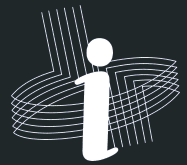
Data Visualization

Introdução à aprendizagem de máquinas INF 792
Prof.: Sabrina de A. Silveira
27 de abril de 2021



What is Data Visualization?

- A relatively new approach to analyzing data that is still under development.
- Two factors were decisive in making Data Visualization possible:
  - Computers (graphics)
  - Availability of big data available
- We can analyze printed data, but how do we interact with that data to answer questions?

The process of data analysis

- Collecting data from various sources, integrating, transforming, filtering and storing it in a database is data analysis?
- Although important, this type of pre-processing, in which we prepare the data to effectively perform the analysis, is not part of the data analysis process.
- In this course, the focus is on data exploration and analysis, in order to understand the semantics present in them.

Why visualize?

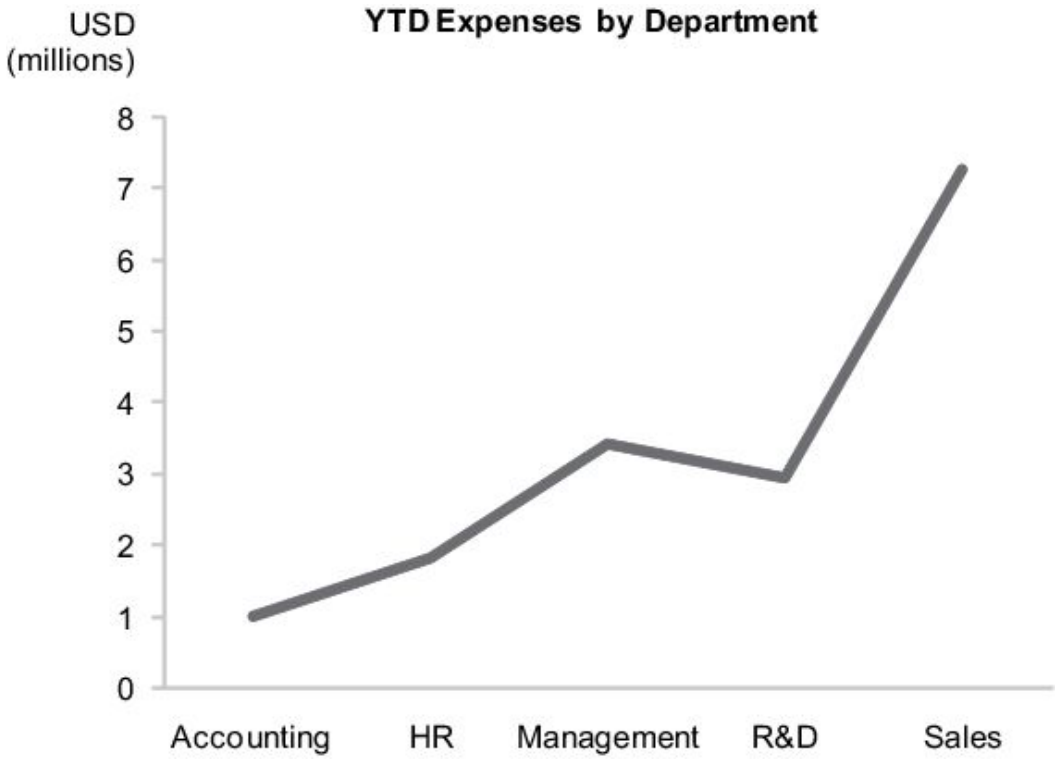
A good sketch is better than a long speech.

https://en.wikiquote.org/wiki/Napoleon_I_of_France

Building core skills for visual analysis

Making abstract data visible

- Does this graph look okay to you?

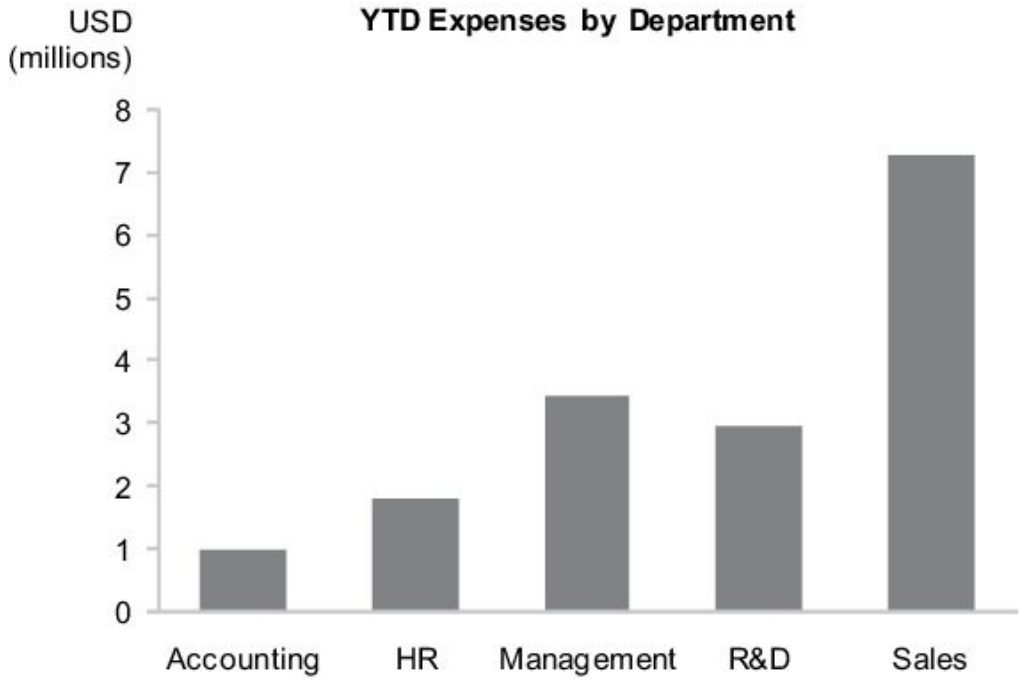


Making abstract data visible

- The line connects values for a series of categorical items that are independent from one another.
- Lines work well for connecting values through time.

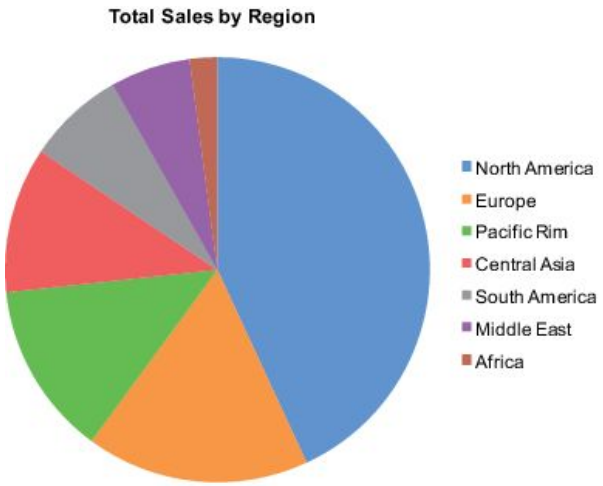
Making abstract data visible

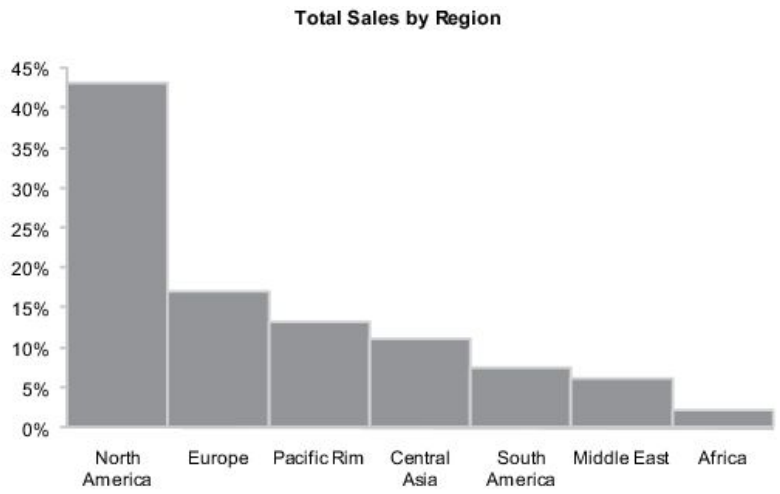
Does this graph look okay to you?



Comparing...

If we wish to rank and compare the sales performance that are displayed in each of the two graphs below, which supports this task more effectively?





Comparing...

- The pie chart doesn't work nearly as well as the bar graph because, to decode it, we must compare the 2-D areas or the angles formed by the slices.
- Visual perception doesn't accurately support either of these tasks (comparing areas or angles).

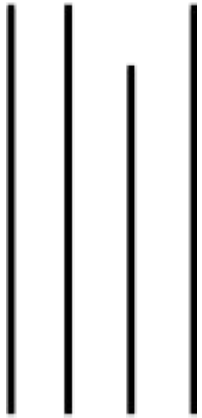
Attributes of visual perception

- Jacques Bertin, in his book *Semiologie graphique*, was the first person to recognize and describe the basic vocabulary of vision, that is, the attributes of visual perception that we can use to display data in a clear accurate, efficient, and intuitive manner.
- We can use the **pre-attentive attributes** of visual perception. The theoretical mechanism underlying pop-out occurs prior to conscious attention.

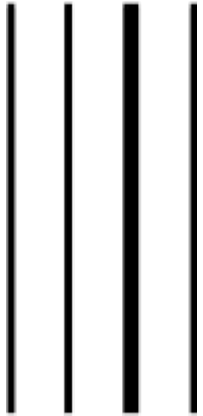
Pre-attentive attributes

Form

Length



Width

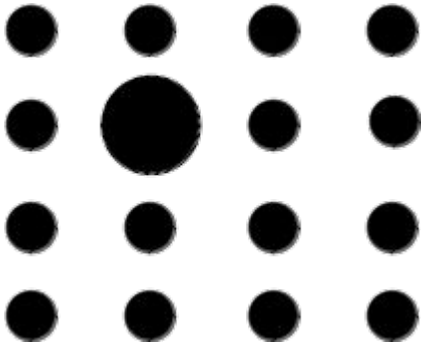


Orientation

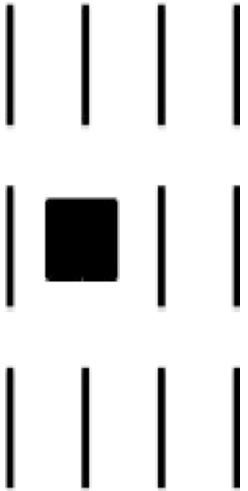
Pre-attentive attributes

Form

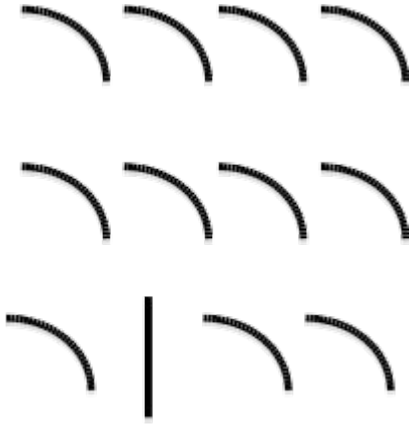
Size



Shape



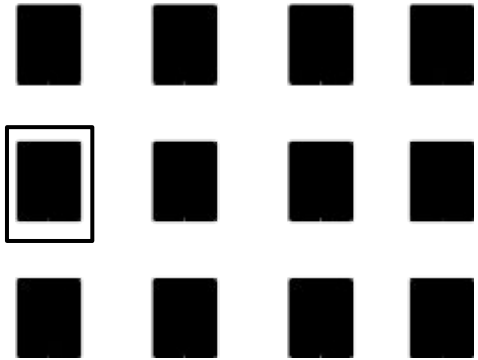
Curvature



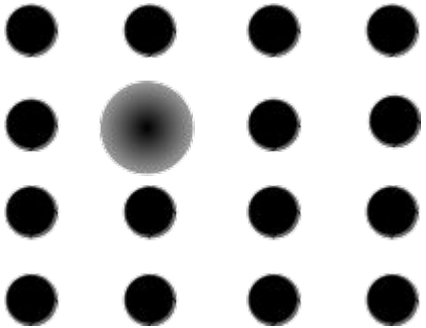
Pre-attentive attributes

Form

Enclosure



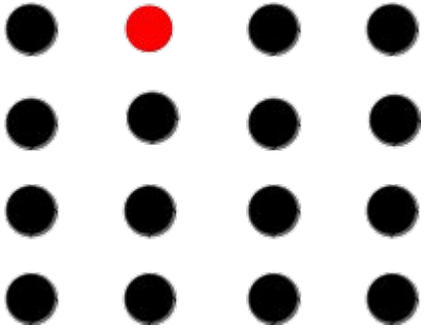
Blur



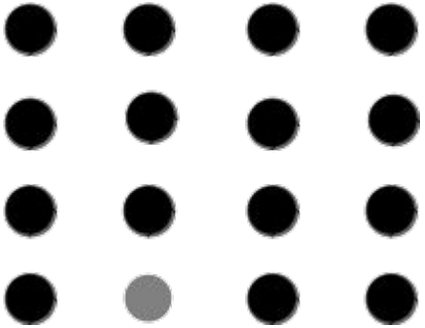
Pre-attentive attributes

Color

Hue



Intensity

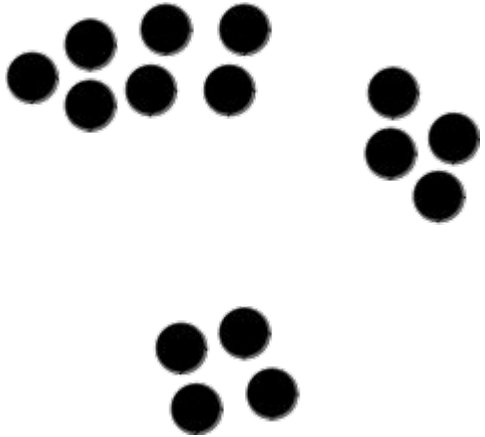


Pre-attentive attributes

Spatial position

2D position

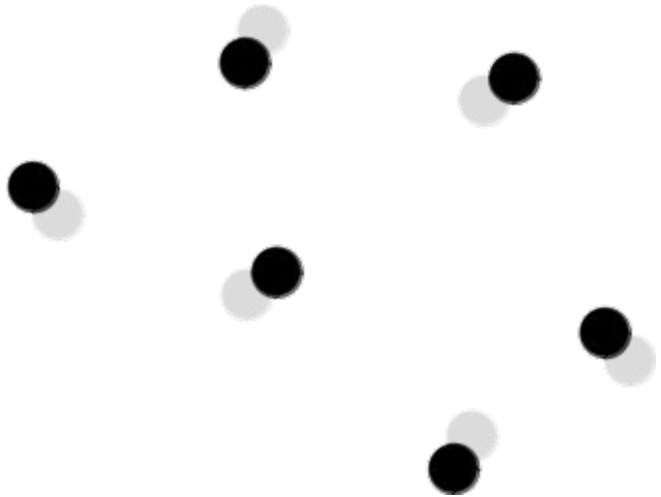
Spatial grouping



Pre-attentive attributes

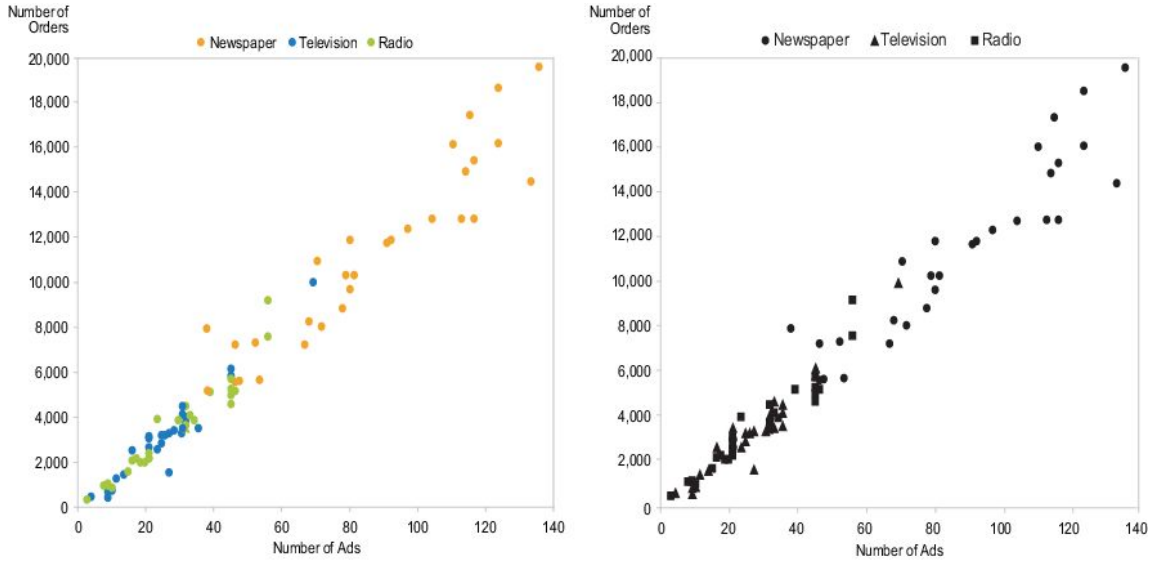
Motion

Direction



How to make data visible

- Some of the pre-attentive attributes are useful for making objects look distinct from one another, which enables us to assign subsets of visual objects to categorical groups



- The best attributes for this are hue and shape.

How much does a figure cost?

- *Colour figures:* Inclusion of colour figures is subject to a special charge (£350/US$600/€525 to print a figure in colour). Online colour figures are free of charge. Orders from the UK will be subject to the current UK VAT charge. For orders from elsewhere in the EU you or your institution should account for VAT by way of a reverse charge. Please provide us with your or your institution's VAT number.

https://academic.oup.com/bioinformatics/pages/instructions_for_authors

How to make data visible

- Hues are usually easier to interpret than shapes.
- What about color blindness?

Color perception

- Red and green are colors that are difficult to distinguish for color blind people.
- 10% of men and 1% of women cannot distinguish between green and red;

globo.com g1 ge gshow globoplay tech E-MAIL

Ministro da Saúde admite dificuldade para fornecer 2ª dose

- 'O pior ainda está por vir', avalia ex-secretário do Ministério da Saúde
- Demitido por Bolsonaro, ex-AGU Levi vira alvo de convocação do PT
- Bolsonaro faz novo discurso contra restrições e 'pseudogovernadores'

Hamilton Mourão toma 2ª dose da vacina em Brasília

- UE processa AstraZeneca por atrasos da vacina

Pazuello é visto sem máscara em shopping no AM

- Sem máscara, Premier da Tailândia é multado

Monique muda versão; compare com depoimento

- Violência atinge mais crianças de até 6 anos

Projeto volta ao Congresso e irrita

Grupo Soma compra a Hering

Estrangeiros tiram US$ 2,1 bi

Votação elege Taffarel nº 1 no Dia do goleiro com ampla vantagem

Restrição pode afetar data e hora de Santa Fe x Flu pela Liberta

- Tabela: Palmeiras, Inter, Santos, Fla e Galo jogam amanhã

Técnico da Seleção

A 8 dias da final, Ju chora por não ganhar nenhuma prova

- Fluk: 'Não sei explicar quando fui tão feliz'; veja raio-x
- Gil encena volta do Paredão Falso: 'O que vocês merecem'

'Surpresa enorme', diz mãe de Gil após nudez de filho e Fiuk

- Assinante Globoplay vê ao vivo
- Dropz: memes, paródias e mais

globo.com g1 ge gshow globoplay tech E-MAIL

Ministro da Saúde admite dificuldade para fornecer 2ª dose

'O pior ainda está por vir', avalia ex-secretário do Ministério da Saúde
Demitido por Bolsonaro, ex-AGU Levi vira alvo de convocação do PT
Bolsonaro faz novo discurso contra restrições e 'pseudogovernadores'

Hamilton Mourão toma 2ª dose da vacina em Brasília
UE processa AstraZeneca por atrasos da vacina

Pazuello é visto sem máscara em shopping no AM
Sem máscara, Premier da Tailândia é multado

Monique muda versão; compare com depoimento
Violência atinge mais crianças de até 6 anos

Projeto volta ao Congresso e irrita

Grupo Soma compra a Hering

Estrangeiros tiram US$ 2,1 bi

Votação elege Taffarel nº 1 no Dia do goleiro com ampla vantagem

Restrição pode afetar data e hora de Santa Fe x Flu pela Liberta
Tabela: Palmeiras, Inter, Santos, Fla e Galo jogam amanhã

Técnico da Seleção

A 8 dias da final, Ju chora por não ganhar nenhuma prova
Fluk: 'Não sei explicar quando fui tão feliz'; veja raio-x
Gil encena volta do Paredão Falso: 'O que vocês merecem'

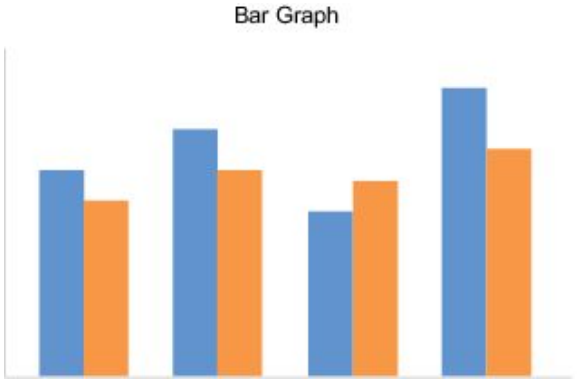
'Surpresa enorme', diz mãe de Gil após nudez de filho e Fiuk
Assinante Globoplay vê ao vivo
Dropz: memes, paródias e mais

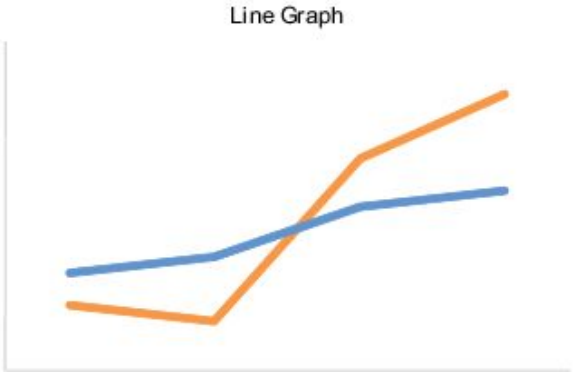
Ready

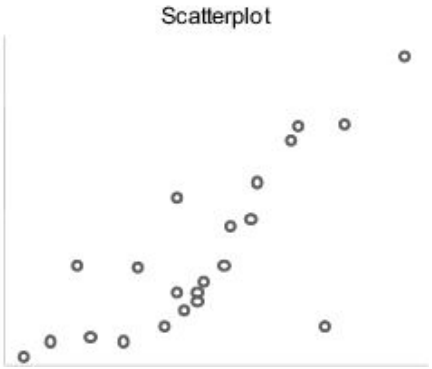
Copy filtered page URL

How to make data visible

- Only two pre-attentive attributes are perceived quantitatively with a high degree of precision: length and 2D position.

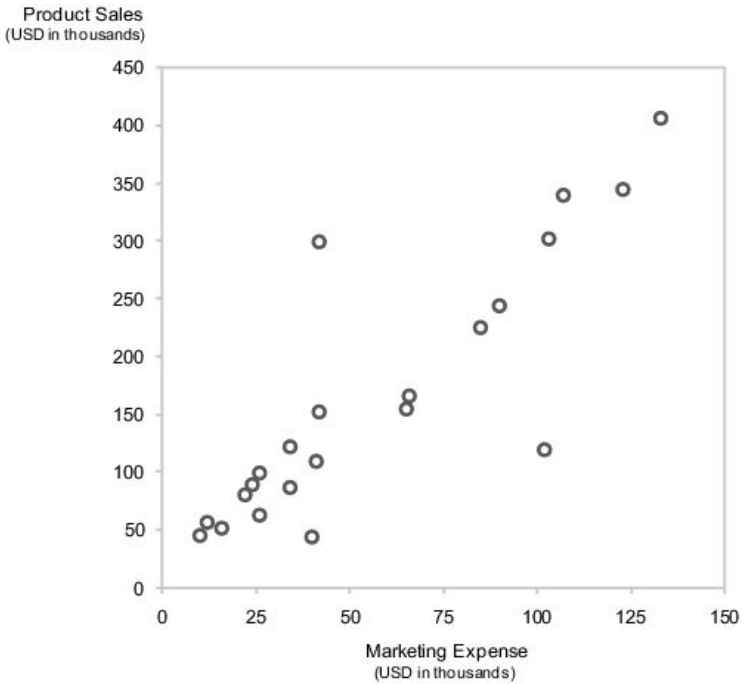






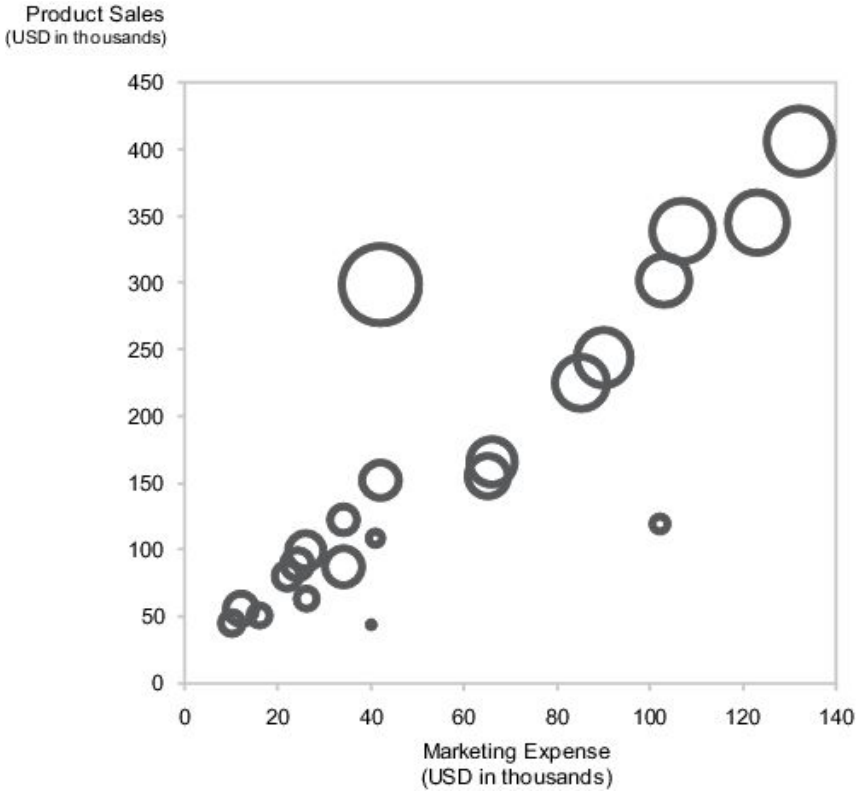
How to make data visible

We need to see the relationship of profits to both sales revenues and marketing expenses. How can we do this?



