resolved.

Re-expressing Values

Square root of the revenue instead



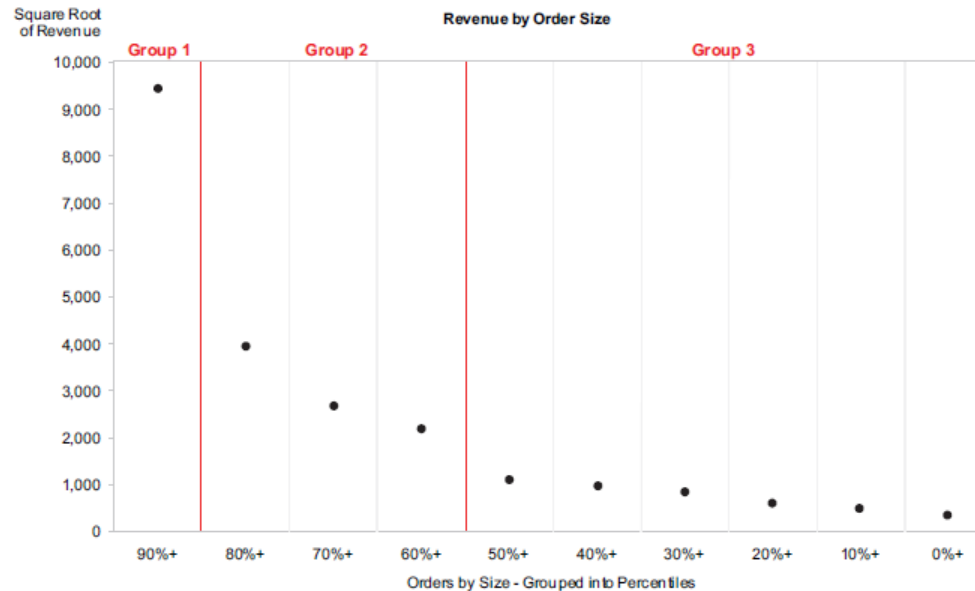
Dot plot is used to discourage ourselves from comparing the heights of the bars as a means of comparing these square root values



Why use dot plot now?

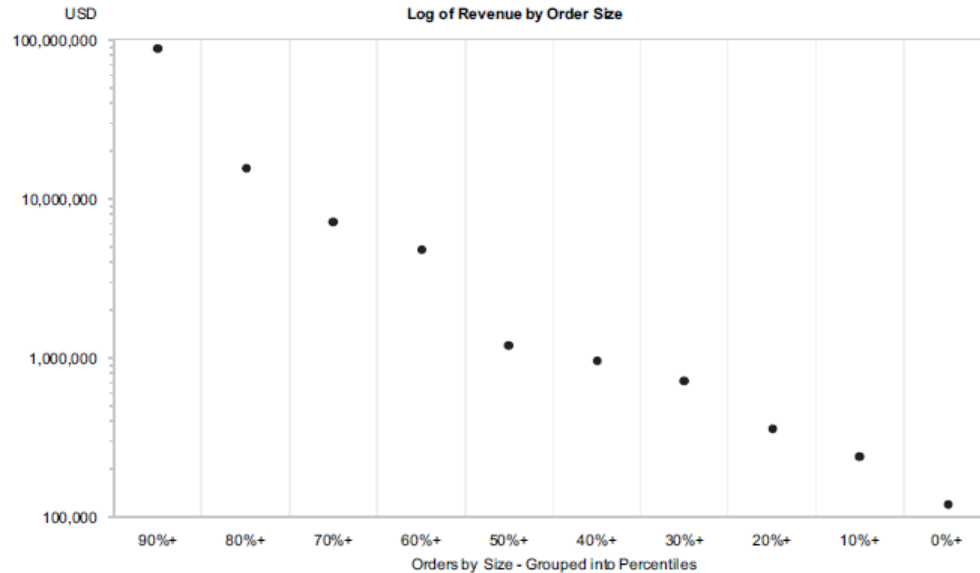
Re-expressing Values

Discover any grouping?



Re-expressing Values

Logarithmic re-expression

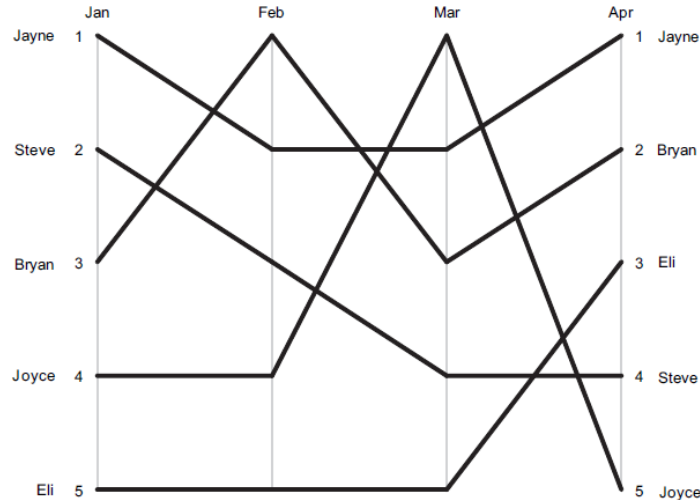


Ranking changes

We should never use line chart to display single ranking relationship but to visualise the changes over time is fine.

Use the combination of a ranking and a time-series relationship

Only show changes in rankings and not values (like the sales amount)



How should we present the following data?