How to make data visible

We can use the size of the point to encode profit.

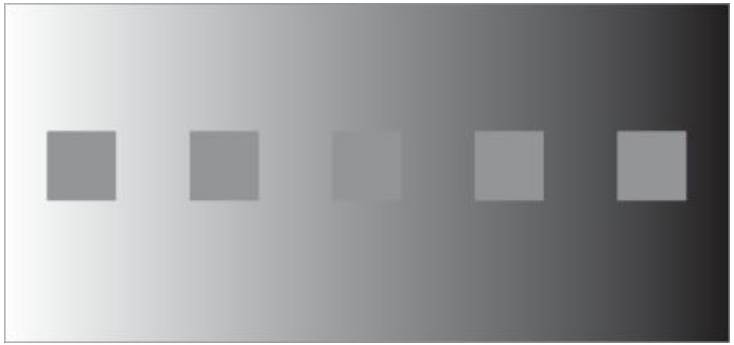


How visual perception works

Visual perception does not measure absolute values, but differences between them.

How visual perception works

What color are these squares?



How visual perception works

We perceive the differences as ratios (percentages) and not as absolute values:

Ratio 2/1

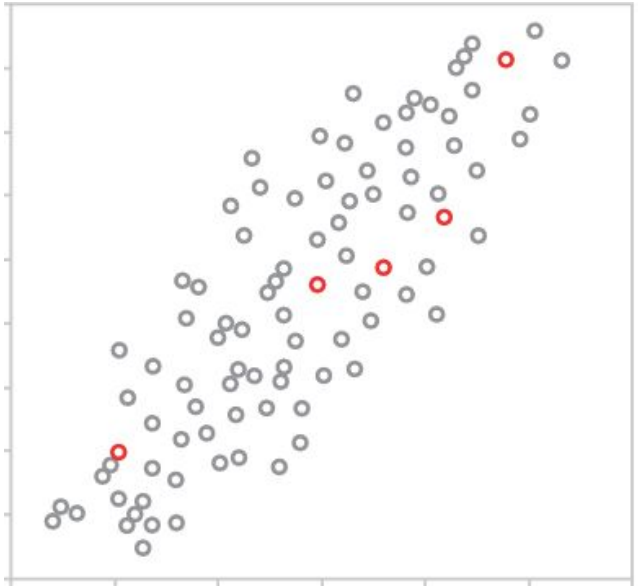
Ratio 100/99

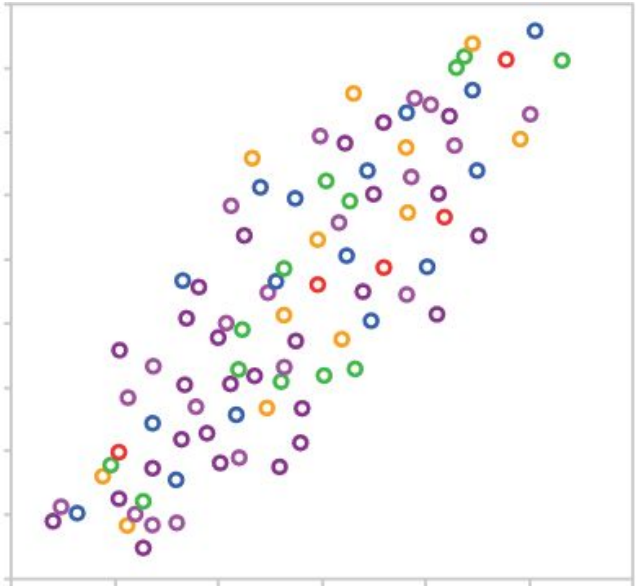
How visual perception works

When we use variations of a pre-attentive attribute to separate objects into a group, we must ensure that these variations are ease to distinguish from each other.

How visual perception works

Our ability to perceive variations of an attribute decreases as distractions pollute our field of view.





Overcoming the limits of memory

- Visual working memory is extremely limited.
- You can work with only 3 units of information at a time.
- But how much is 1 unit of information?

Overcoming the limits of memory

It depends on what we are observing and our skill/knowledge for that.

For example:

2007 Call Volume (in thousands)

	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
United States	1,983	2,343	2,593	2,283	2,574	2,838	2,382	2,634	2,938	2,739	2,983	3,493
Europe	574	636	673	593	644	679	593	139	599	583	602	690

In the table above, in general, each person stores a value as 1 unit of information.

Overcoming the limits of memory

If we show the same data as a line graph, each line can be considered 1 unit of information.

Advantage of visualization: when quantitative values are displayed as images with representative patterns, we can work with more information simultaneously, which multiplies the number and complexity of the insights.

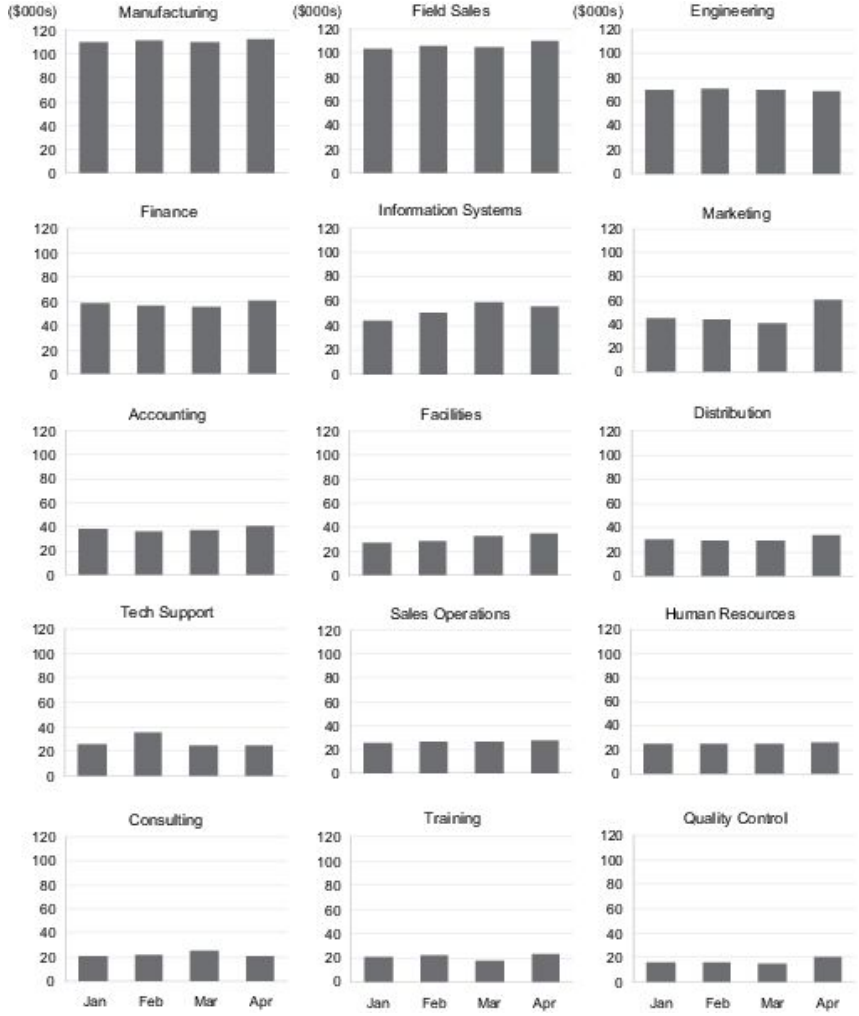
Overcoming the limits of memory

- Even so, the limits of visual working memory are considerable, so we use the “external” storage strategy: paper or computer.
- We must put the data we want to observe and compare on a single sheet or screen, avoiding page/scroll changes.

Overcoming the limits of memory

The figure represents 4 months of expenses for 15 departments so that they can be compared and provide an overview for the user.

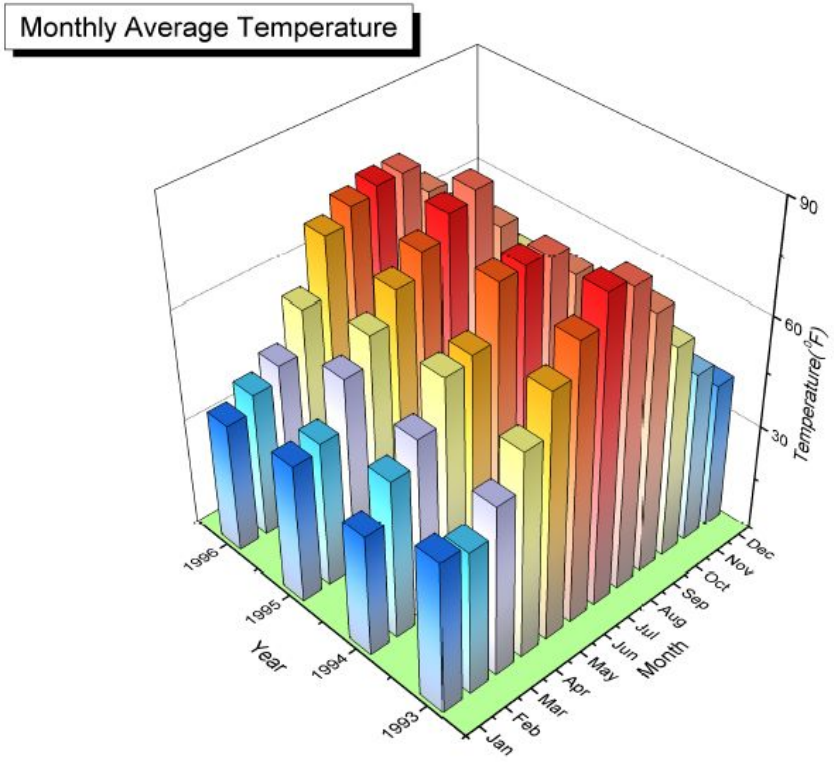
Small multiples



The building blocks of information visualization

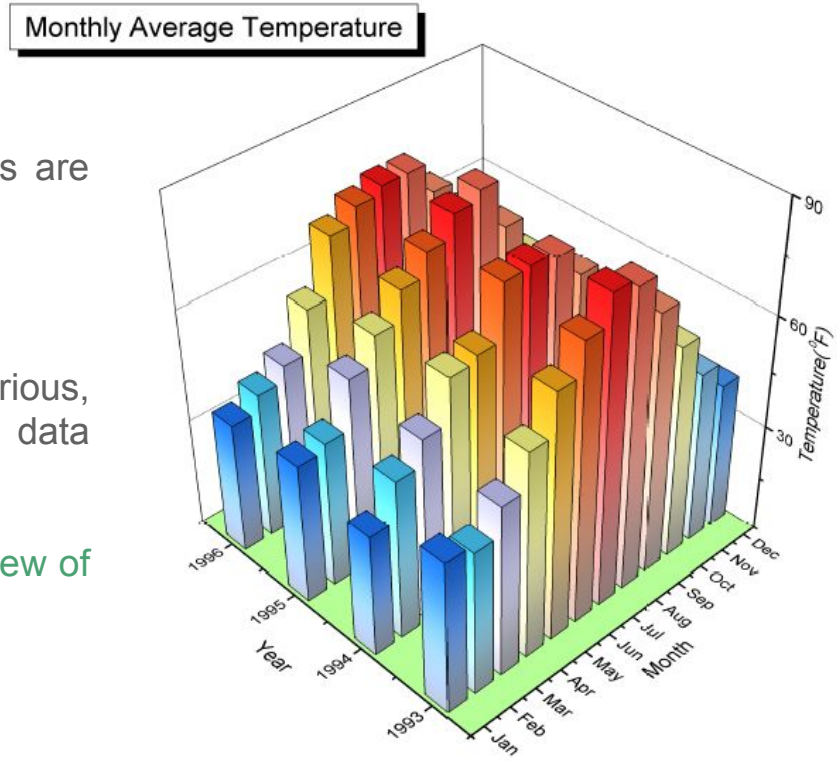
- Effective visualization is built based on the understanding of how we see and think.
- The blocks of visual perception consist of objects and properties (for example, pre-attentive attributes) that can represent quantitative data visually.
- A visualization shows quantitative relationships such as patterns, trends and exceptions in a visual way.
- An image of the data is not the goal, it is only the means. Information visualization is all about gaining understanding so we can make good decisions.

Comparing


Monthly Average Temperature
1996
1995
1994
1993
Year
Jan
Feb
Mar
Apr
May
Jun
Jul
Aug
Sep
Oct
Nov
Dec
Month
30
60
90
Temperature(°F)

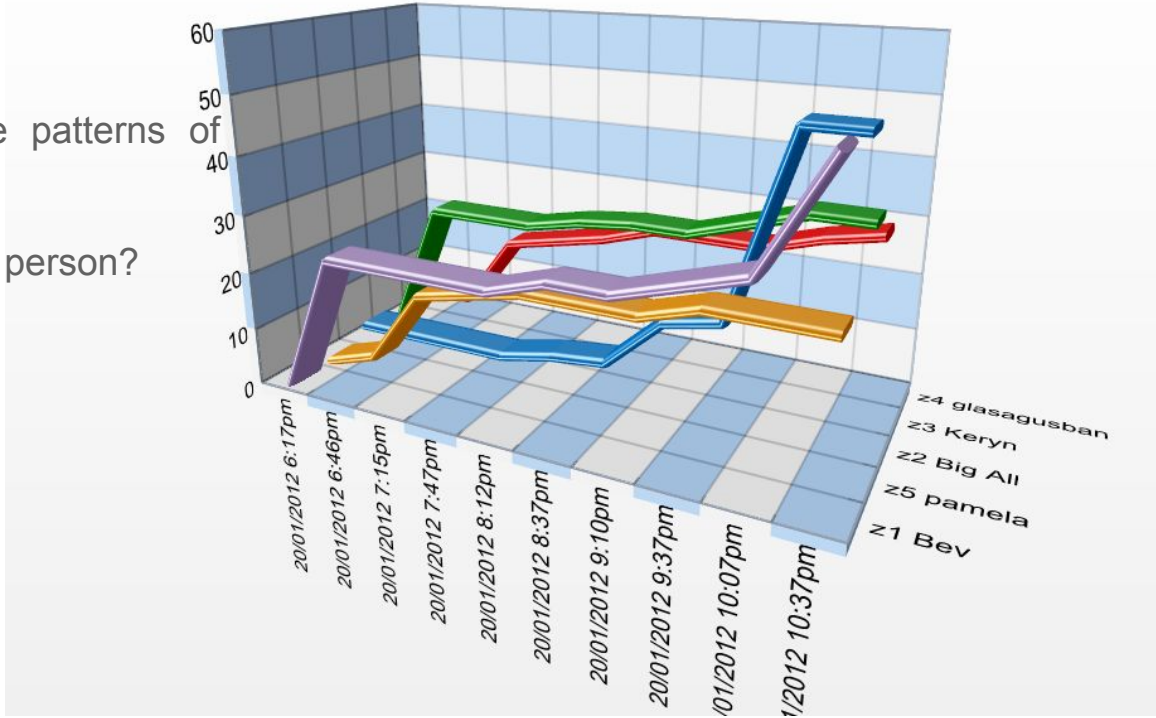
Comparing

- This graph suffers from occlusion: some bars are hidden behind others.
- But the original graph could be rotated!
- In addition to consuming time and being laborious, it undermines an important principle of data visualization:
    - See everything at once, giving an overview of the relationships between data.



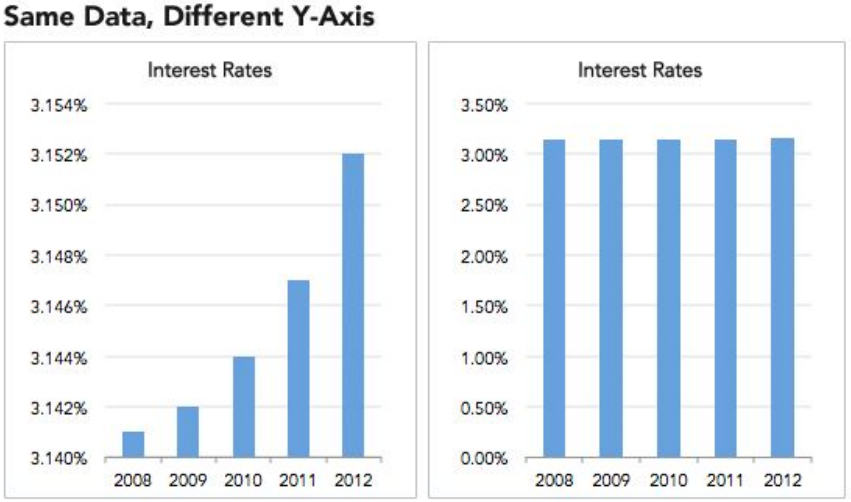
Comparing

- Interpret the graph and compare patterns of change over time.
- Which of the lines represents each person?
- How good is the graph?
- Is it possible to improve?



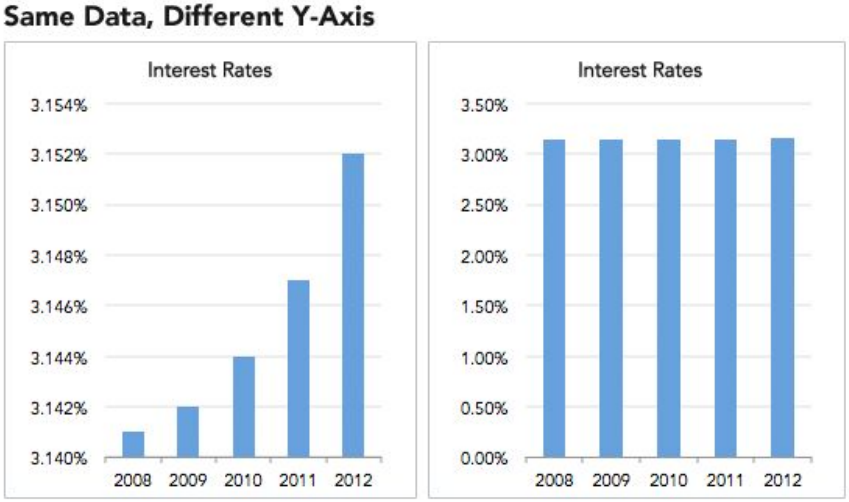
Comparing

How much bigger/smaller are the bars?



Comparing

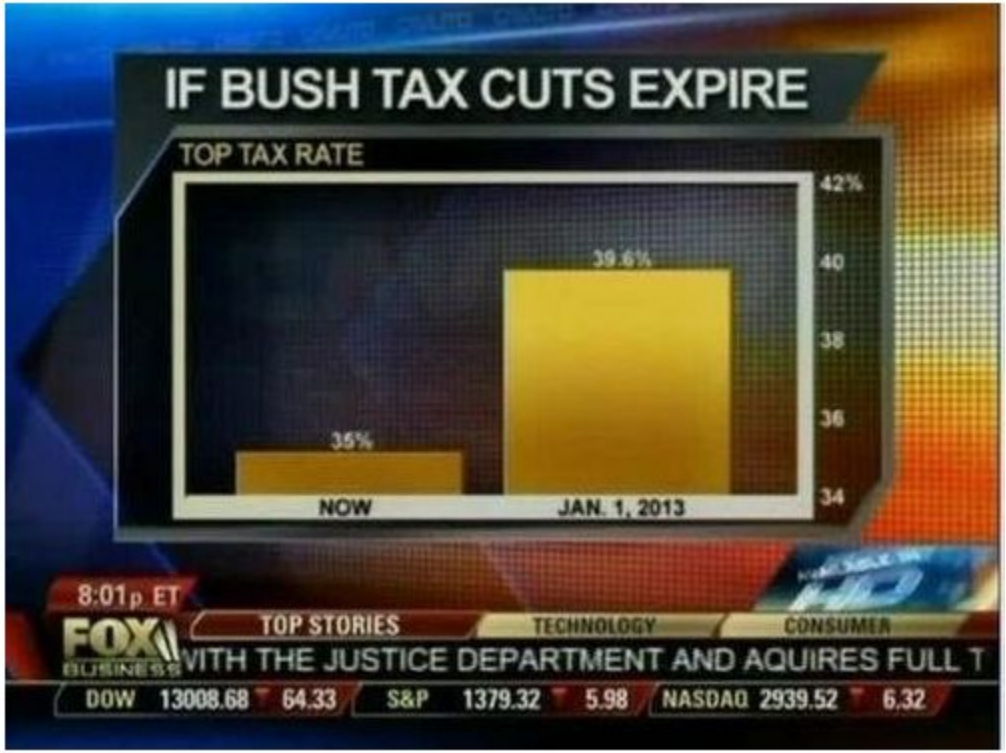
How much bigger/smaller are the bars?



Avoid truncating axes, especially the y. If you do, make it explicit.

Does it seem too much?

Comparing


IF BUSH TAX CUTS EXPIRE
TOP TAX RATE
42%
40
38
36
34
39.6%
35%
NOW
JAN. 1, 2013
8:01p ET
FOX BUSINESS
TOP STORIES
TECHNOLOGY
CONSUMER
WITH THE JUSTICE DEPARTMENT AND AQUIRES FULL T
DOW 13008.68 64.33
S&P 1379.32 5.98
NASDAQ 2939.52 6.32

Comparing

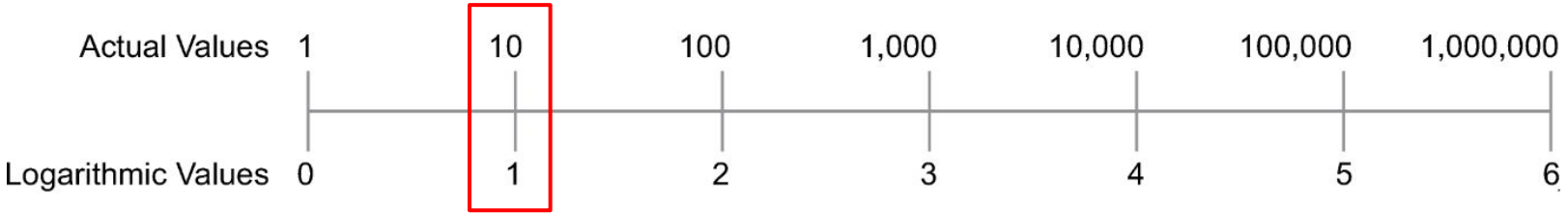

VOTO 2013
El pueblo decide
ELECCIÓN PRESIDENCIAL 2013
Venezolana de Televisión
PORCENTAJES
50,66
%
49,07
Nicolás Maduro Moros
Henrique Capriles Radonski

Re-scaling

It consists of changing the scale type of a quantitative graph (linear / logarithmic), which can reveal trends implicit in the graph.

Re-scaling

What is a logarithmic scale?



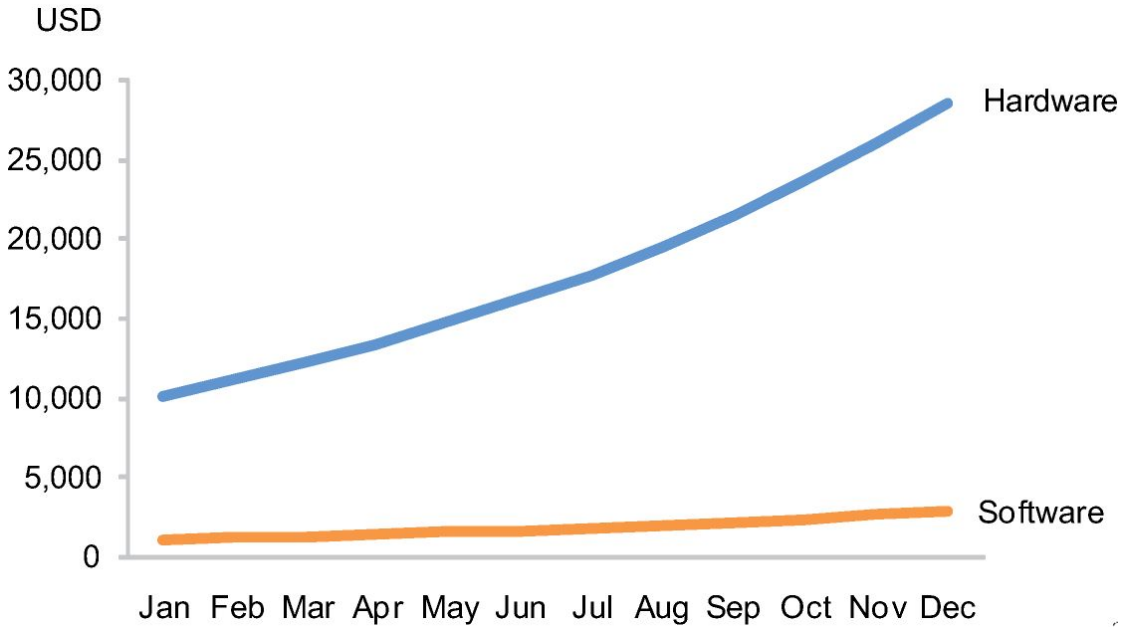
Remembering that logarithm is exponent, the value 1 on the logarithmic scale (base 10) corresponds to 10 on the linear scale because:

$\log 10 = 1$ e $10^1 = 10$

$\log 100 = 2$ e $10^2 = 100$

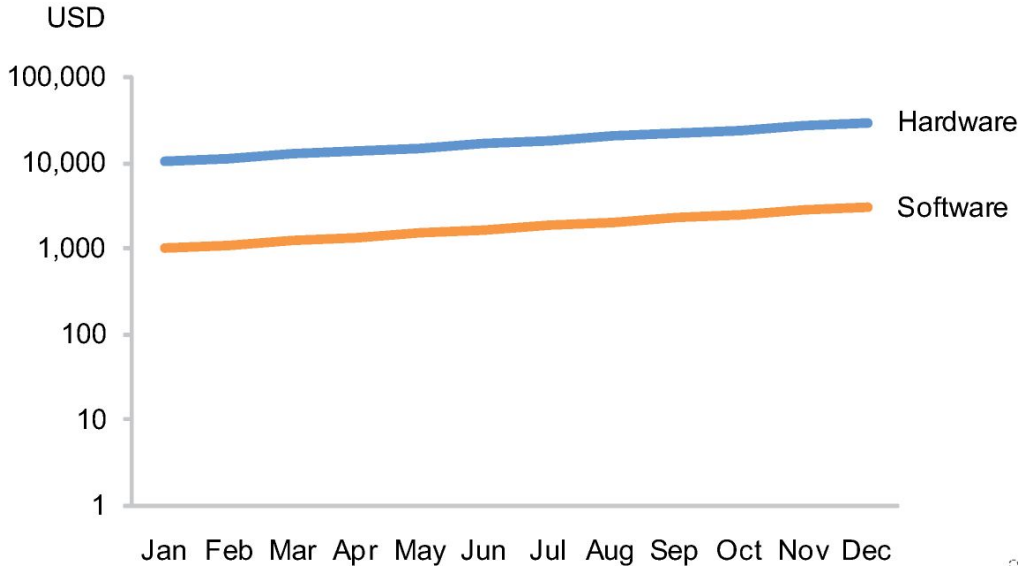
Re-scaling

Below we see the sales of hardware and software. Which grows faster?



Re-scaling

Using logarithmic scale ...



Directed vs. exploratory navigation

Navigation can be divided into directed vs. exploratory:

Directed: when there is a specific question that must be answered and the purpose of navigation is to find an answer.

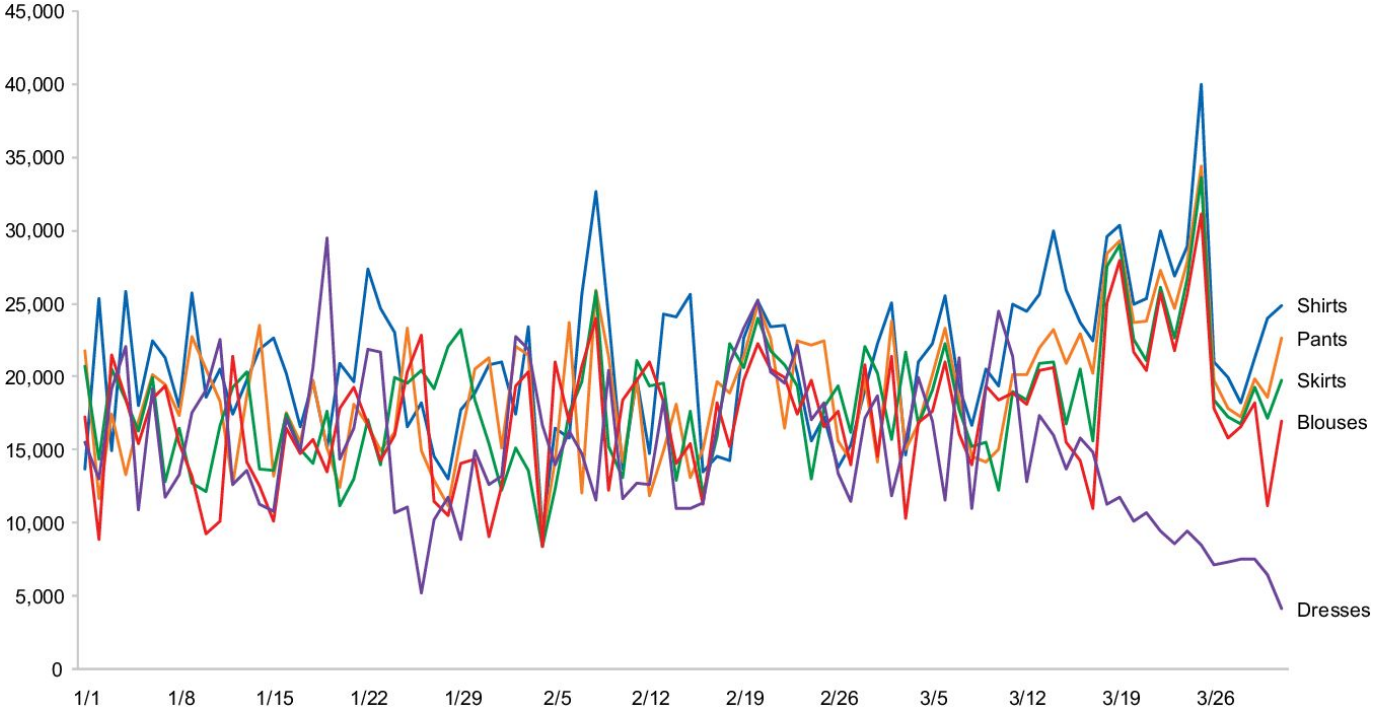
Exploratory: navigation is carried out without previously determining what you want to find and, in the process, questions arise and we try to find an answer.

Shneiderman's Mantra

Overview first, zoom and filter, then details-on-demand.

Overview

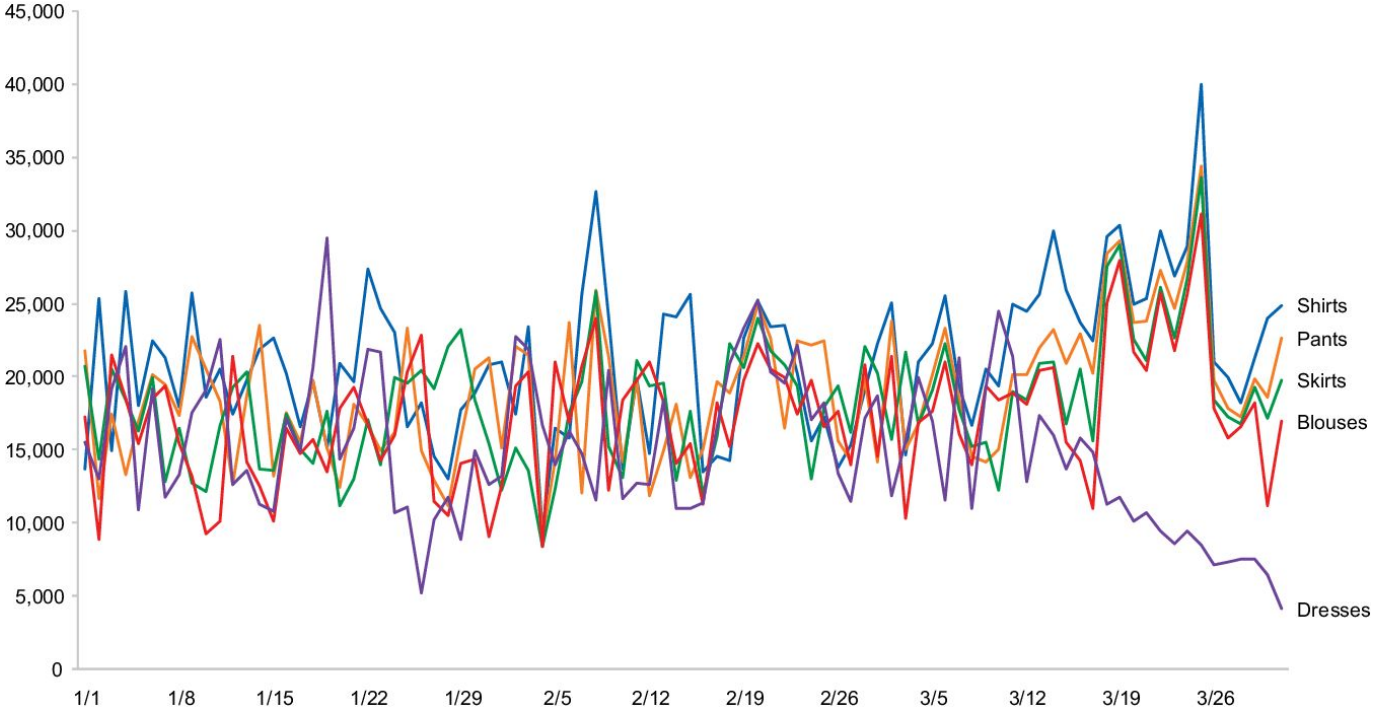
Observe the chart on the side that represents the daily sale of 5 types of clothing for 3 months.



Overview

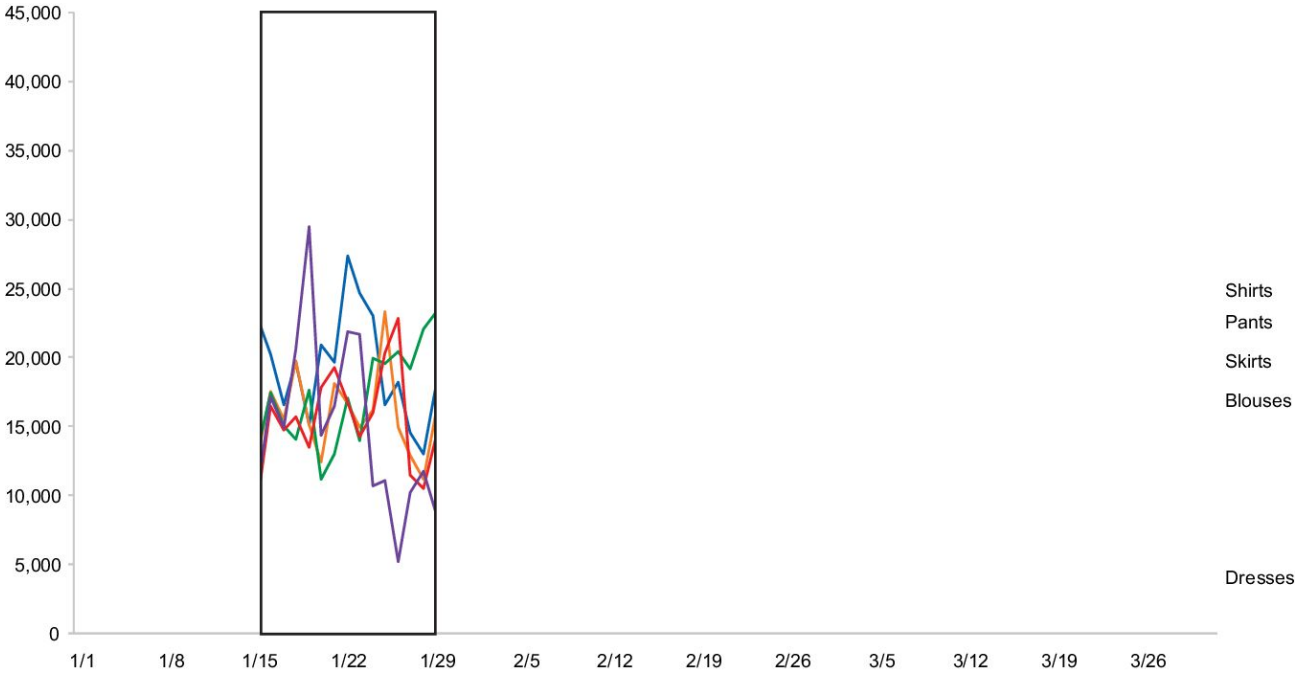
Observe the chart on the side that represents the daily sale of 5 types of clothing for 3 months.

Can we compare blouses and dresses?



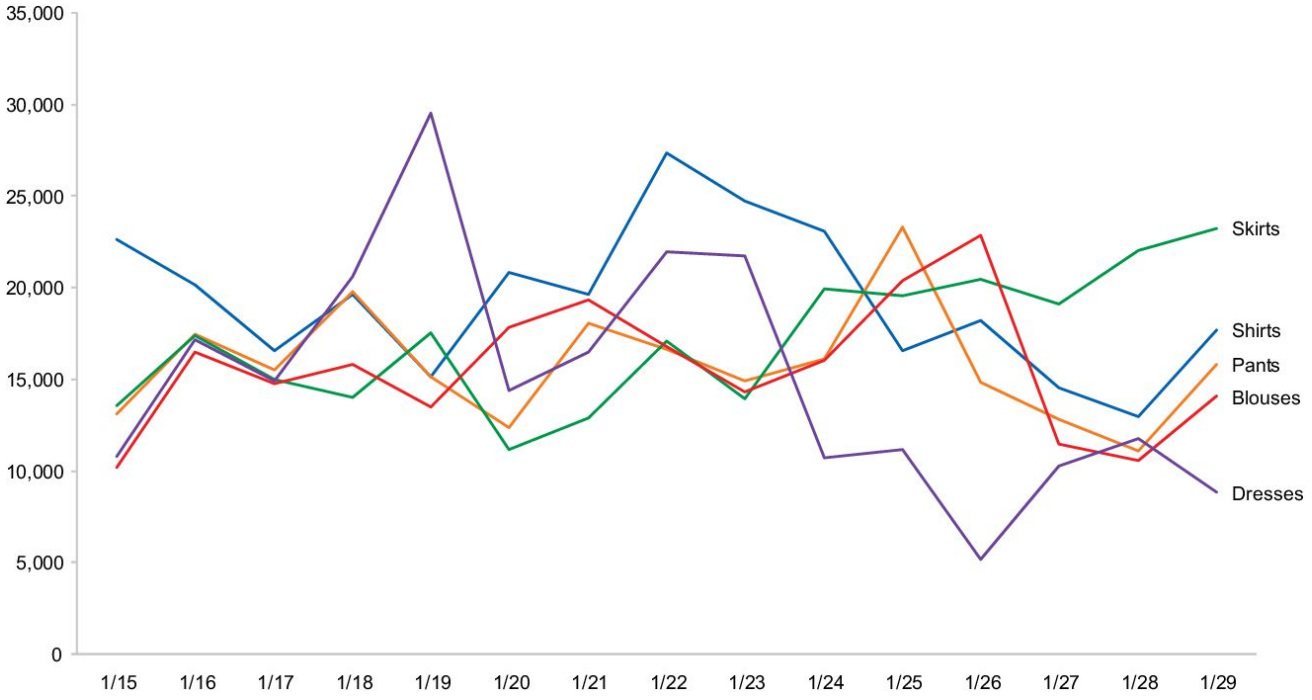
Data set reduction

When we spot a particular point of interest, we can zoom in on it.



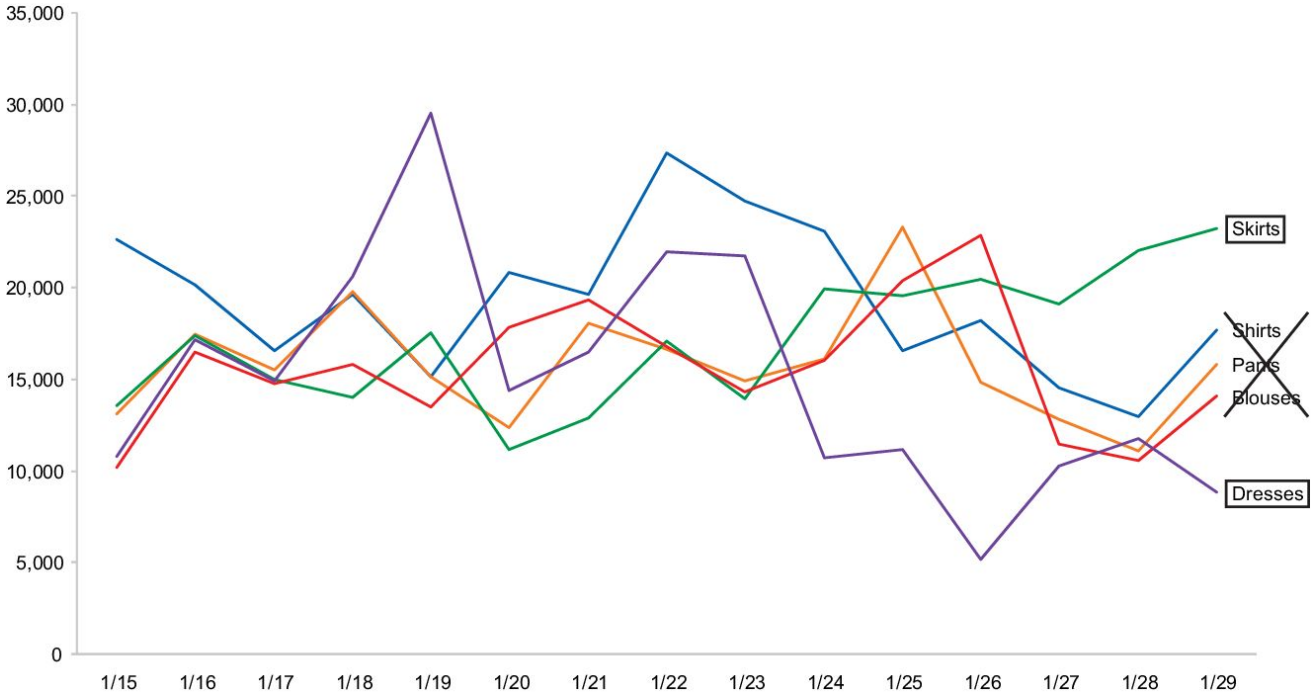
Zoom

Allow us to examine it more closely and in detail.



Filtering

To better focus on the relevant data, we must remove what is extraneous to our investigation.



Filtering

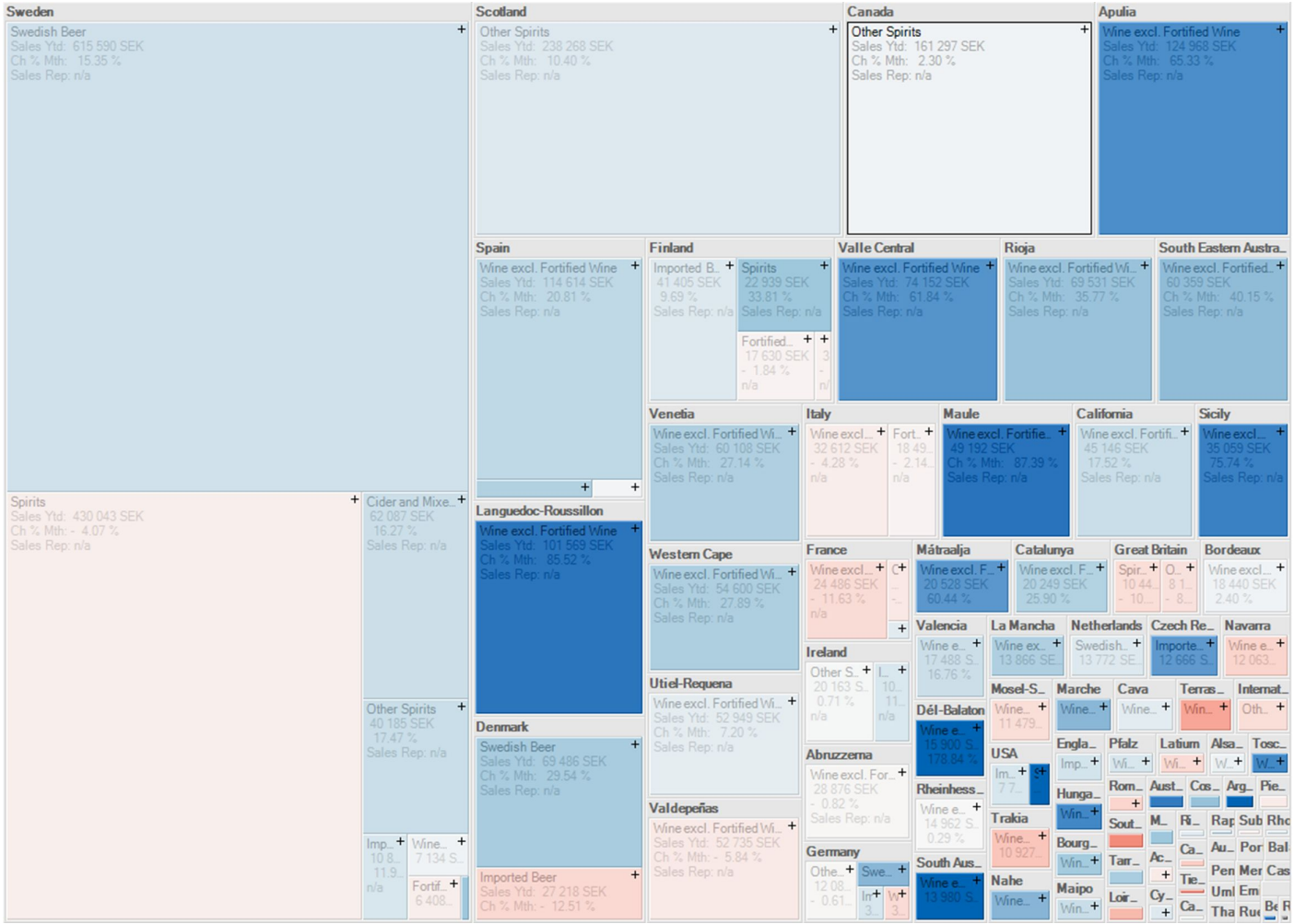
Removing data that is not the focus at the moment avoids distractions and allows you to focus your analysis on what is relevant.

Details-on-demand

In some cases, we want precise details that are not provided by the visualization. One possible solution is to use a pop-up.

Treemap

This treemap represents Swedish beverage sales.

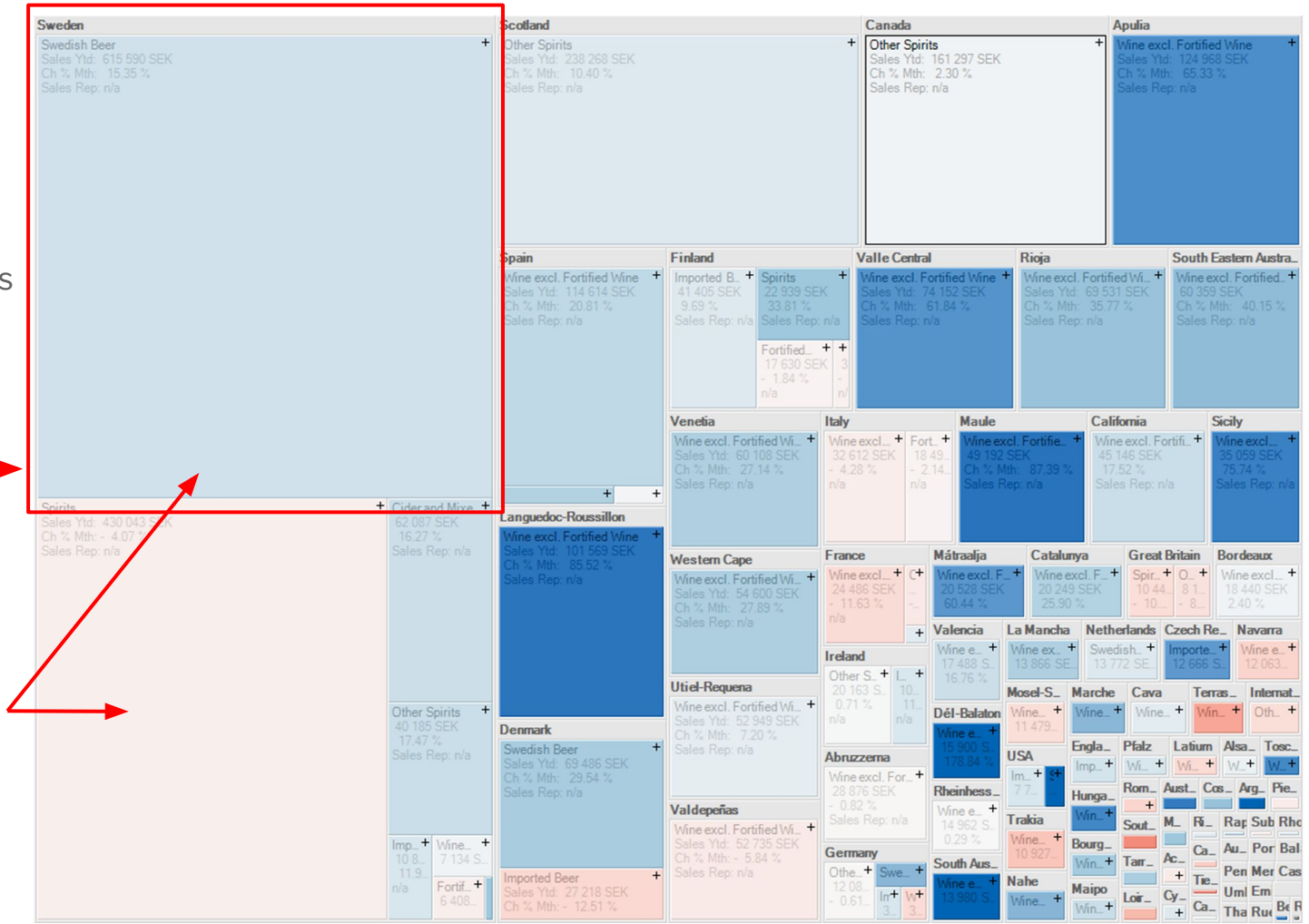


Treemap

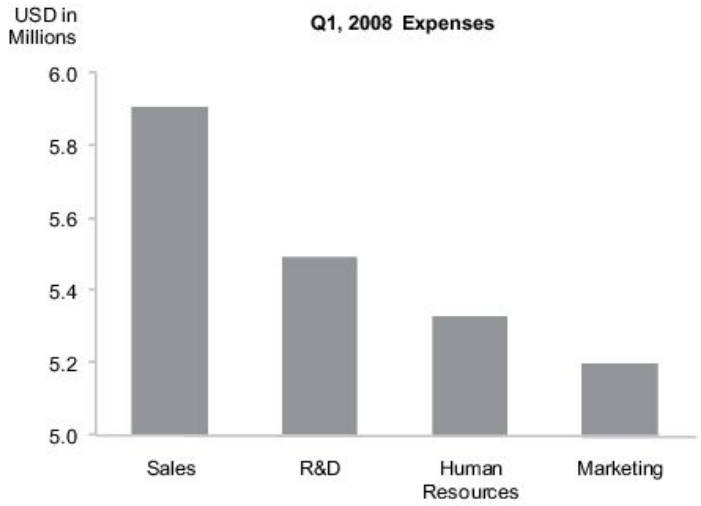
A treemap can display 2 quantitative variables simultaneously.