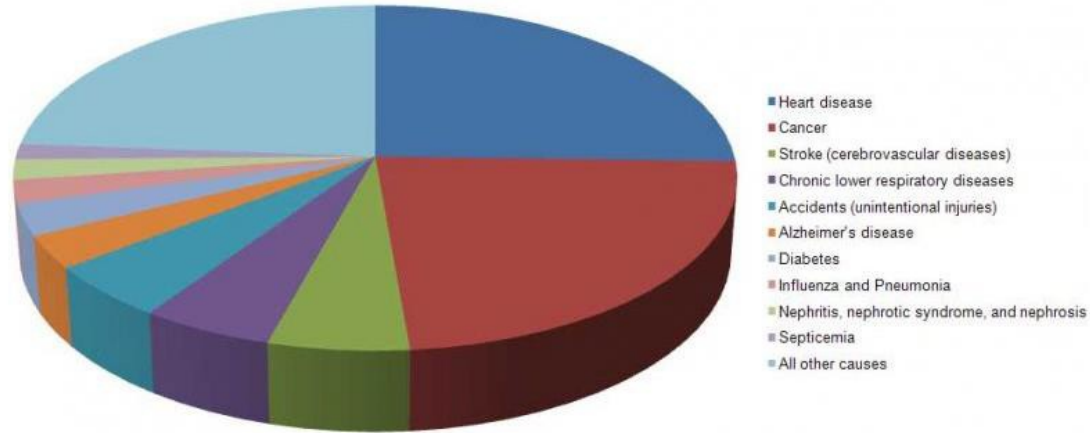
Causes of Death	Deaths per Year
Heart disease	616,067
Cancer	562,875
Stroke (cerebrovascular diseases)	135,952
Chronic lower respiratory diseases	127,924
Accidents (unintentional injuries)	123,706
Alzheimer's disease	74,632
Diabetes	71,382
Influenza and Pneumonia	52,717
Nephritis, nephrotic syndrome, and nephrosis	46,448
Septicemia	34,828
All other causes	577,181
Total	2,423,712

The display should achieve the following:

- Clearly indicates how the values relate to one another, which in this case is a **part-to-whole relationship** - the number of deaths per cause, when summed, equal all deaths during theyear.
- Represents the quantities accurately.
- Makes it easy to **compare** the quantities.
- Makes it easy to see the **ranked order of values**, such as from the leading cause of death to the least.
- Makes it obvious on how people should use the information - what they should use it to accomplish - and encourages them to do this.

http://www.interaction-design.org/encyclopedia/data_visualization_for_human_perception.html

Total Deaths in American by Cause in 2007



Copyright status: Unknown (pending investigation). See section "Exceptions" in the copyright terms below.

Clearly indicates the nature of the relationship?

Yes.

Represents the quantities accurately?

No.

Makes it easy to compare the quantities? Makes it easy to see the ranked order of values?

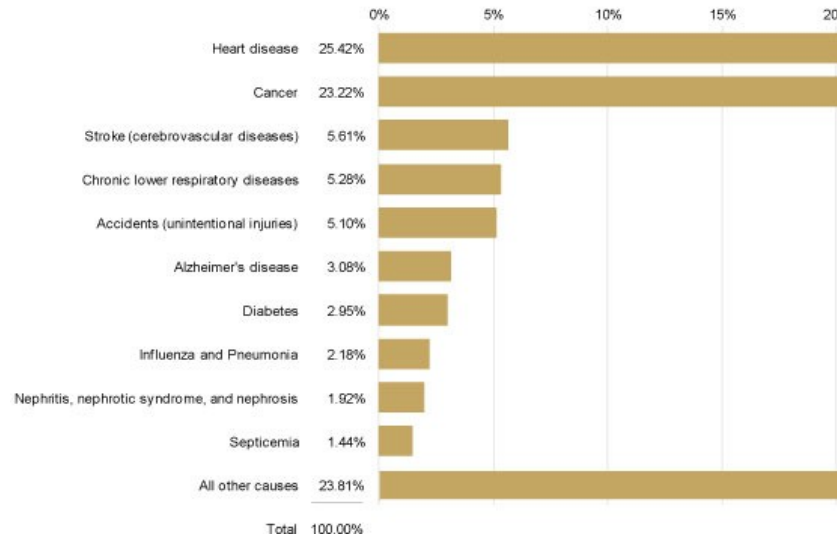
No.

No.

Makes obvious how people should use the information?

Partially.

Total Deaths in America by Cause in 2007



Clearly indicates the nature of the relationship?

Represents the quantities accurately?

Makes it easy to compare the quantities? Makes it
easy to see the ranked order of values?

Makes obvious how people should use the information?

Yes.

Yes.

Yes.


Yes.

Yes.

Summary

- Proportion or Part to Whole pattern
- How to display proportion data
- Techniques and best practices to consider for proportion data

Reference

 <http://flowingdata.com/2009/11/25/9-ways-to-visualize-proportions-a-guide/>

 Now You See It : Simple Visualisation Techniques for Quantitative Analysis/ Stephen Few, Analytics Press, c2009

P.189 – P.202 Chapter 8 Part-to-Whole and Ranking Analysis

▶ All the images are extracted from the book unless it is explicitly stated

 Visualize This / Nathan Yau, Wiley, c2011

<http://www.npr.org/blogs/deceptivecadence/2013/05/27/186461168/watch-a-mind-blowing-visualisation-of-the-rite-of-spring>