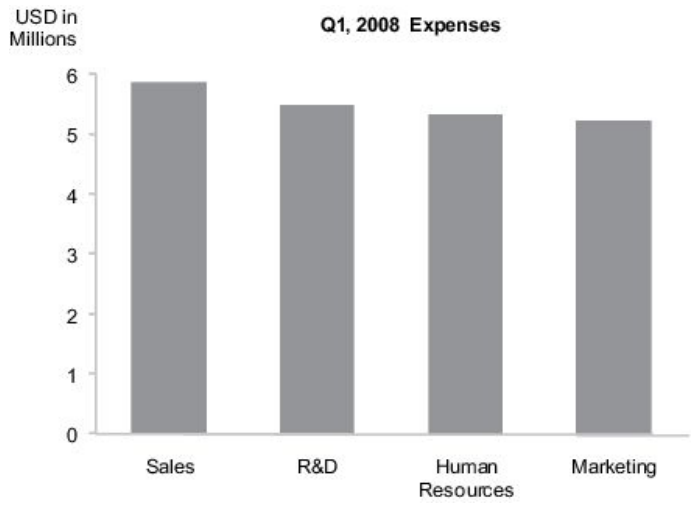
Size: Year-to-date sales

Color: percentage of change in sales from last month



Optimal quantitative scales





Which bar graph is most appropriate?

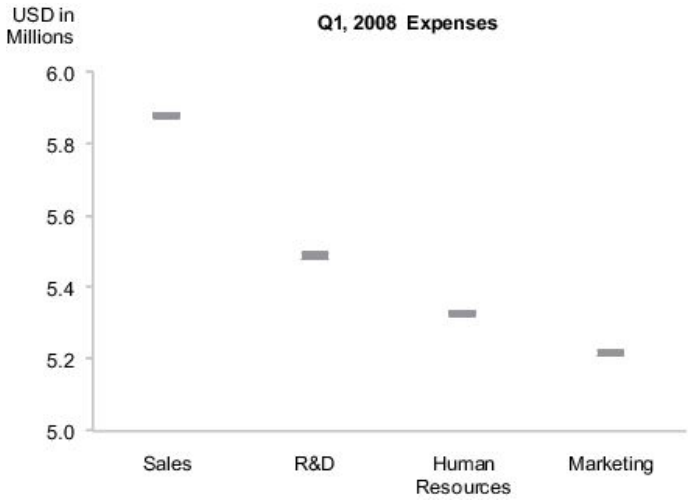
Optimal quantitative scales

- Our perception captures differences, contrasts, not absolute values. We perceive these differences proportionally.
- When we need to compare values and patterns, if the differences between them stand out, our task is made easier.
- In a graph that uses positions to represent values, it is interesting to distribute objects as much as possible in the available space, avoiding crowding them.

Optimal quantitative scales

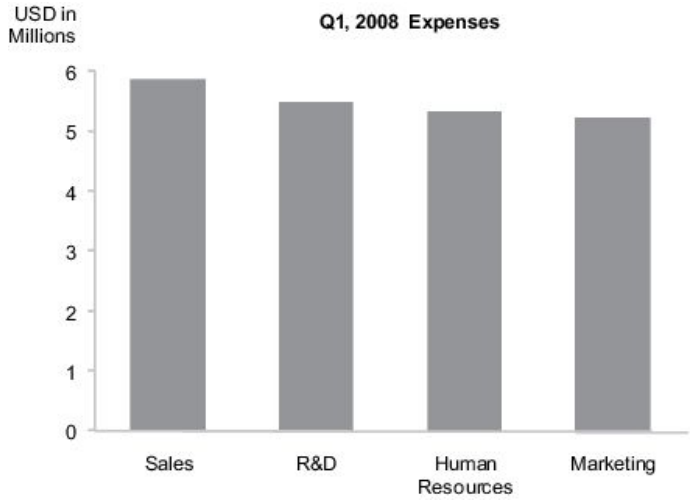
- The strategy that we are going to see is called narrowing of scale.
- The scale should start just below the lowest value and end just above the highest value.
- In a bar graph, there should be no narrowing of the scale.
- How to represent the data in our example, highlighting the differences between expenses and still maintaining a fair comparison?

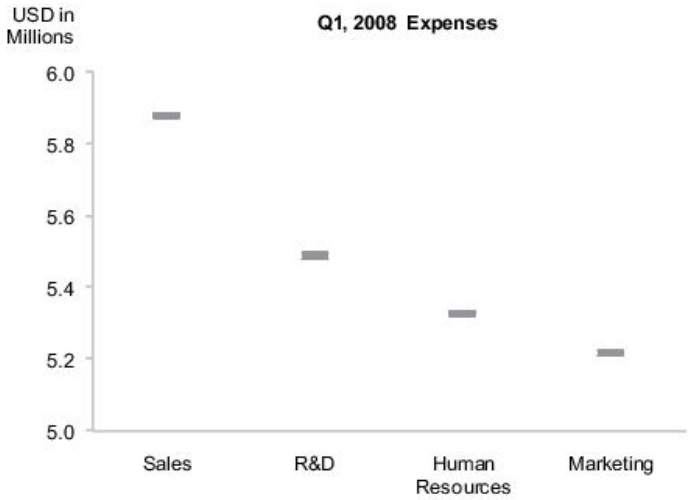
Optimal quantitative scales



We can replace the bars with data points and narrow the scale.

Optimal quantitative scales

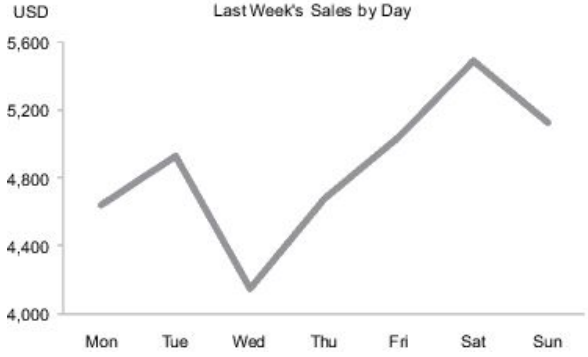


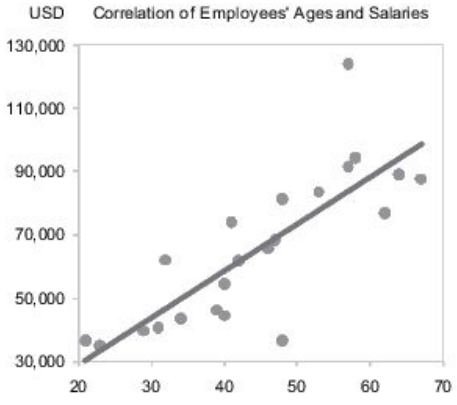


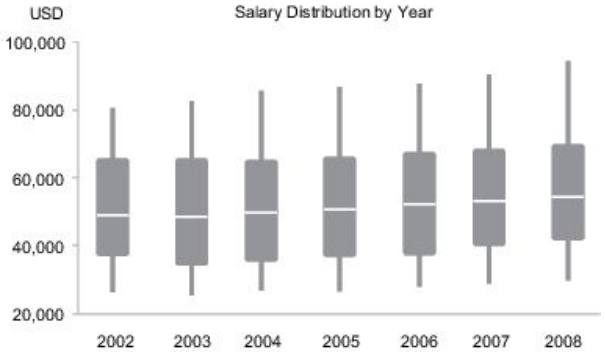
Note that it is easier to compare the values and patterns of the graph on the left when represented as points with values between 0 and 6 on the y axis.

Optimal quantitative scales

Scales can be narrowed in line graphs, scatter plots and box plots.







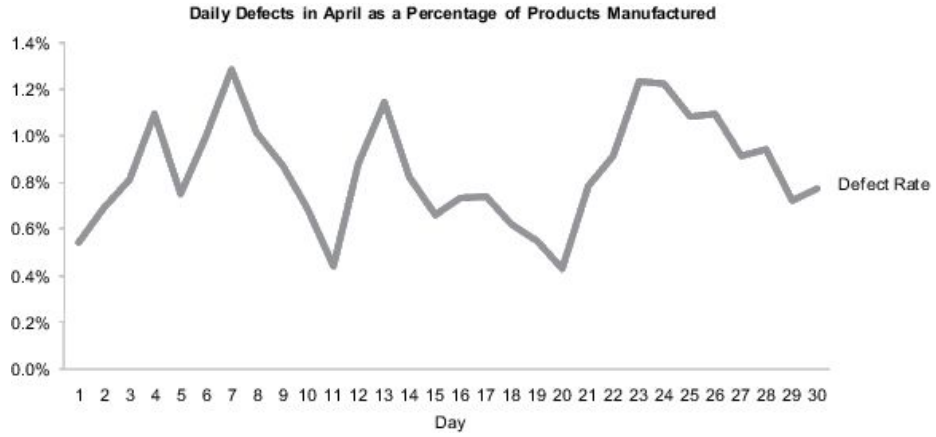
Optimal quantitative scales

Rule of thumb

- For bar charts, start the scale at 0 and end just above the highest value.
- For all other types of charts, start the scale just below the lowest value and end just above the highest value.
- Start and finish the scale in round numbers. Do the same for breaks.

Reference lines and regions

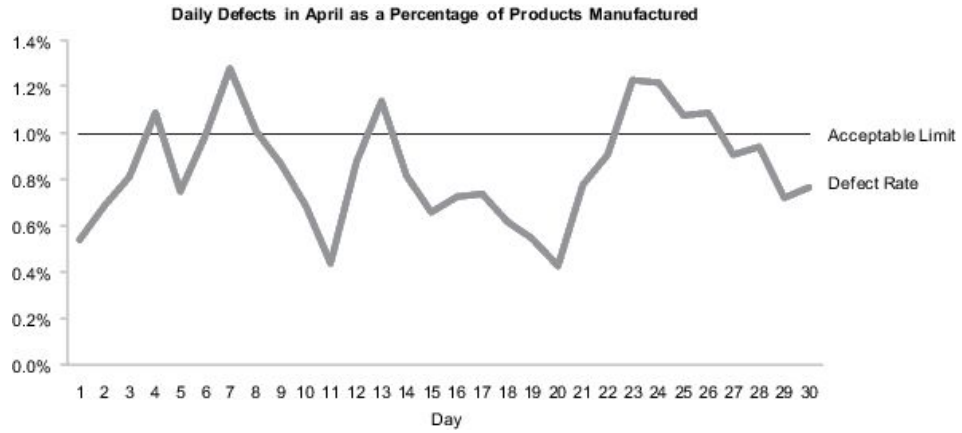
In this example, defects should not exceed 1% of products manufactured in one day. Which days exceeded 1%?



Reference lines and regions

With the reference line it's easier:

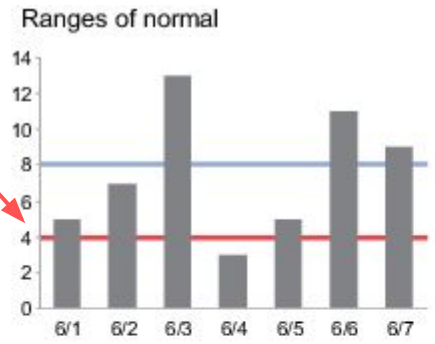
- to count the days when the number of defects exceeds the threshold;
- to see how much the limit has been exceeded

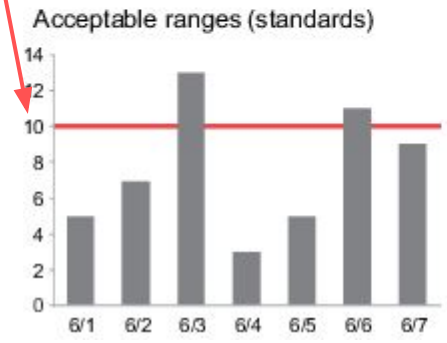


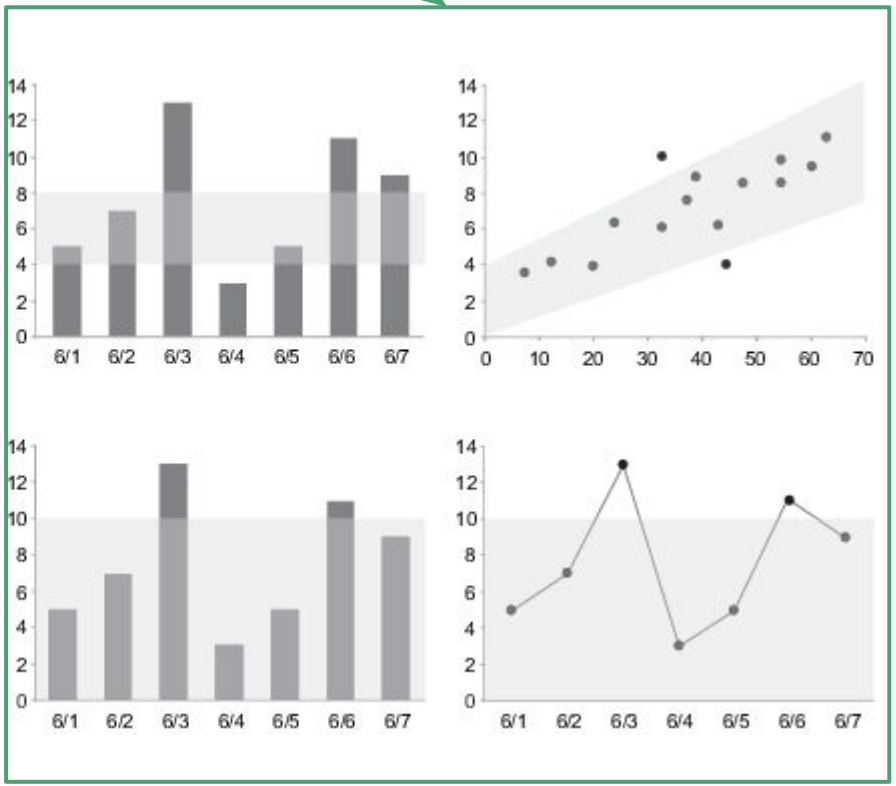
Trend and exception

Reference region

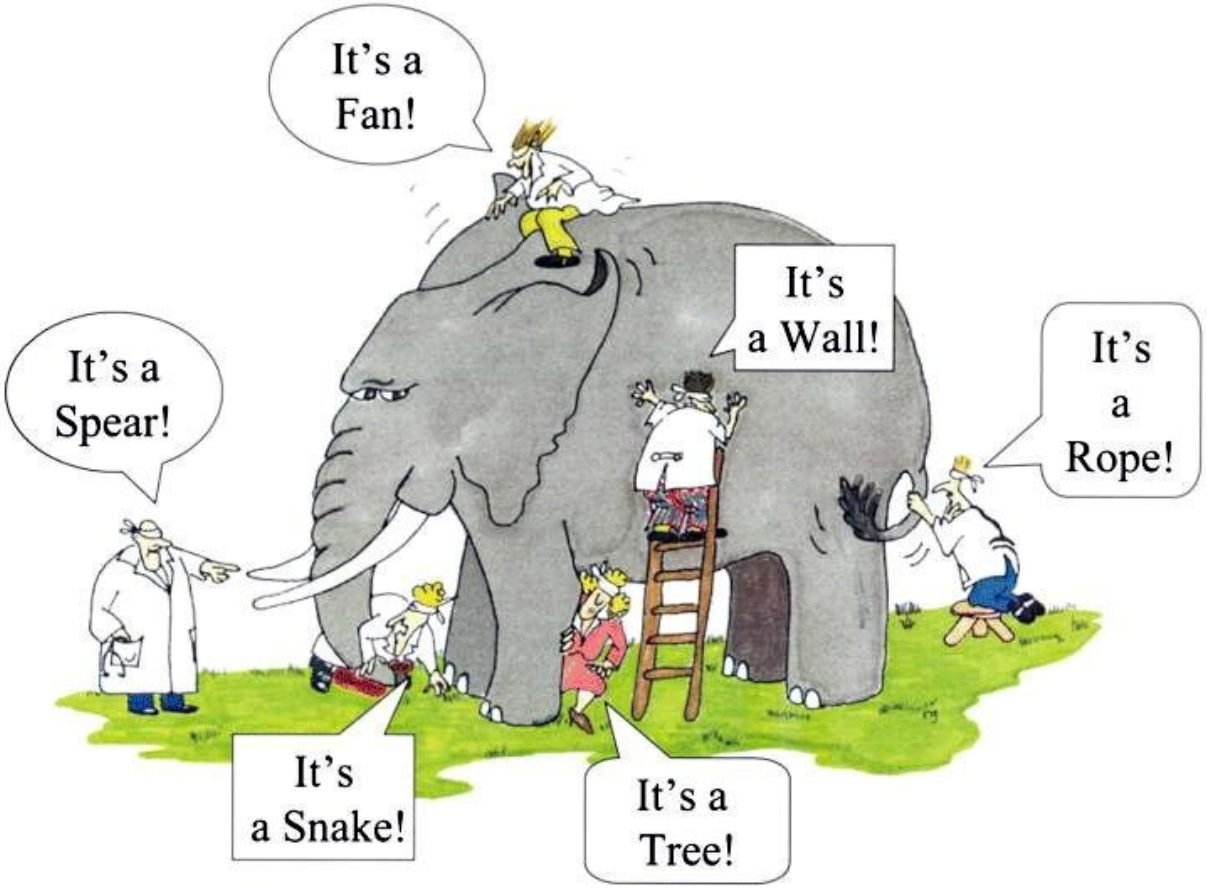
Reference lines



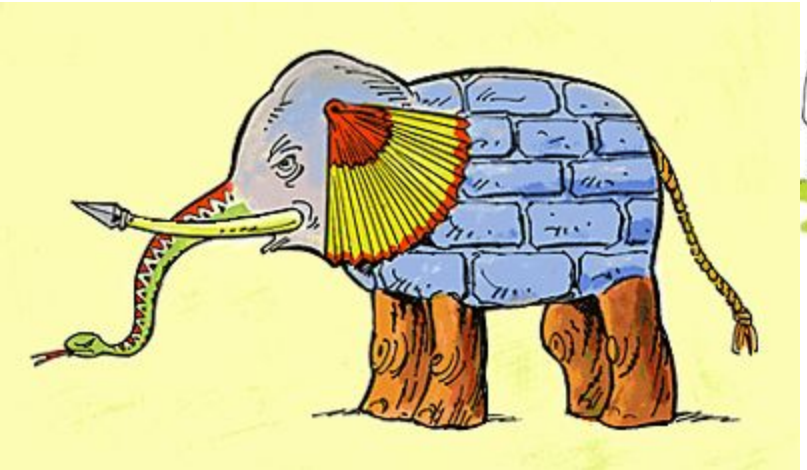


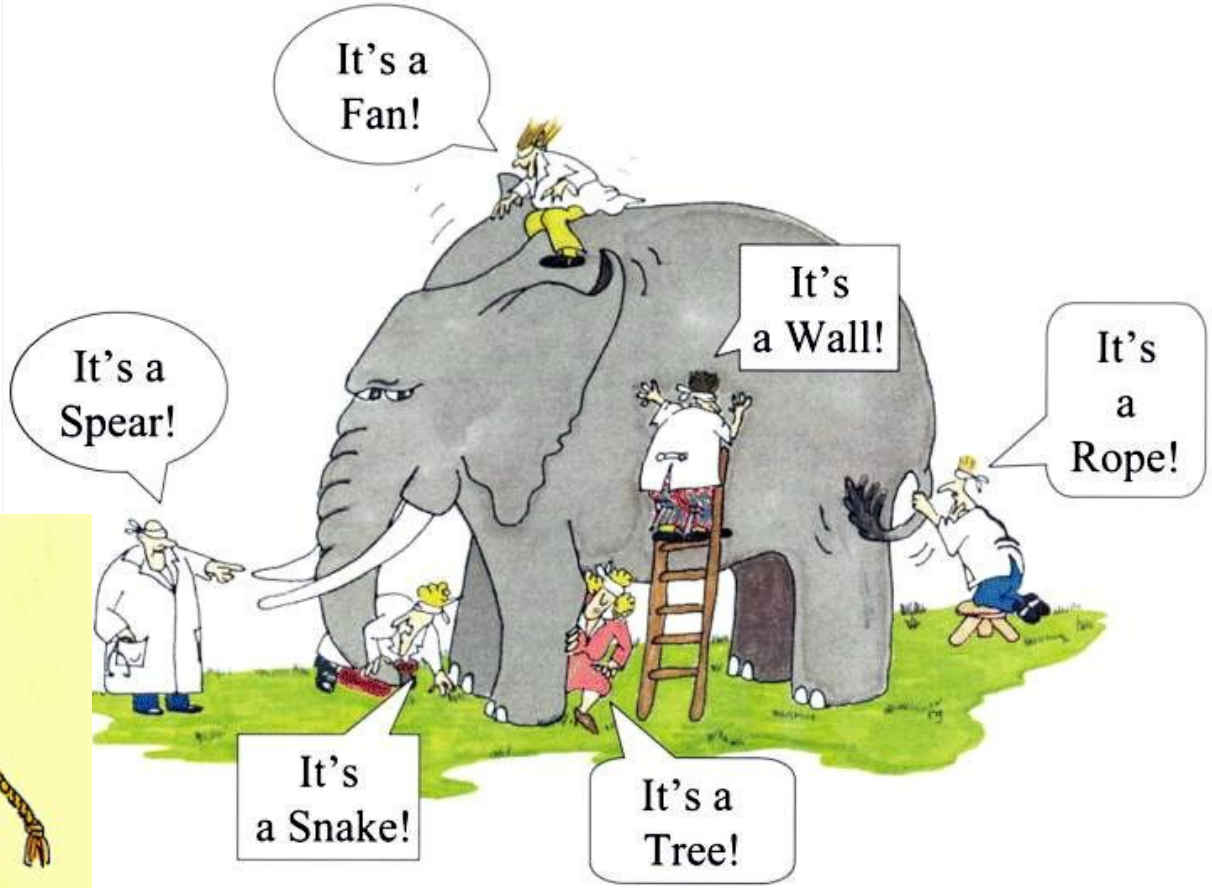


Multiple concurrent views and brushing



Multiple concurrent views and brushing

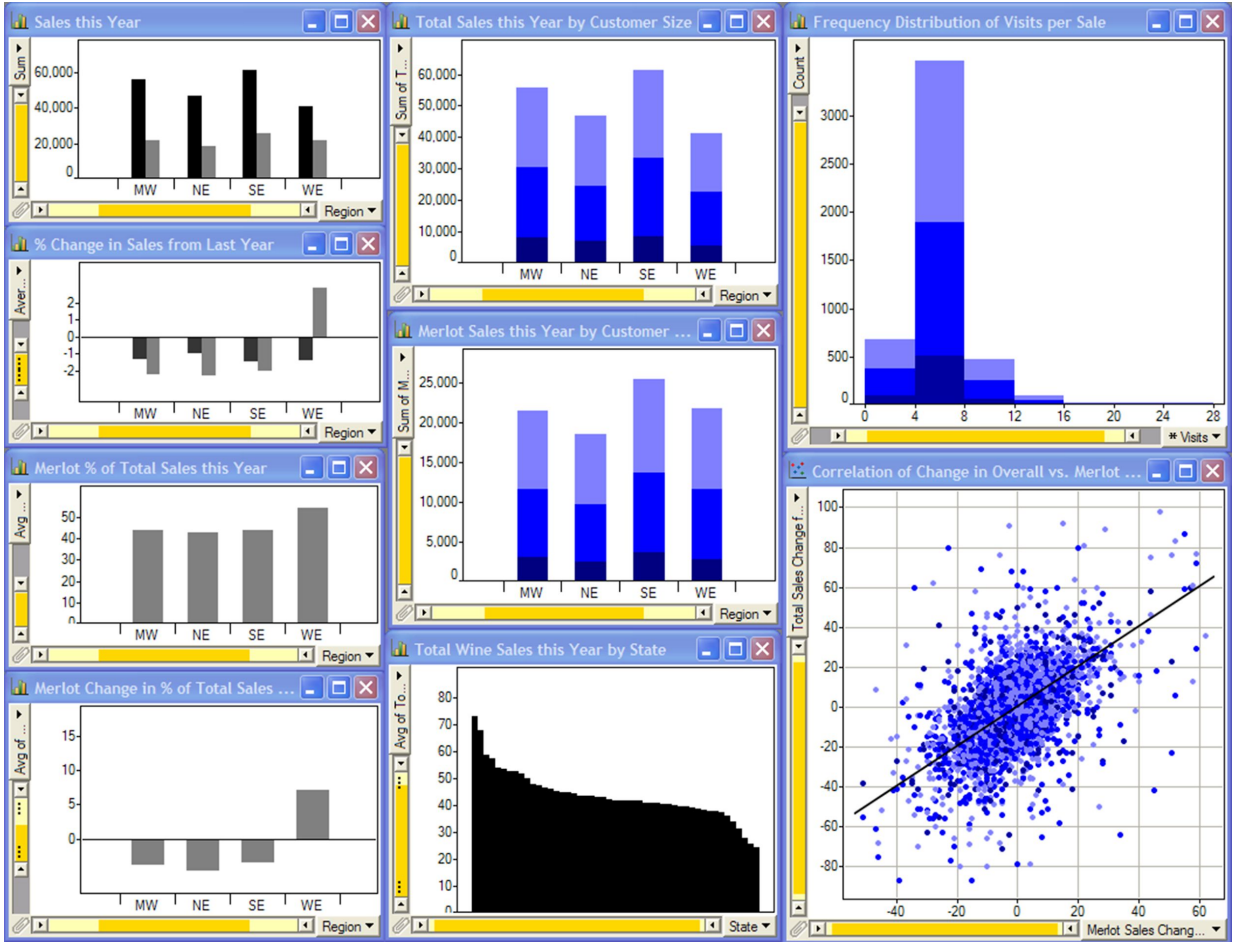





A
Monthly Sales
West
Central
East
South
Espresso
Coffee
Herbal Tea
Tea
3K
2K
1K
0K
Fe..
May
Aug..
Nov..
B
Budget vs. Actual
Columbian
Lemon
Caffe Mocha
DecafEspresso
Chamomile
Darjeeling
Earl Grey
Decaf Irish Crea..
Caffe Latte
Mint
Green Tea
Amaretto
RegularEspresso
0K
20K
40K
60K
Measure Names
Budget
Actual Sales
Product Type
Coffee
Espresso
Herbal Tea
Tea
Product Type
Coffee
Espresso
Herbal Tea
Tea
Product
Amaretto
Caffe Latte
Caffe Mocha
Chamomile
Columbian
Darjeeling
Decaf Espres
Decaf Irish Cr
C
Sales by State
California
New York
Illinois
Nevada
Iowa
Colorado
Oregon
Washington
Texas
Florida
Utah
Ohio
Wisconsin
Massachusetts
Oklahoma
Connecticut
Missouri
Louisiana
New Mexico
New Hampshire
0K
10K
20K
30K
40K
50K
D
Marketing and Sales
Sum of Sales
8K
6K
4K
2K
0K
0
500
1000
1500
Sum of Marketing
E
Product Sales
Columbian
Lemon
Caffe Mocha
DecafEspresso
Chamomile
Darjeeling
Earl Grey
Decaf Irish Cream
Caffe Latte
Mint
Green Tea
Amaretto
RegularEspresso
0K
50K
100K
State
California
Colorado
Connecticut
Florida
Illinois
Sales
17
912
Profit
-638.14
778.41

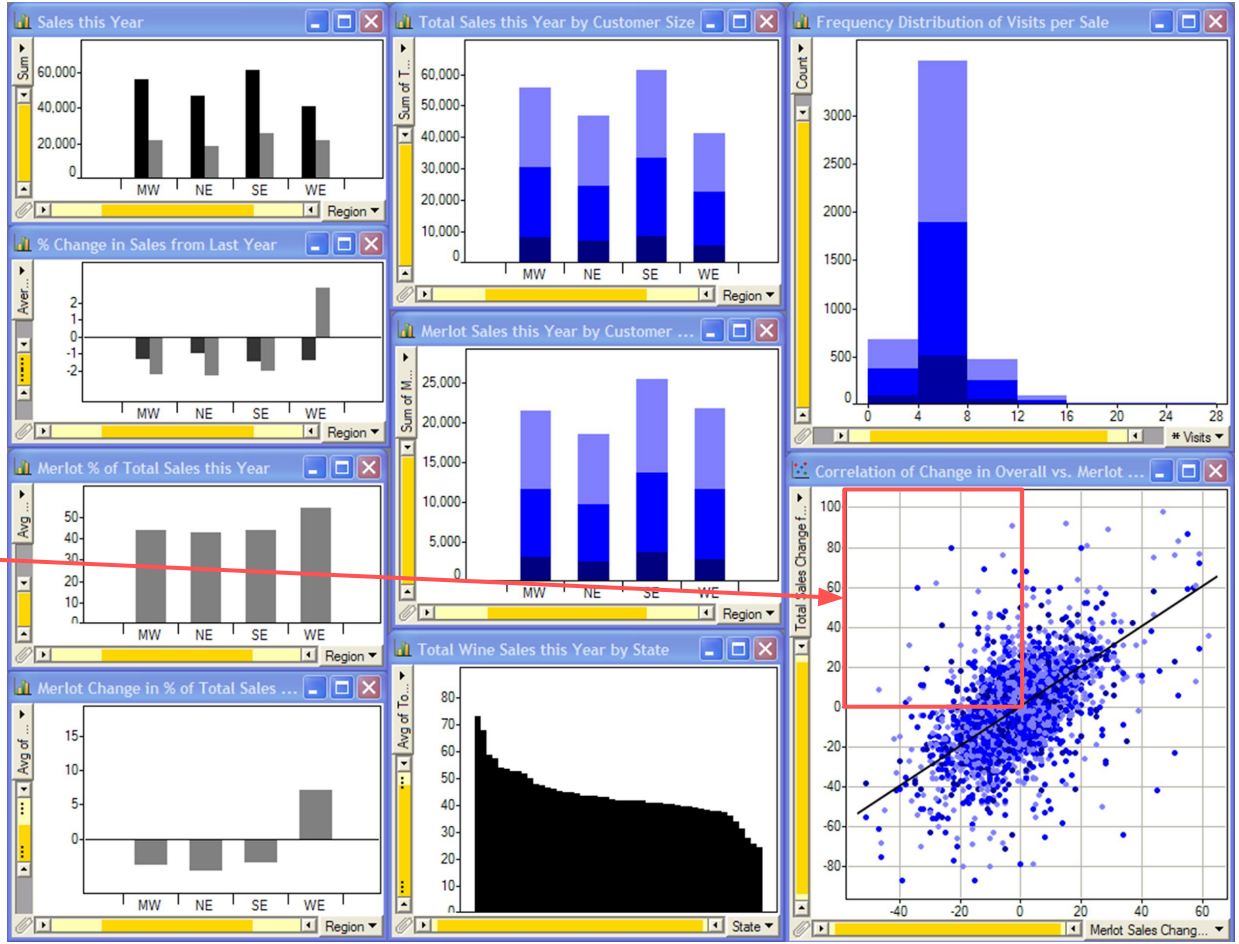
Brushing

We want to highlight customers whose sales of Merlot have declined despite the fact that wine sales, in general, have increased



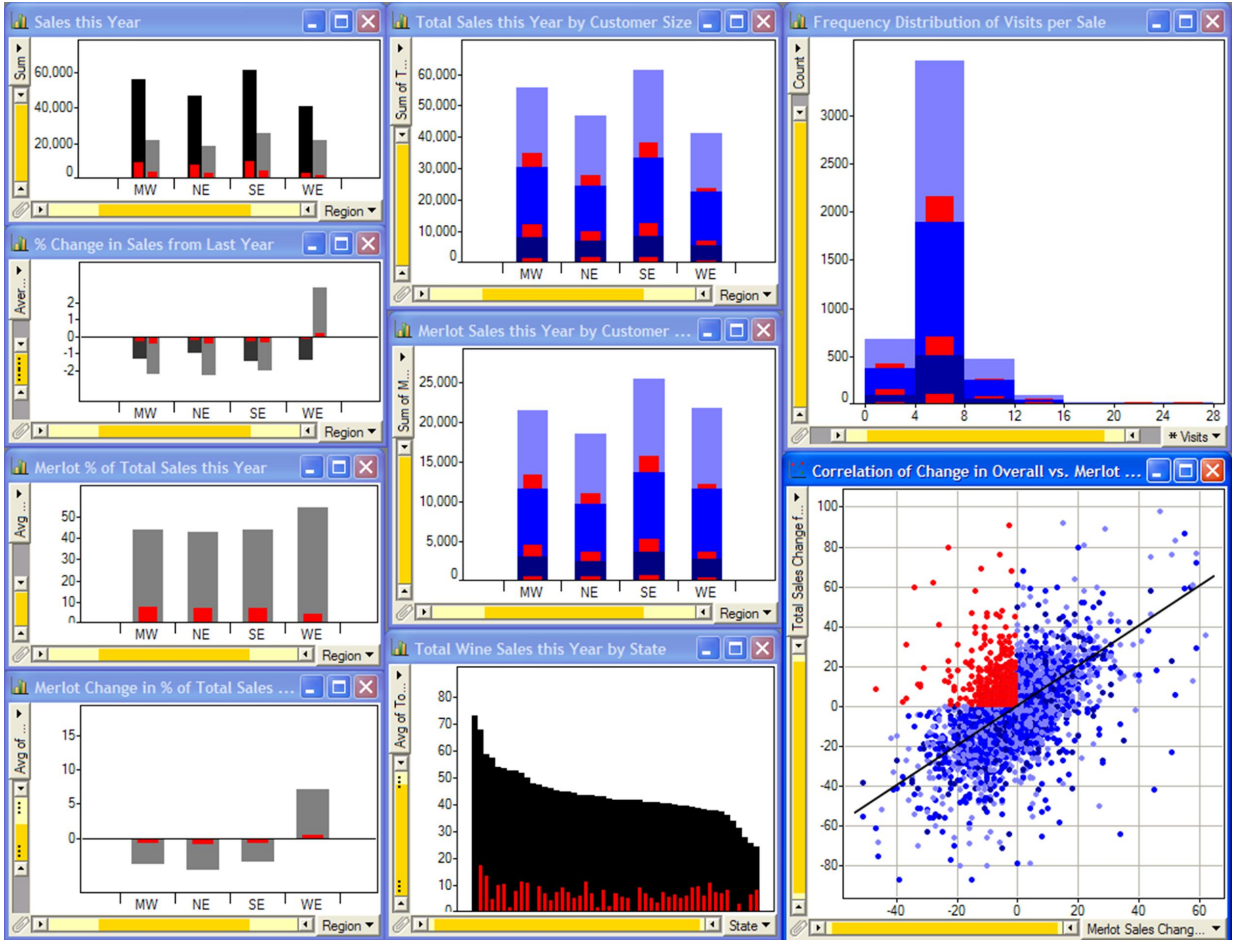
Brushing

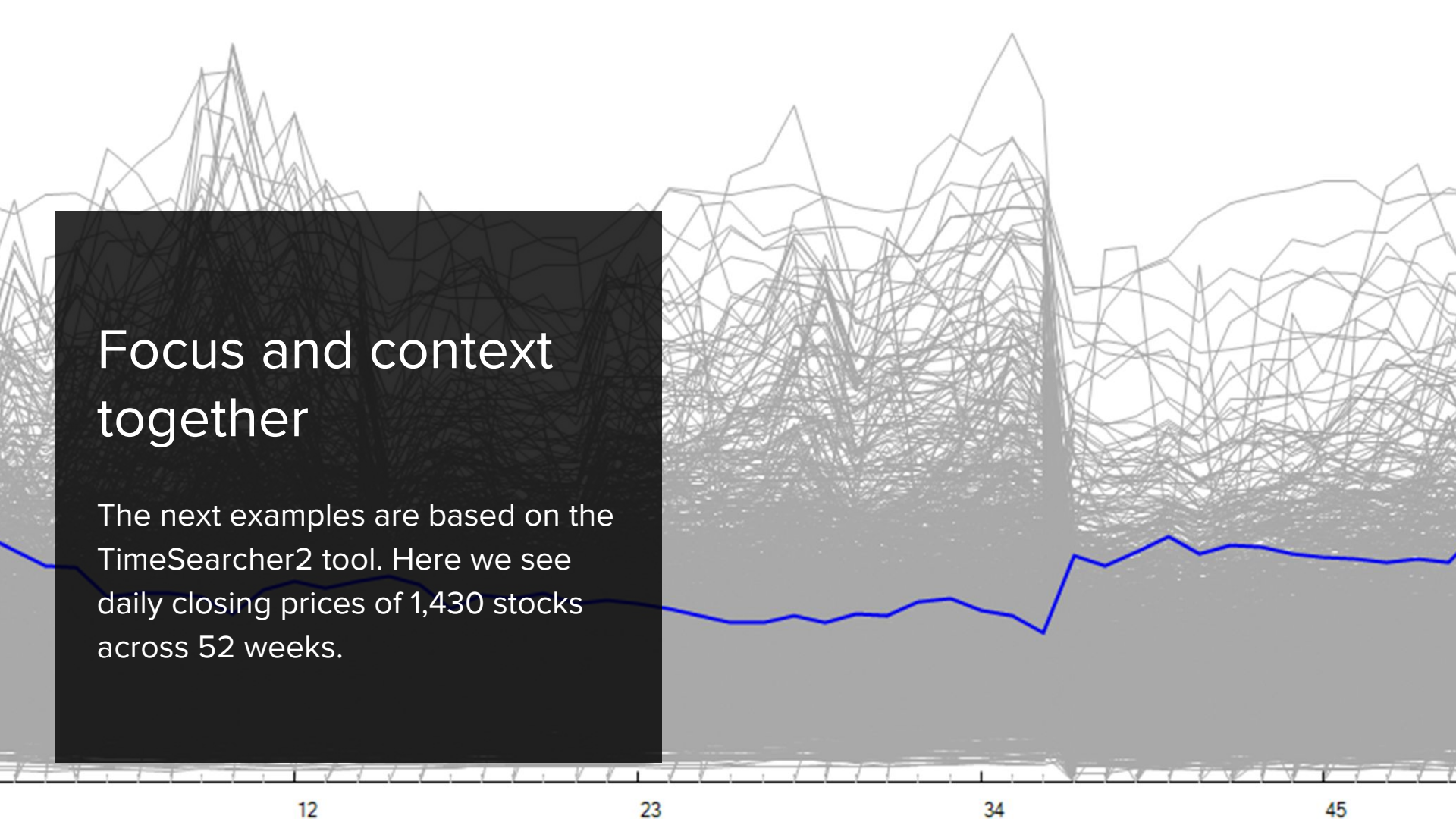
What if we could select that data on the scatterplot?



Brushing

Did brushing help us to notice something interesting here?





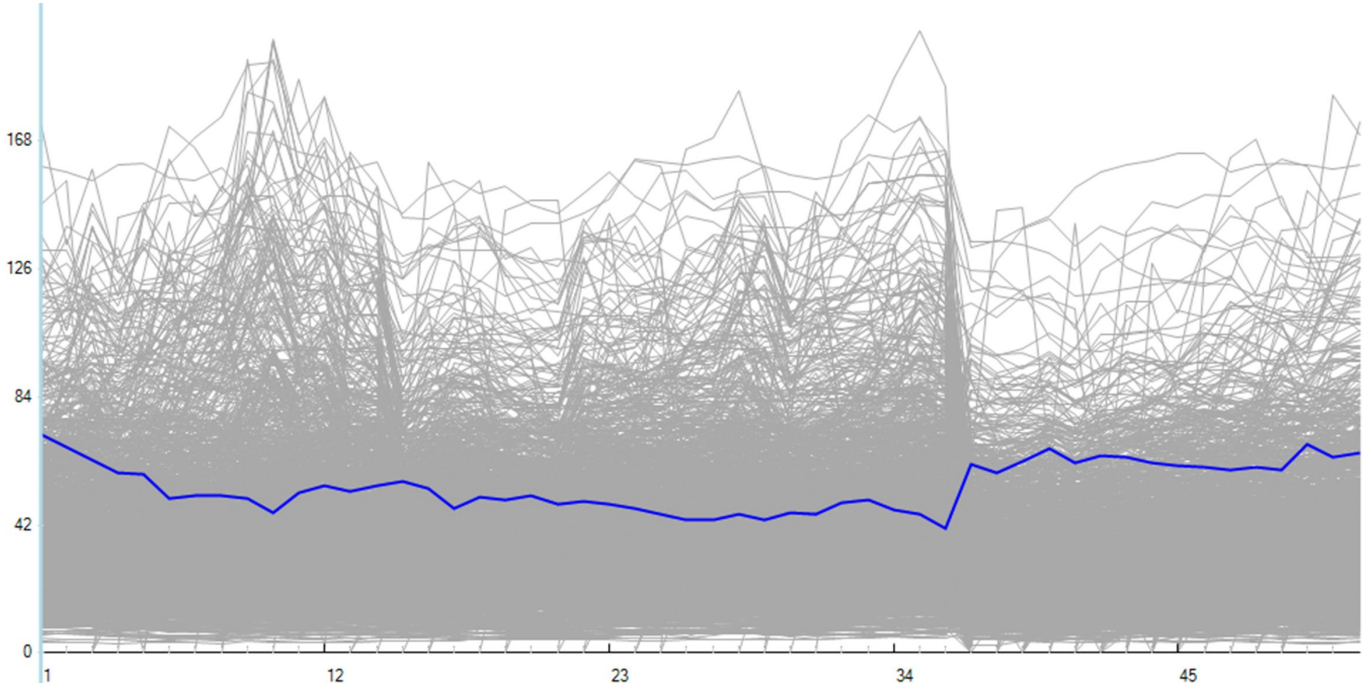
Focus and context together

The next examples are based on the TimeSearcher2 tool. Here we see daily closing prices of 1,430 stocks across 52 weeks.

Focus and context together

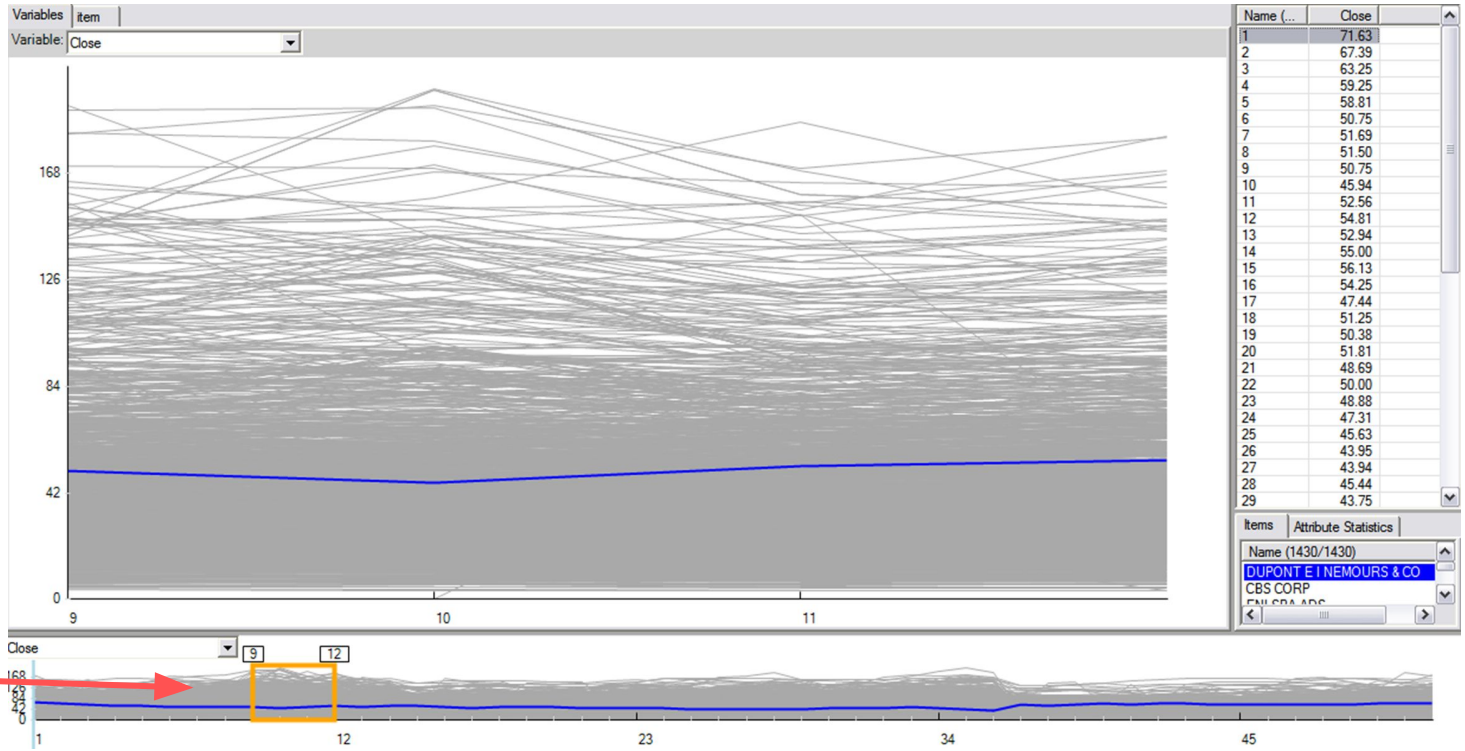
The blue line represents a single stock.

What if we want to investigate further what happened between weeks 9 and 12?



Focus and context together

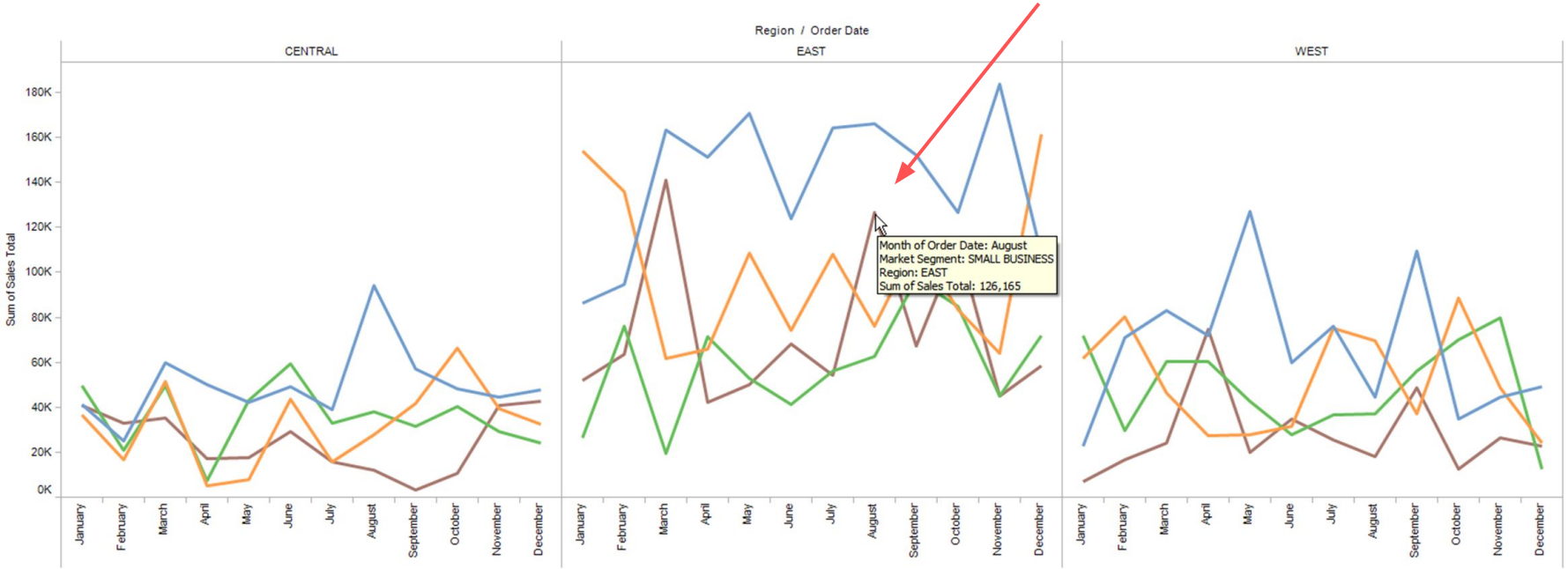
The tool allows us to select a time interval and represent only that interval.



Without losing context

Details on demand

When you hover over a point, a pop-up with data appears. When you move the mouse, the pop-up disappears.

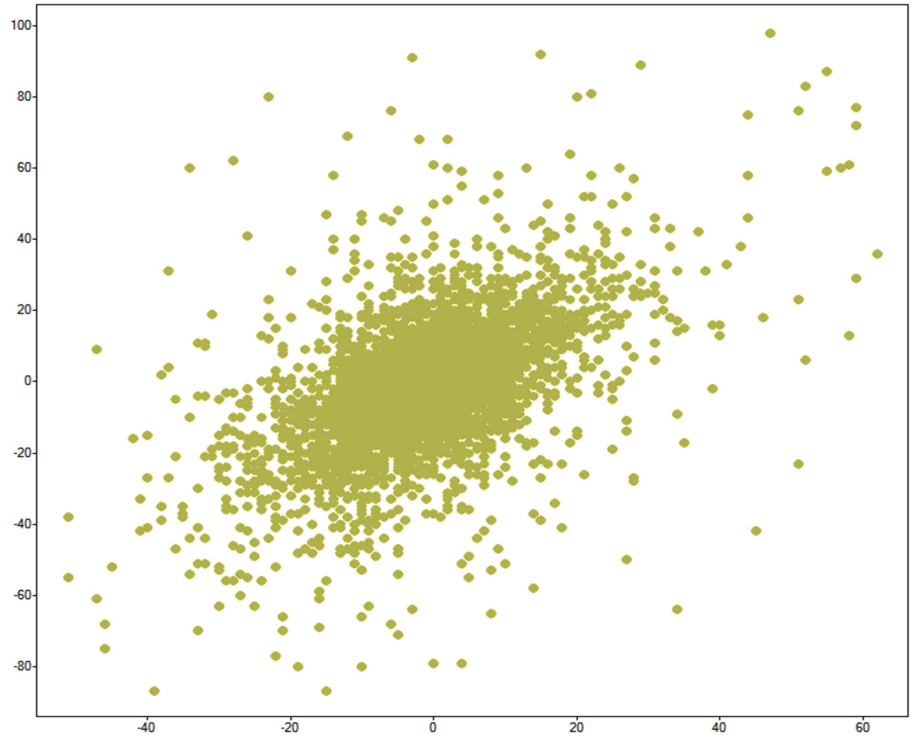


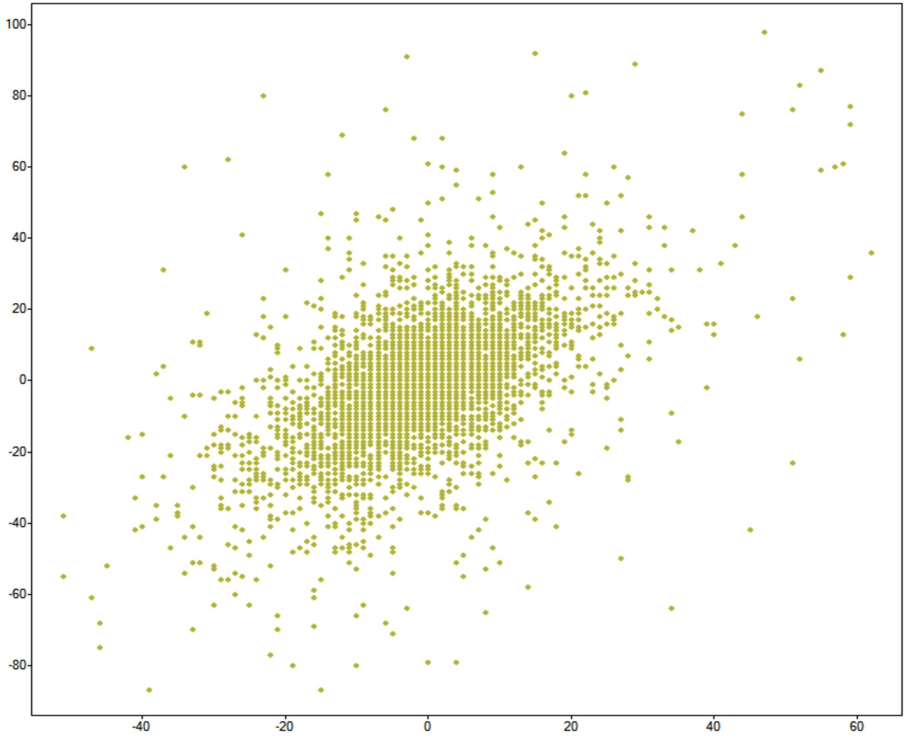
Over-plotting reduction

In some graphs, especially those that use points or lines, objects can overlap, which does not allow to distinguish individual values and makes analysis difficult. This problem is called **over-plotting**. We will take a look at some alternatives to eliminate or at least reduce it.

Reducing the size of data objects

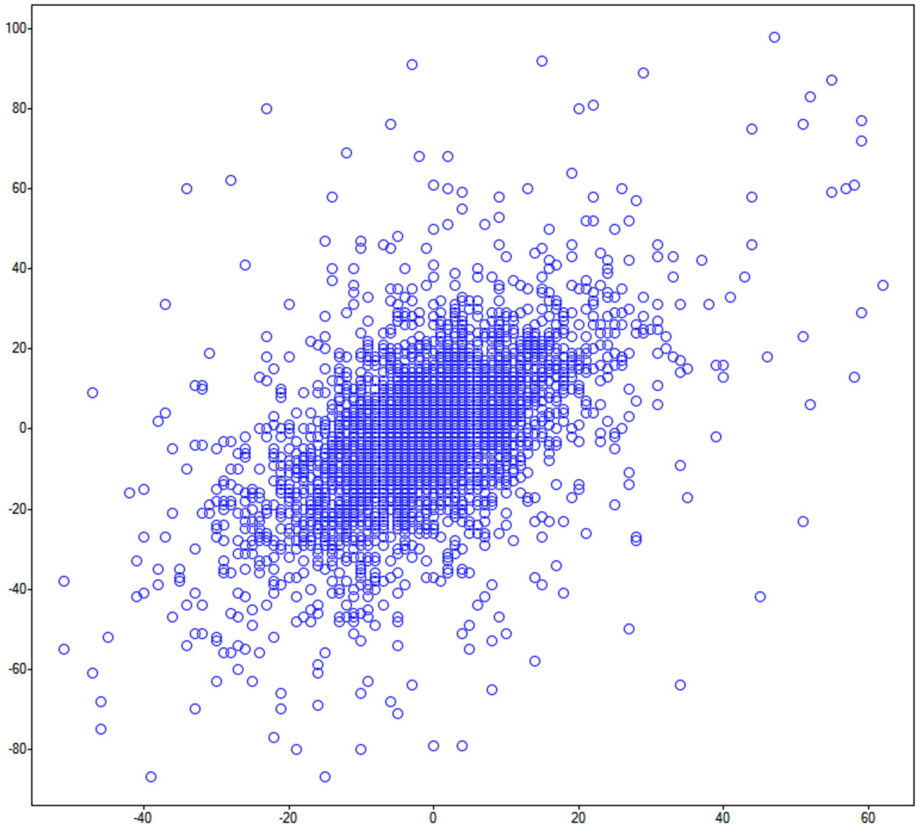
Over-plotting reduction





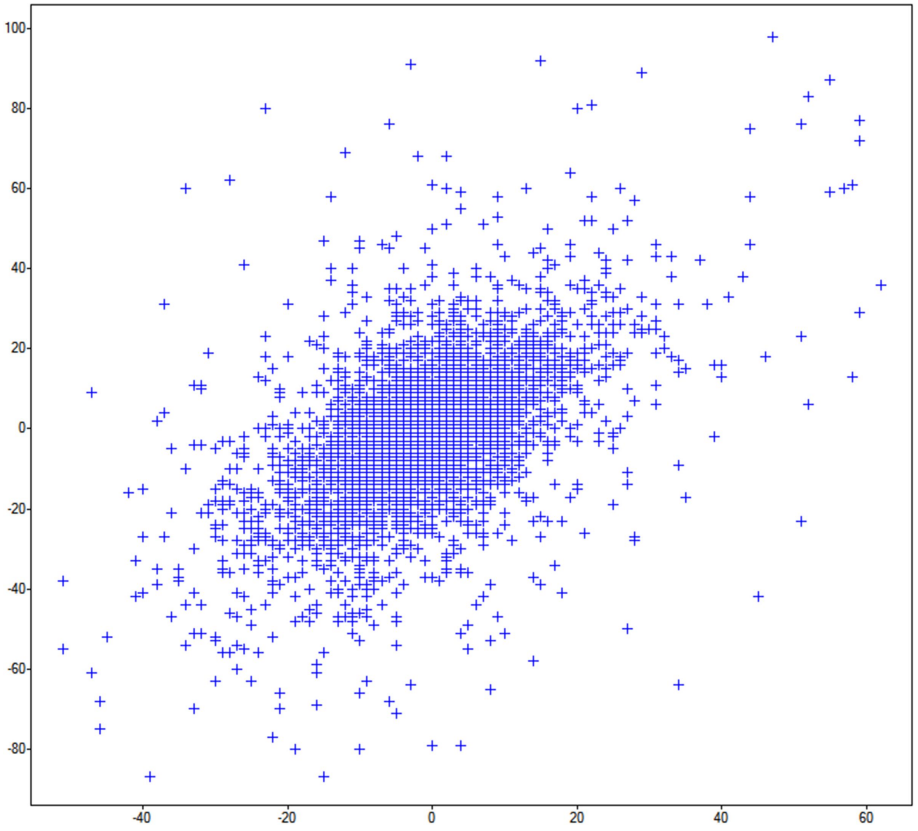
Removing fill color from data objects

Over-plotting reduction



Changing the shape of data objects

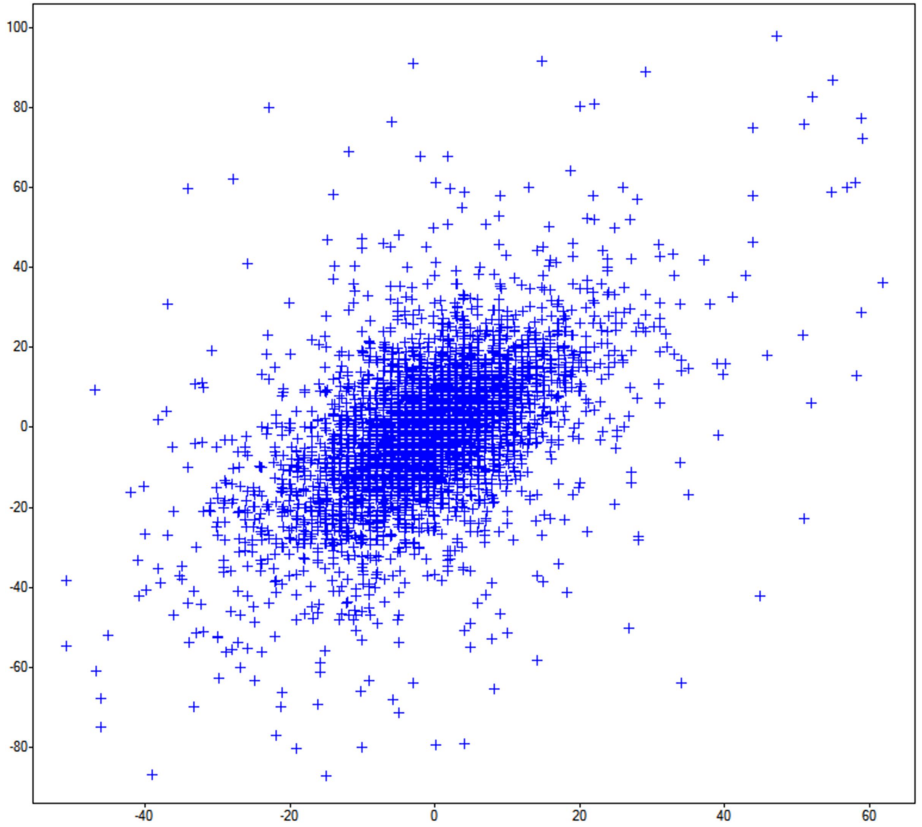
Over-plotting reduction



Jittering

Over-plotting reduction

It consists of changing the actual values, moving the points to slightly different positions.

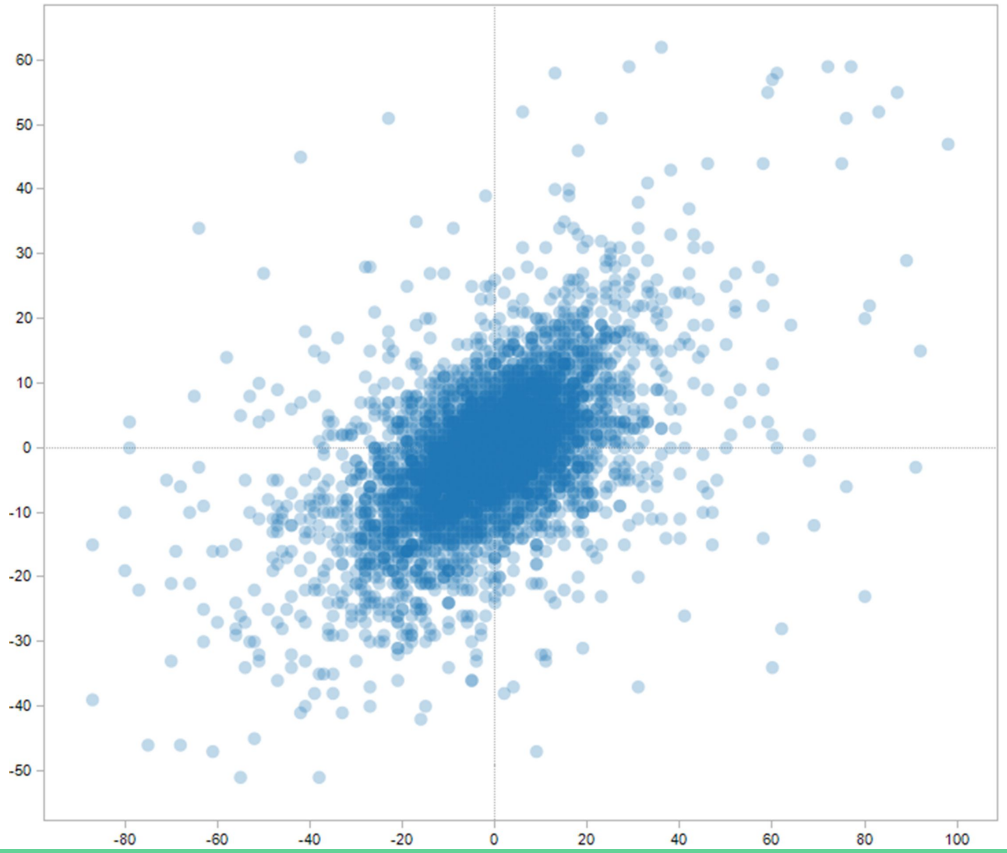


Making data objects transparent

Over-plotting reduction

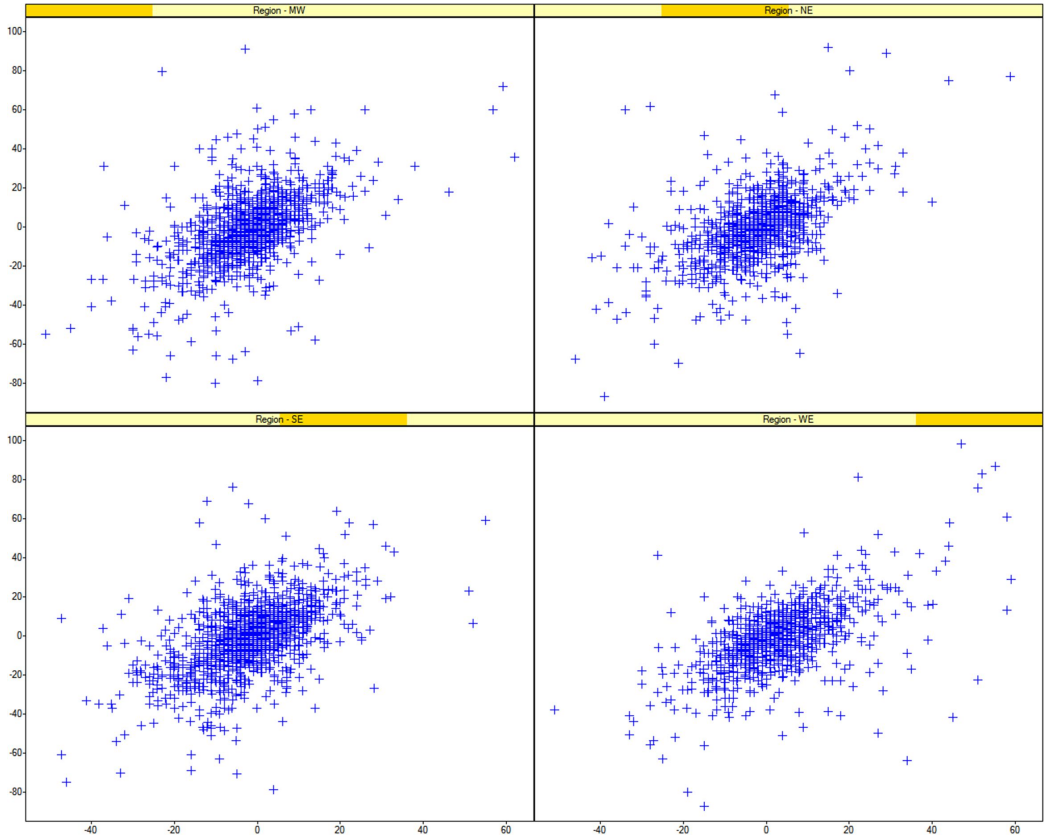
Giving a color with a degree of transparency to the points helps us to notice differences in the overlap as variations in the intensity of the color.

We can also use a slider to vary the degree of transparency.

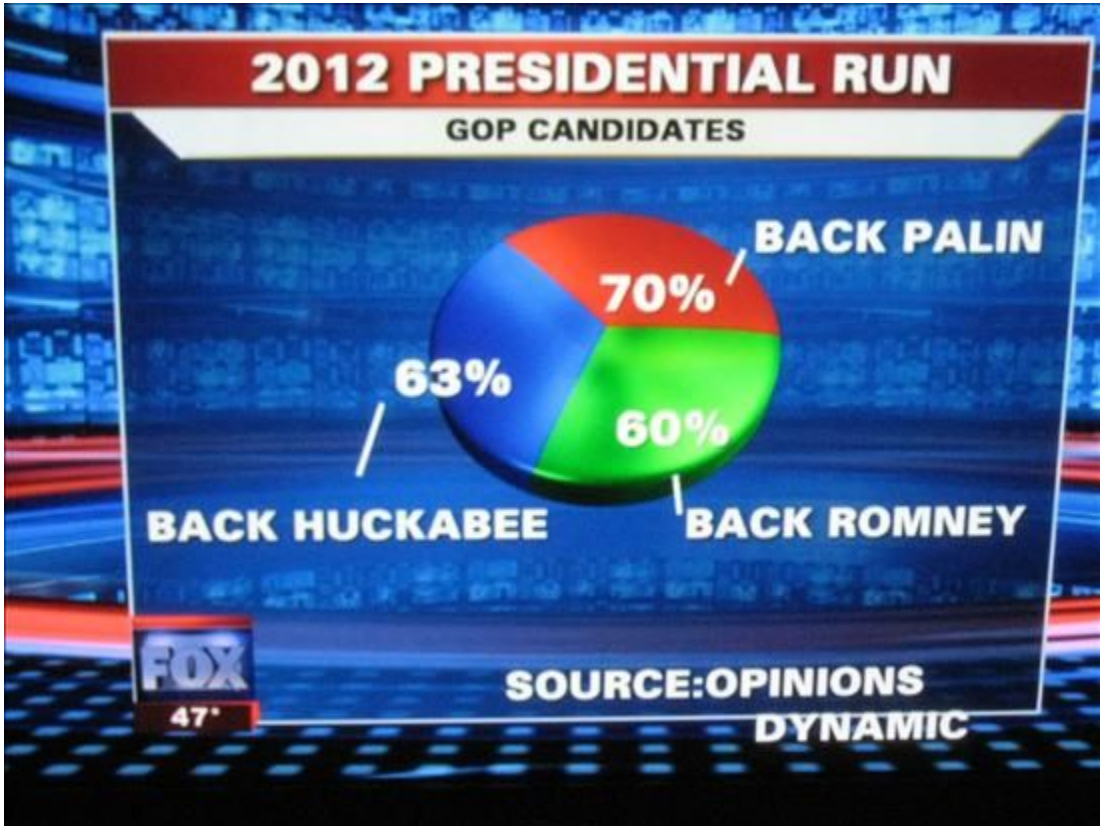


Reducing the number of values

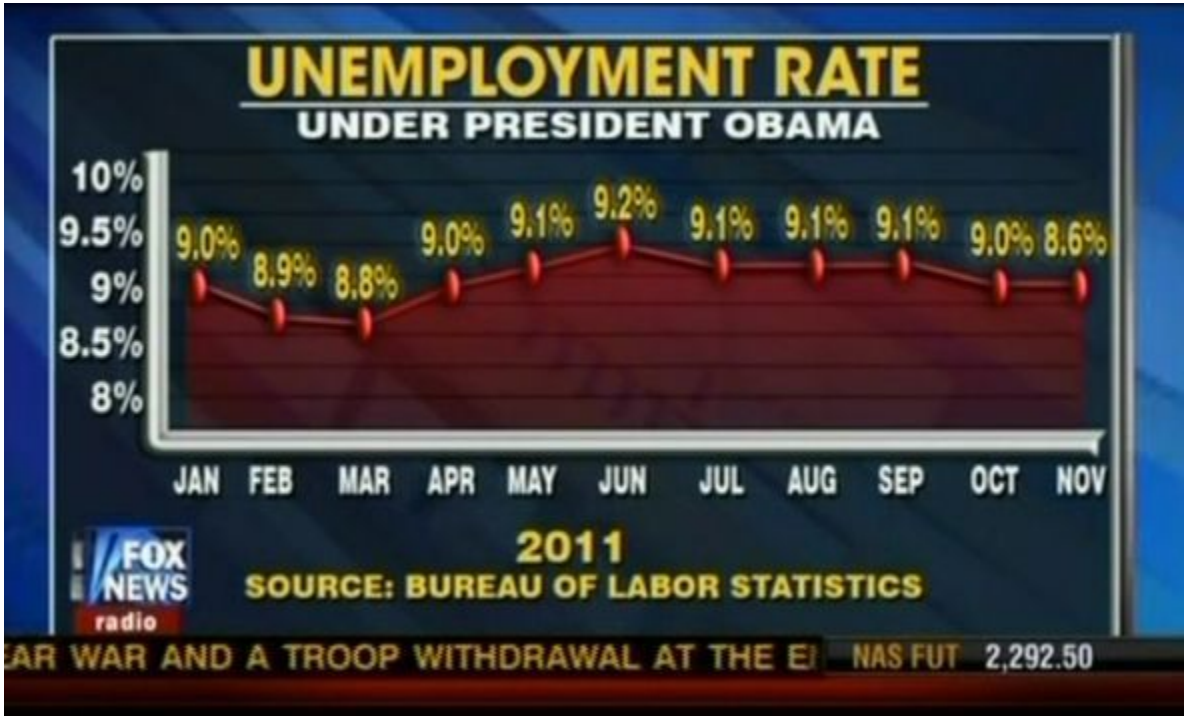
The same information from the previous charts was divided into four charts, one for each region.



Exercise


2012 PRESIDENTIAL RUN
GOP CANDIDATES
BACK PALIN
70%
63%
60%
BACK HUCKABEE
BACK ROMNEY
FOX
47'
SOURCE:OPINIONS DYNAMIC

Exercise

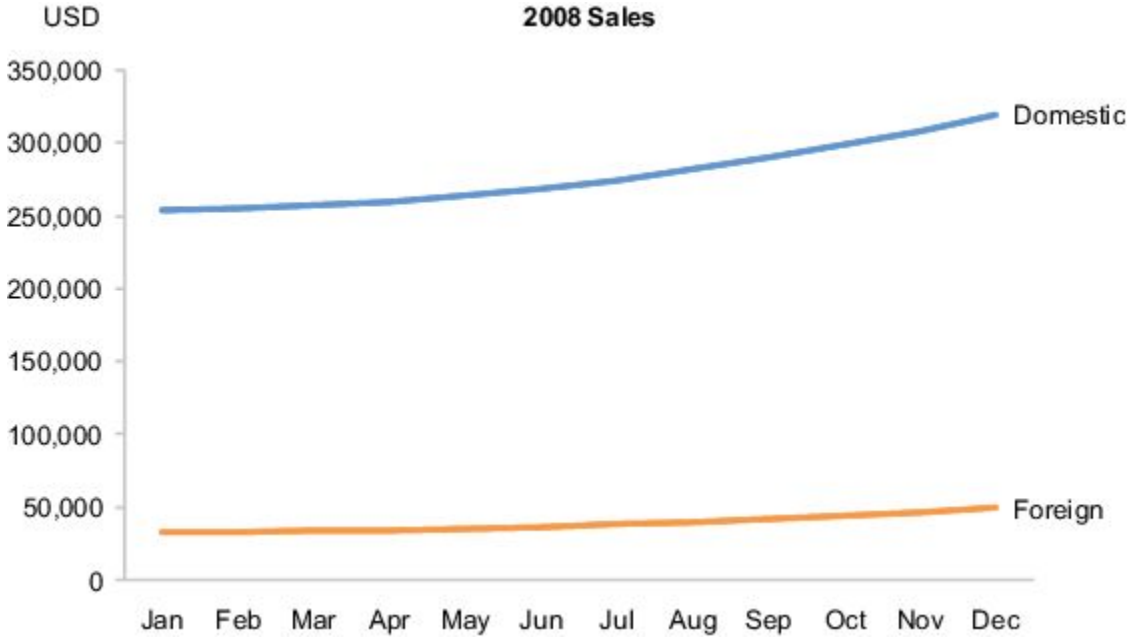

UNEMPLOYMENT RATE
UNDER PRESIDENT OBAMA
10%
9.5%
9%
8.5%
8%
9.0%
8.9%
8.8%
9.0%
9.1%
9.2%
9.1%
9.1%
9.1%
9.0%
8.6%
JAN
FEB
MAR
APR
MAY
JUN
JUL
AUG
SEP
OCT
NOV
2011
SOURCE: BUREAU OF LABOR STATISTICS
FOX NEWS radio
AR WAR AND A TROOP WITHDRAWAL AT THE E
NAS FUT 2,292.50

Honing skills for diverse types of visual analysis

Time series

Rate of change

We see that domestic sales have considerably greater value. But which of the regions shows the greatest growth?



Rate of change

Here we see the same data, but as a rate of change (percentage) from one month to the next.

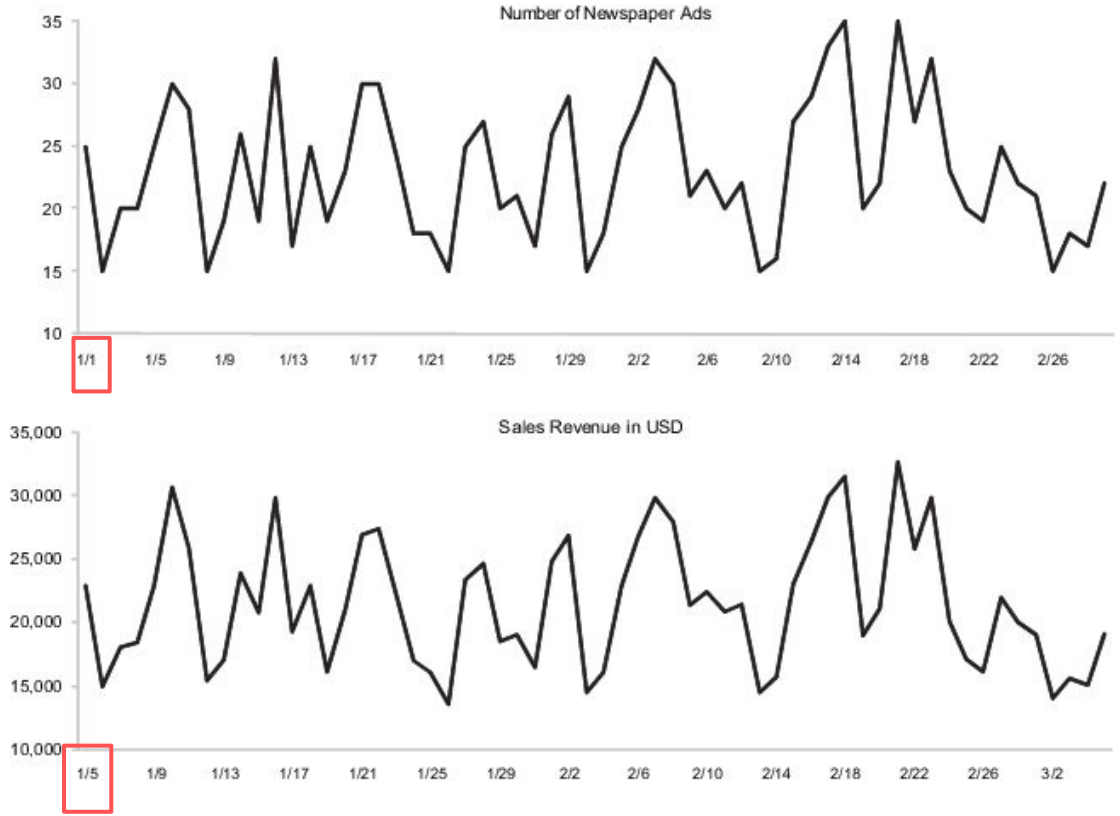
Foreign sales may have better potential.

Co-variation

When changes in one time series happen before or after the related changes in another series, we have leading indicators or lagging indicators.

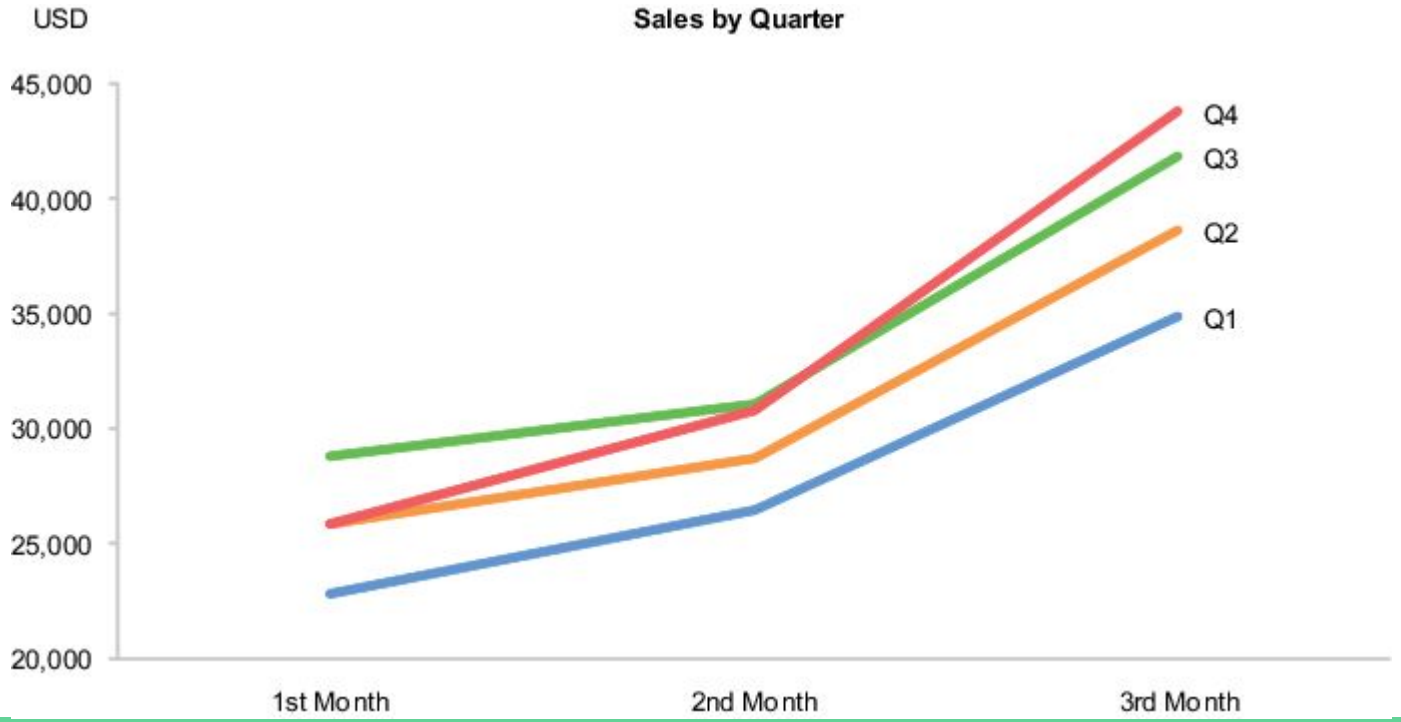
Co-variation

Correlation between newspaper ads (leading indicator) and sales (lagging indicator).



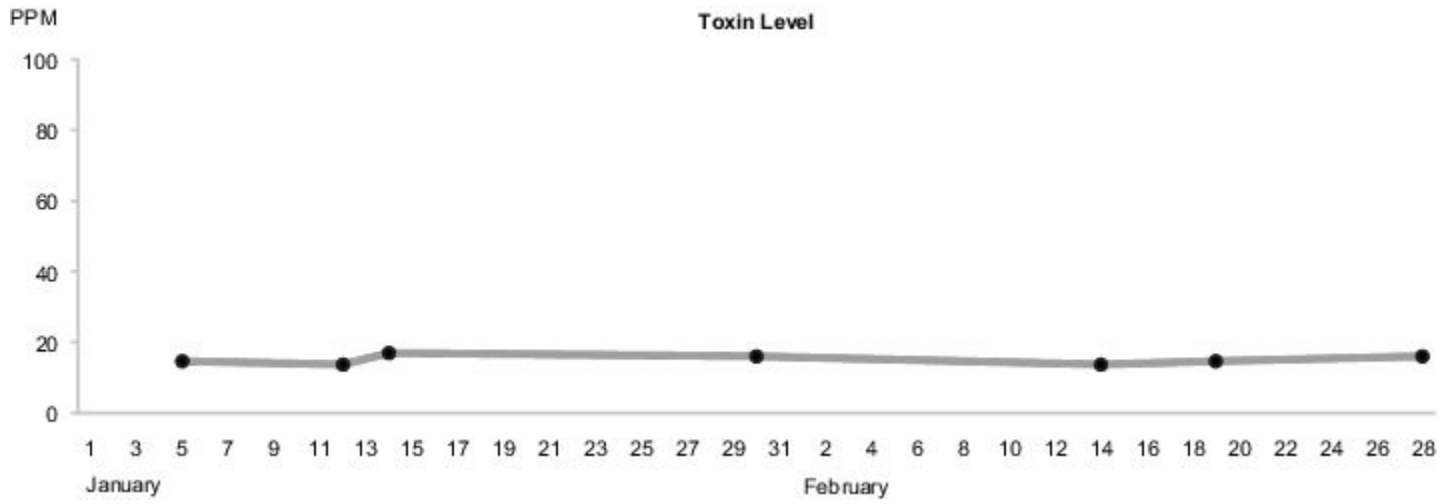
Cycles

Here we show 1 year as 4 quarters and not linearly.



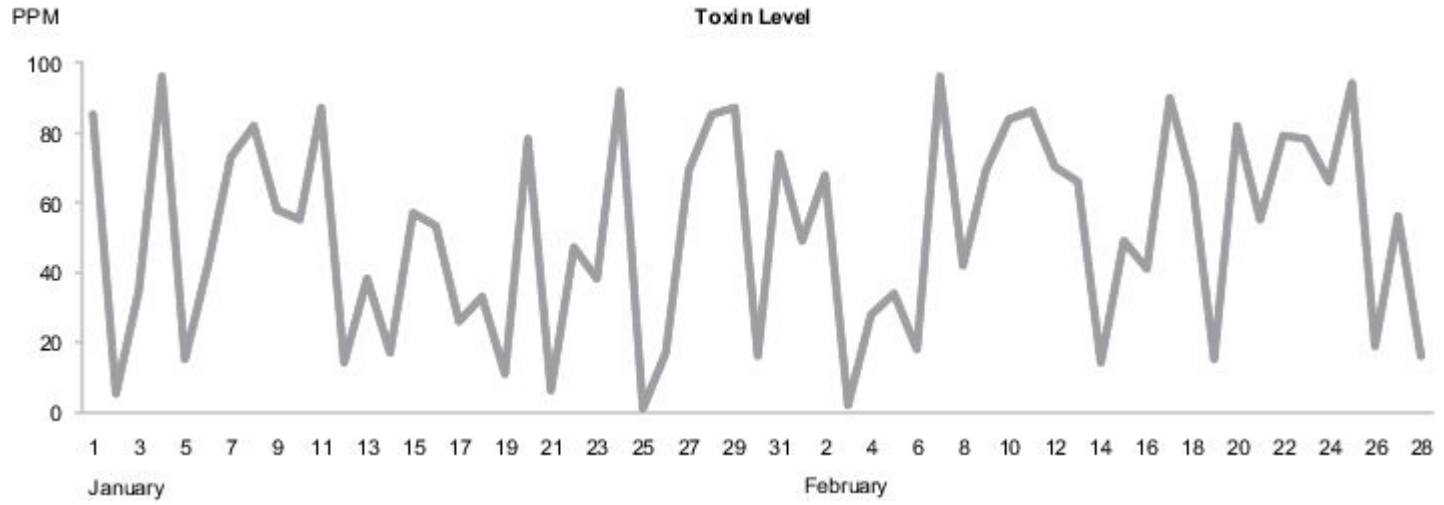
Dot plots for analyzing irregular intervals

Observe the time series. Is this chart appropriate?



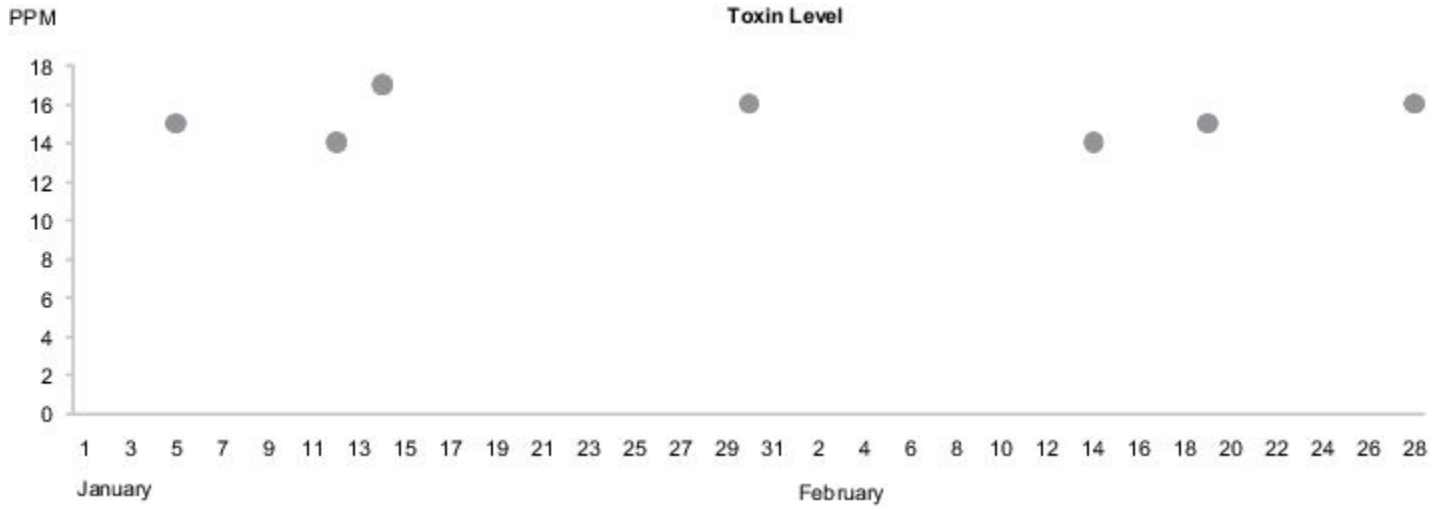
Dot plots for analyzing irregular intervals

But the data collected daily shows the following profile.



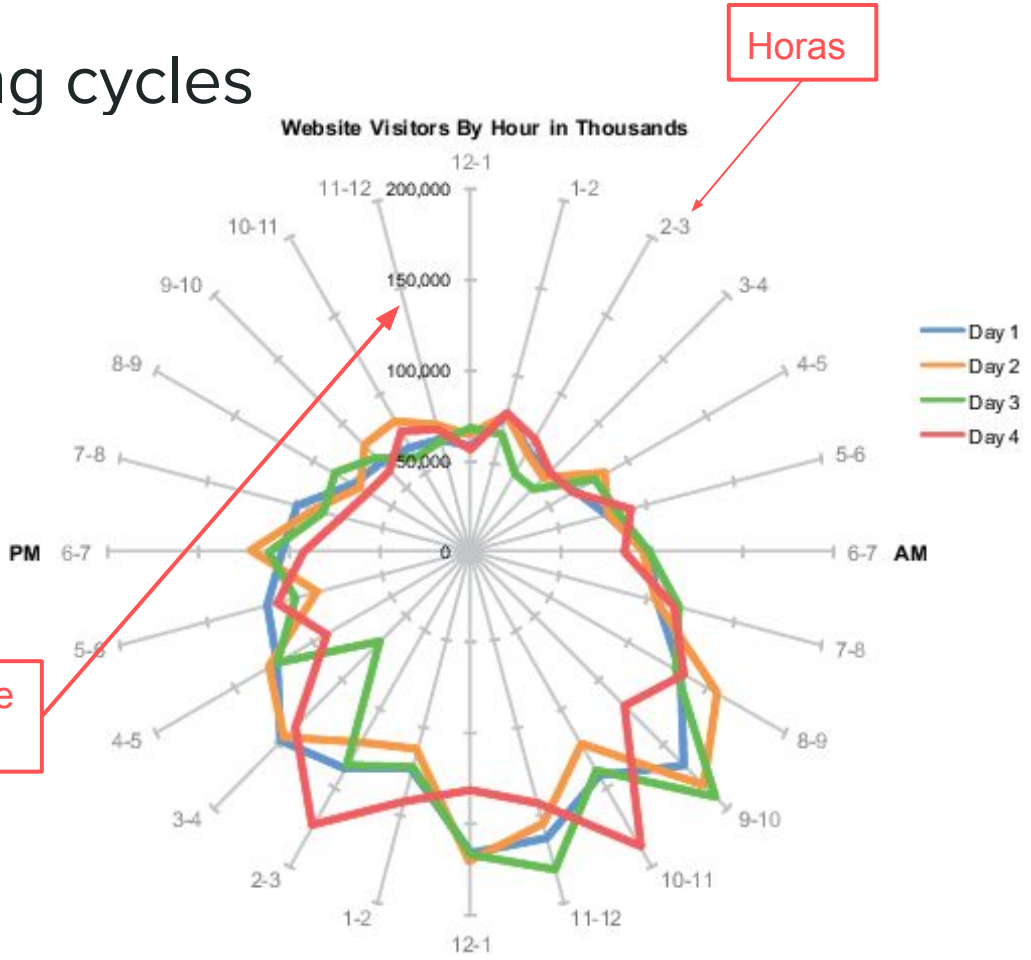
Dot plots for analyzing irregular intervals

When we have irregular intervals, we must not connect the dots. The points do not encourage us to “see” a direct transition between them.



Radar graphs for comparing cycles

The circular shape of radar graphs or spider graphs can be used to represent the cyclical nature of time.

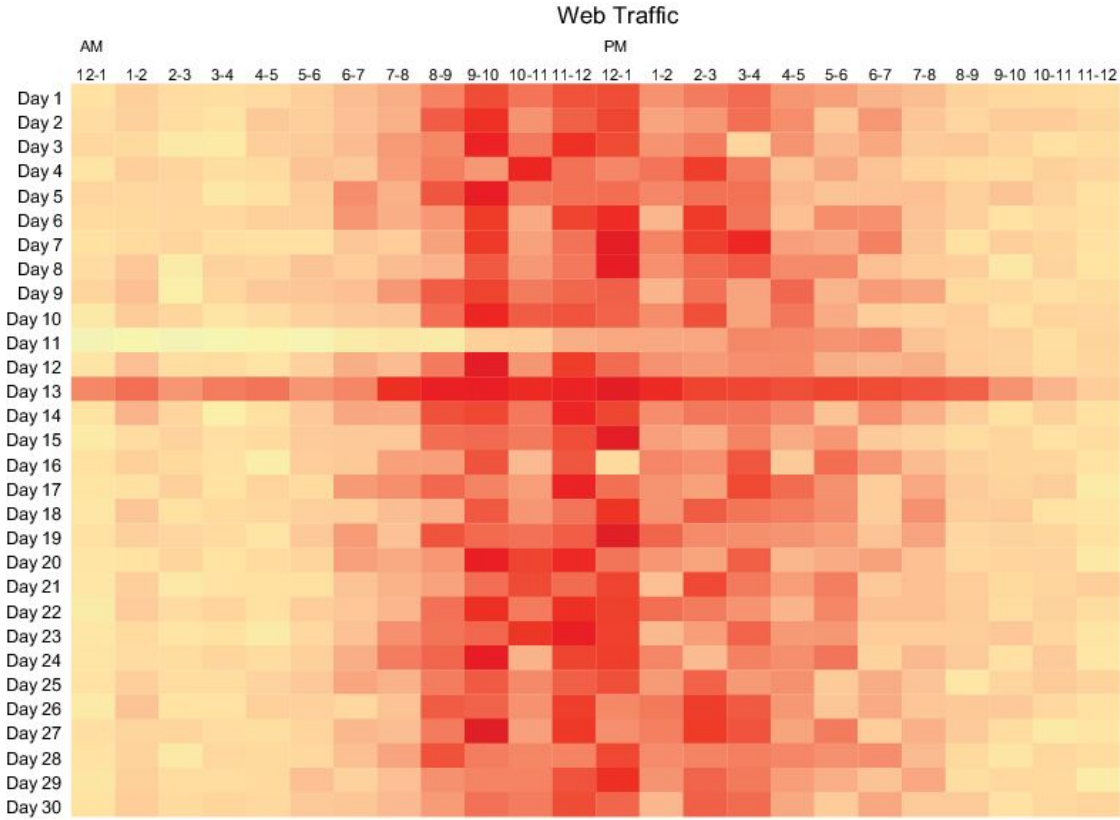


Radar graphs for comparing cycles

The same data can be represented by a line graph.

Heatmaps for analysing high-volume cyclical patterns and exceptions

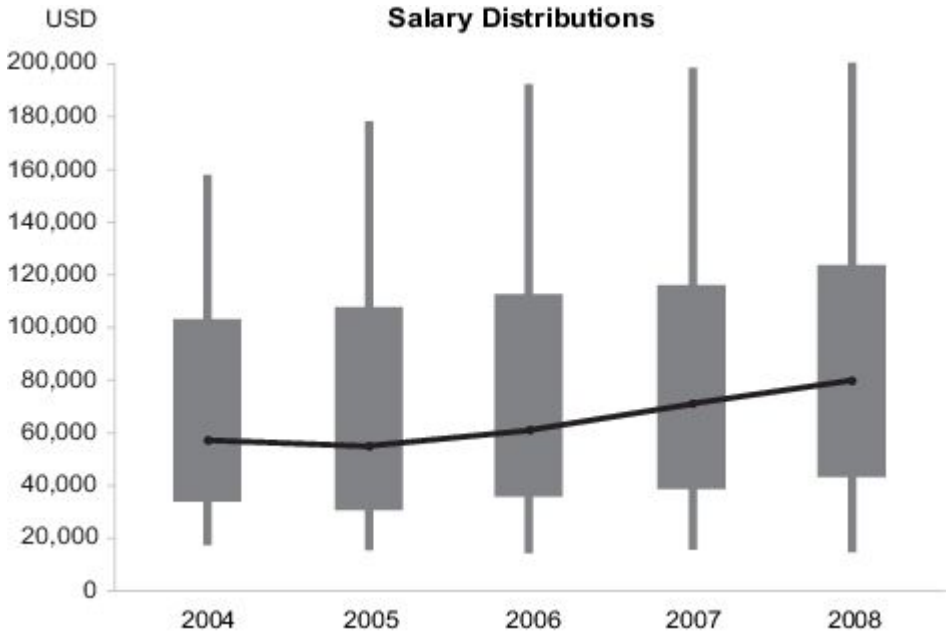
In this example we show web traffic as the number of visits to a website over the course of the day, during 30 days.



Heatmaps for analysing high-volume cyclical patterns and exceptions

The heatmap should be used to represent time series when it can display cyclical data that cannot be clearly displayed using line graph or radar due to over-plotting.

Box plots for analyzing distribution changes over time



Animated scatterplots for analyzing correlation changes

In 2006, social scientist Hans Rosling of GapMinder had 20 minutes at the Technology, Entertainment, and Design (TED) conference in California to tell the story of the relationship between fertility rate and life expectancy over the years.

And he told this story in a very interesting way.

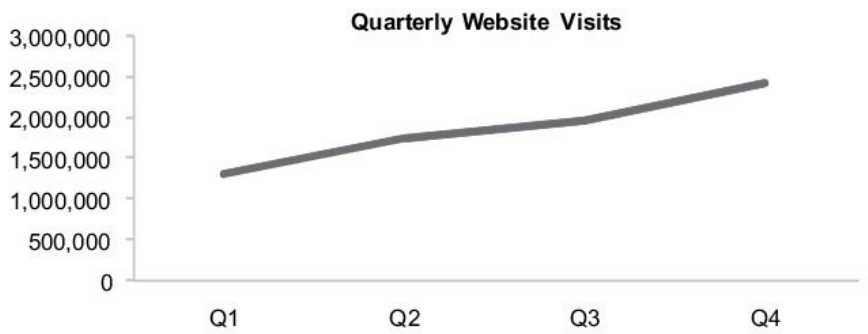
Animated scatterplots for analyzing correlation changes

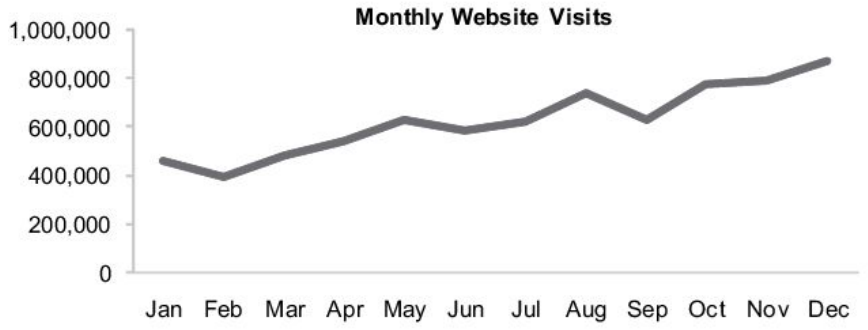
https://www.youtube.com/watch?v=FOSBZxNEBdE

Aggregating to various time intervals

Don't be restricted to just one level of aggregation when working with time series.

A good alternative can be a slider that controls the time interval.

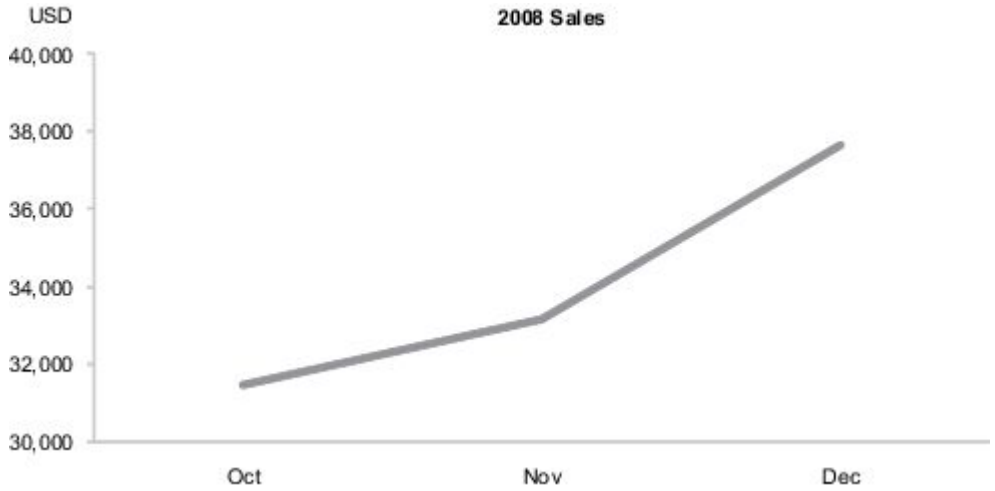




Viewing time periods in context

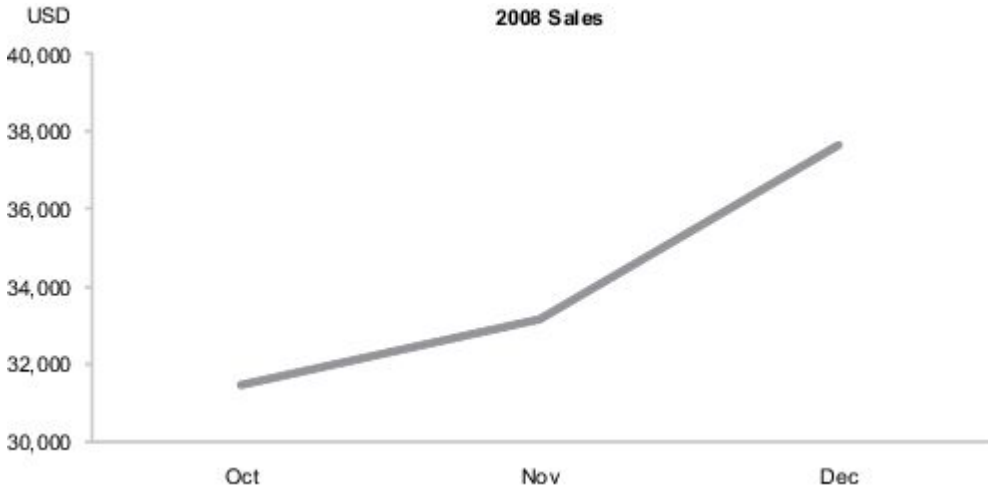
If the analysis of the time series is restricted to a very short time interval, we can draw incorrect general conclusions, as we do not see the data in a more general context.

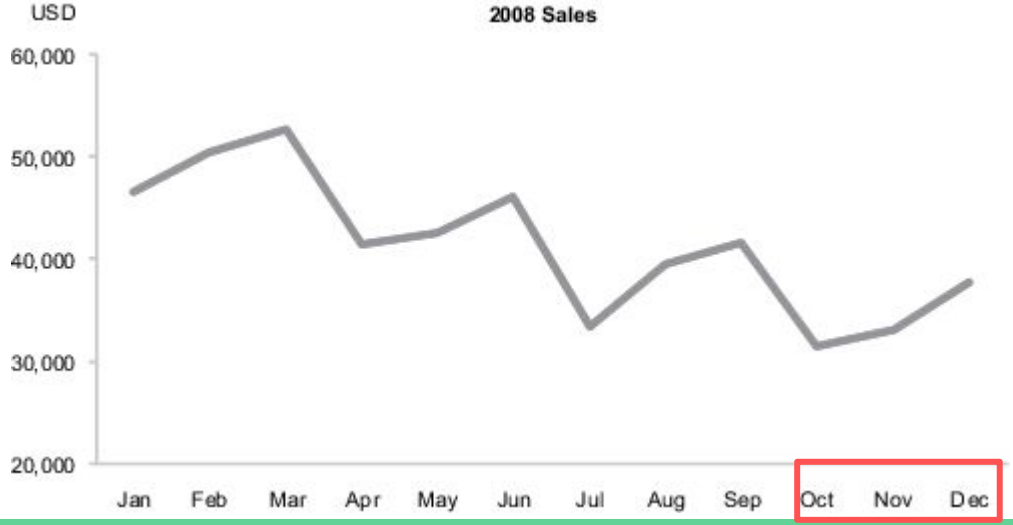
What conclusion can you draw from this graph?



Viewing time periods in context

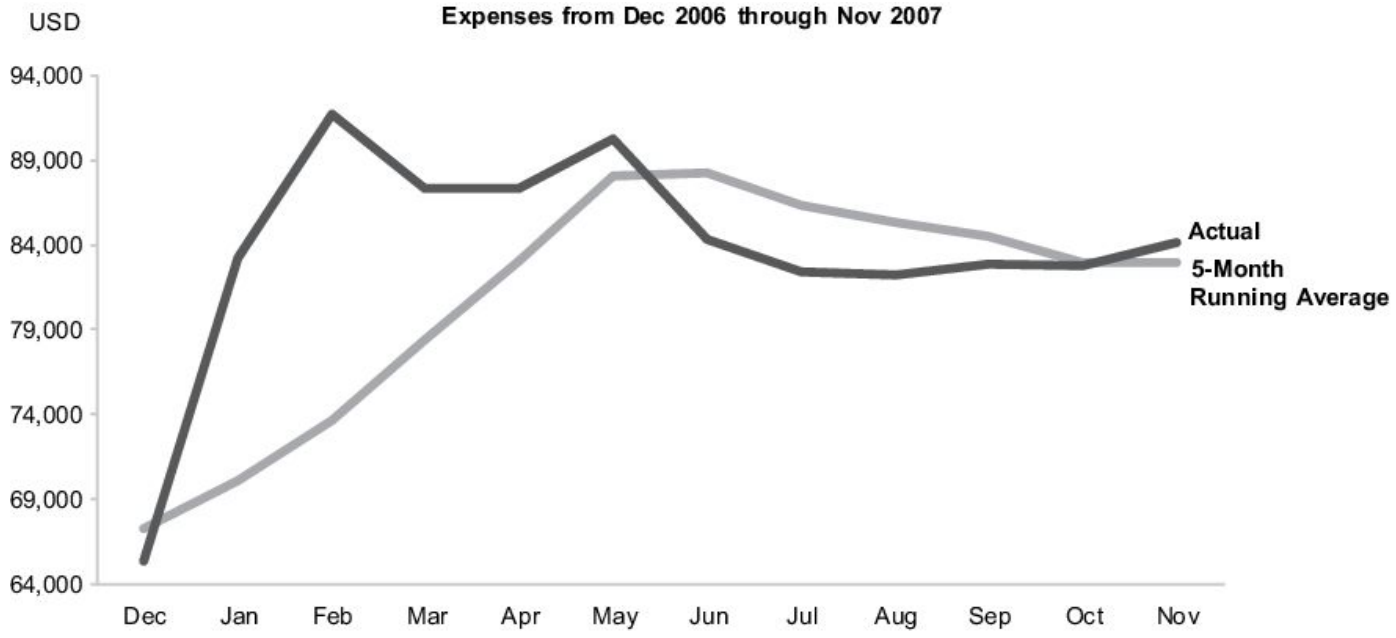
And now, what conclusion can you draw?





Using running averages to enhance perception of high-level patterns

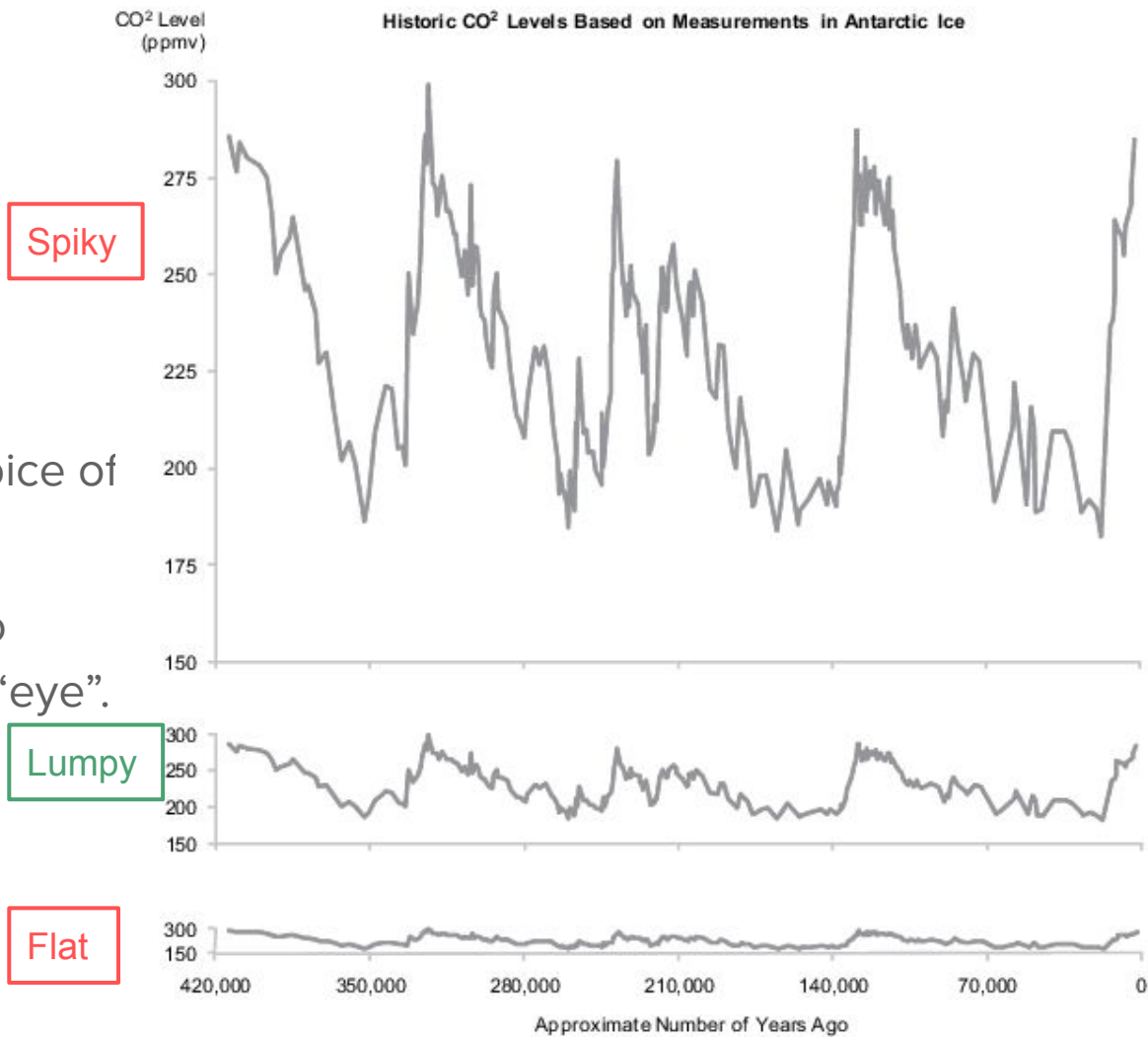
We can also examine time series from two perspectives: smoothed (high level) and actual (low level). This helps us not to draw inappropriate conclusions.



Optimizing a graph's aspect ratio

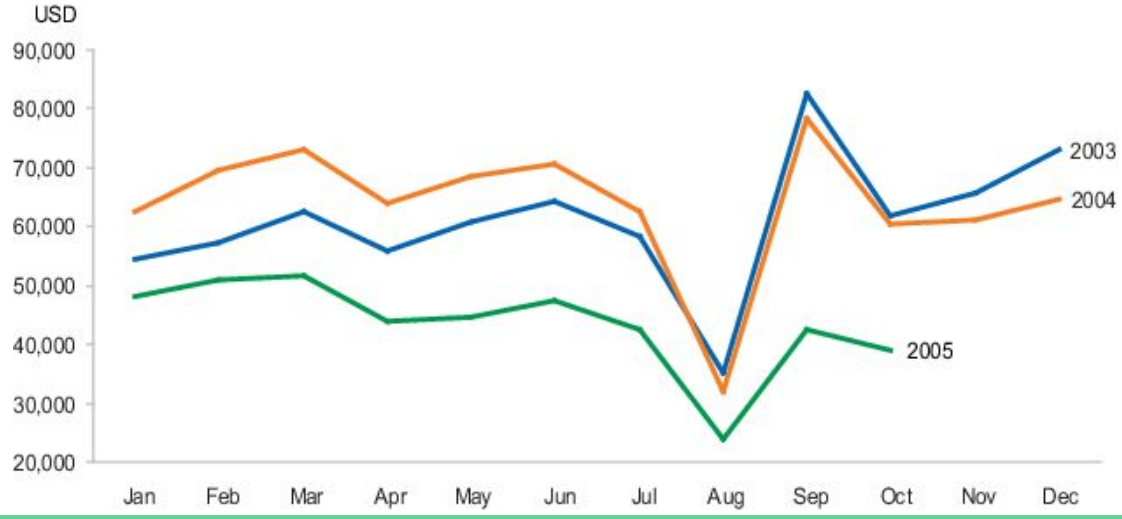
Some programs support the choice of 45 degrees, such as the R.

However, in general, we need to adjust the ratio manually, in the “eye”.



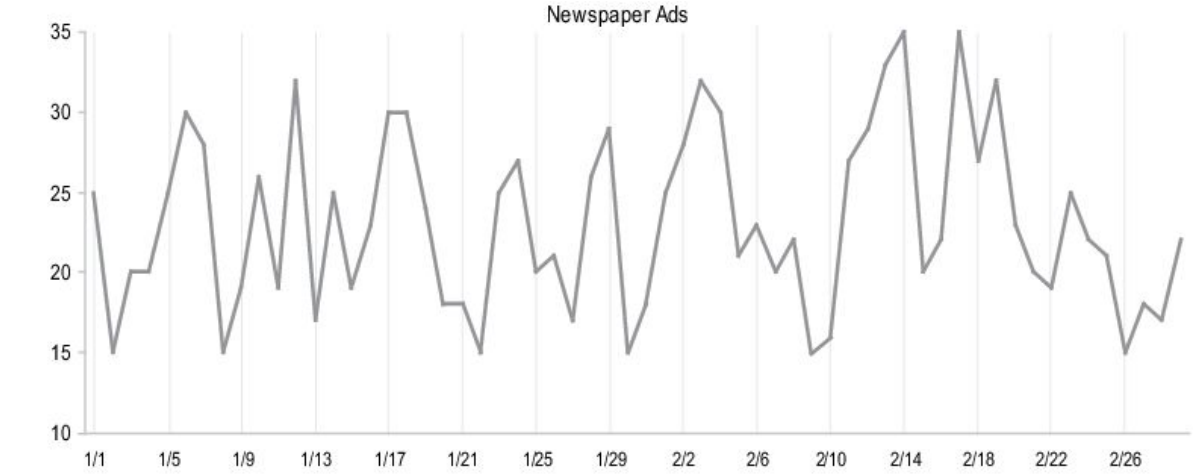
Overlapping time scales to compare cyclical patterns

Showing each cycle as a separate line helps us to detect cycle patterns.



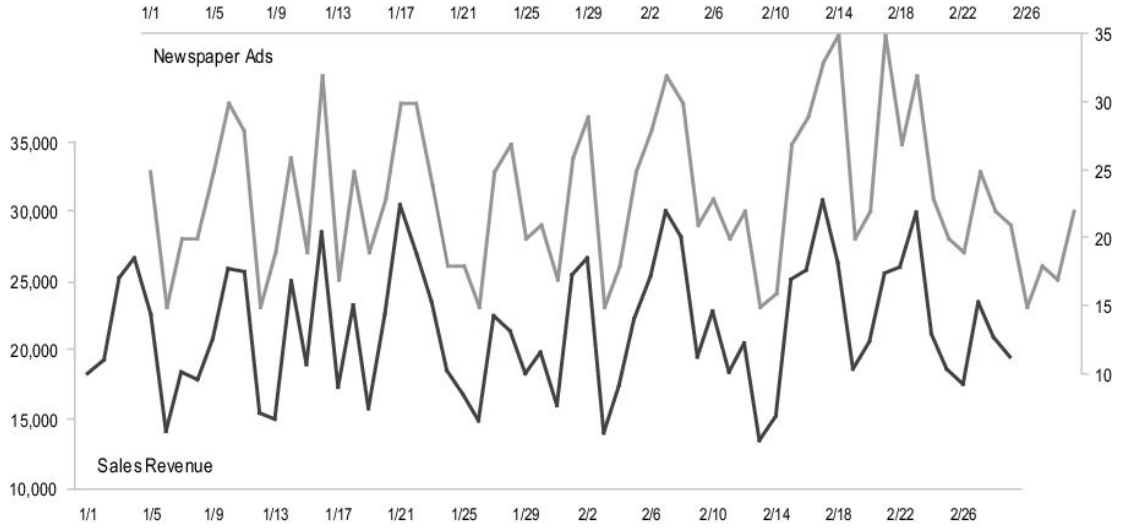
Shifting time to compare leading and lagging indicators

We would like to understand how one variable influences the other, which can be difficult when there is a time lag between the cause and the consequence.



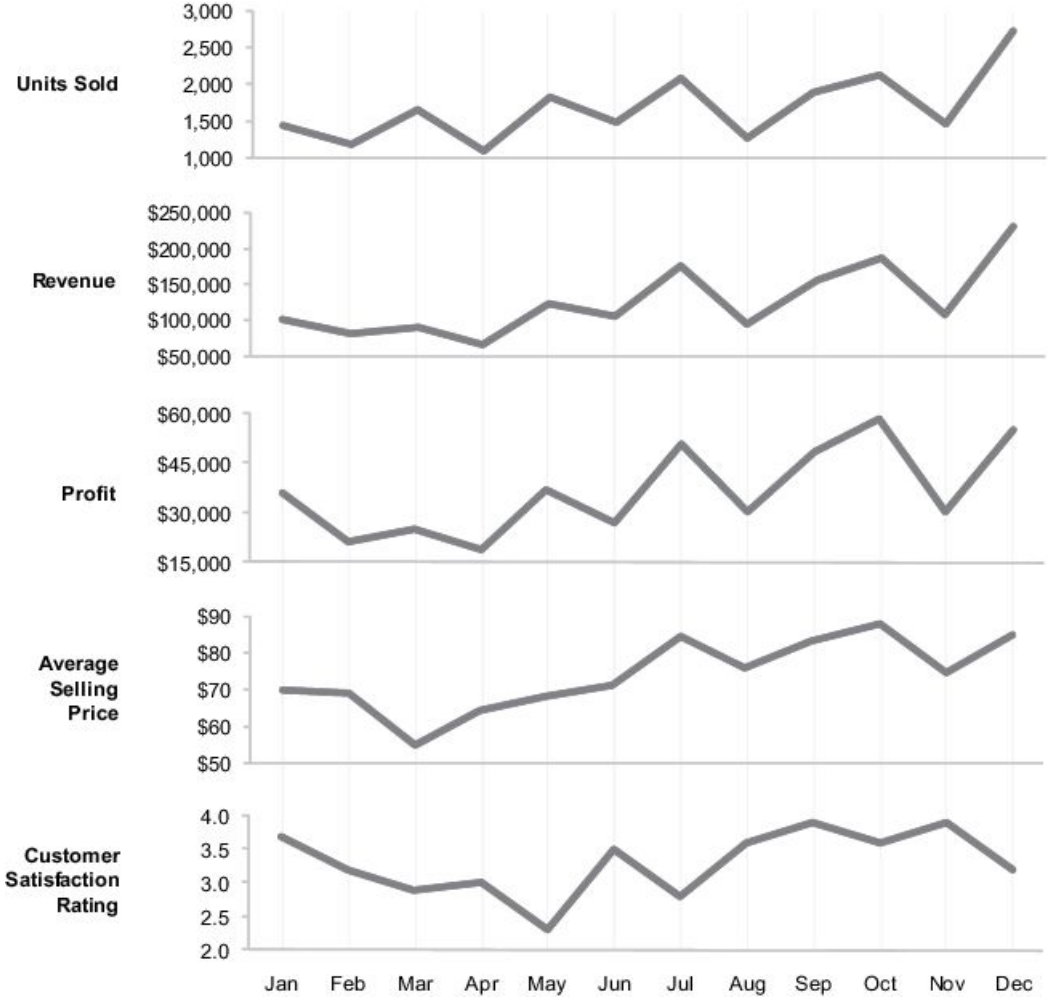
Shifting time to compare leading and lagging indicators

We can align the events to examine the relationship.



Stacking line graphs to compare multiple variables

Even though the data cannot be compared directly, we can compare patterns of change.

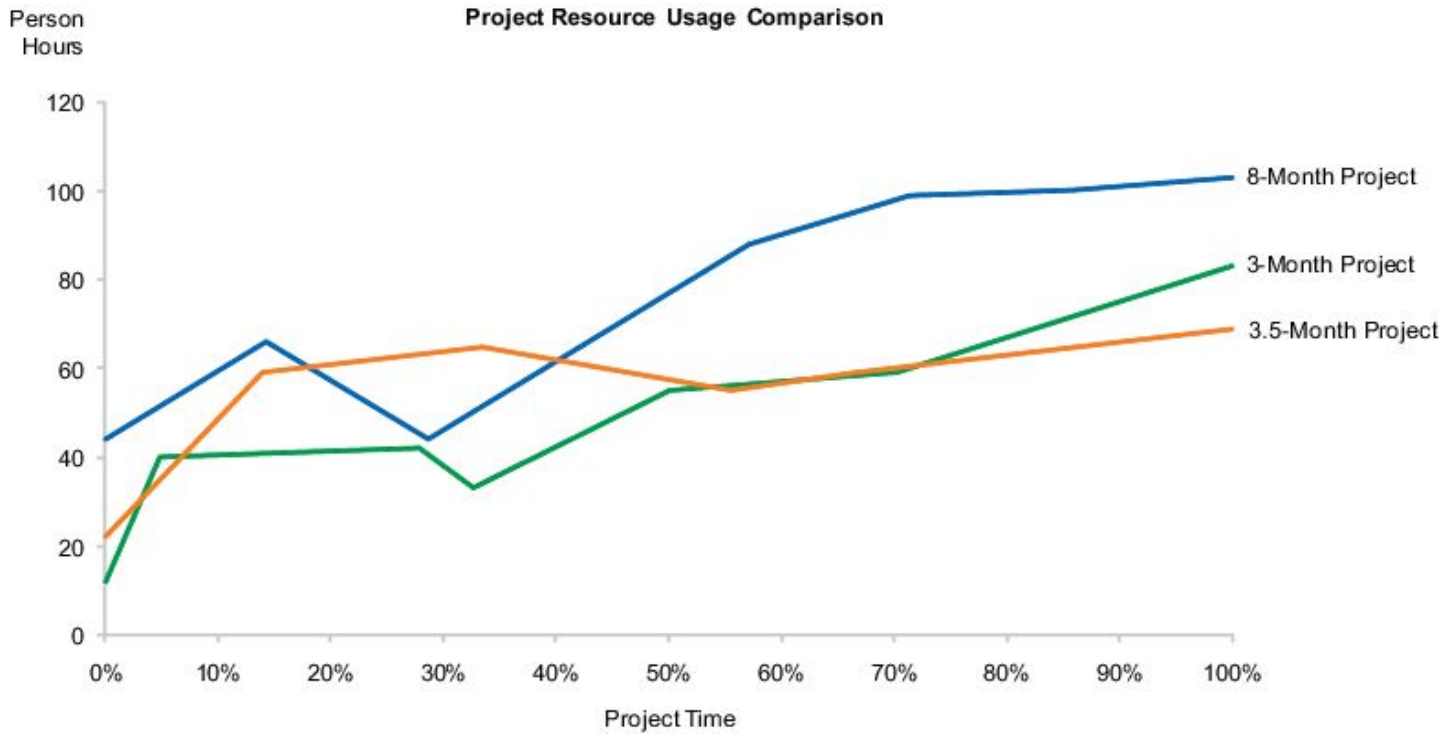


Expressing time as 0-100% to compare asynchronous processes

Imagine that an IT company wants to compare 50 projects over the past 5 years. They start at different times and have a different duration.

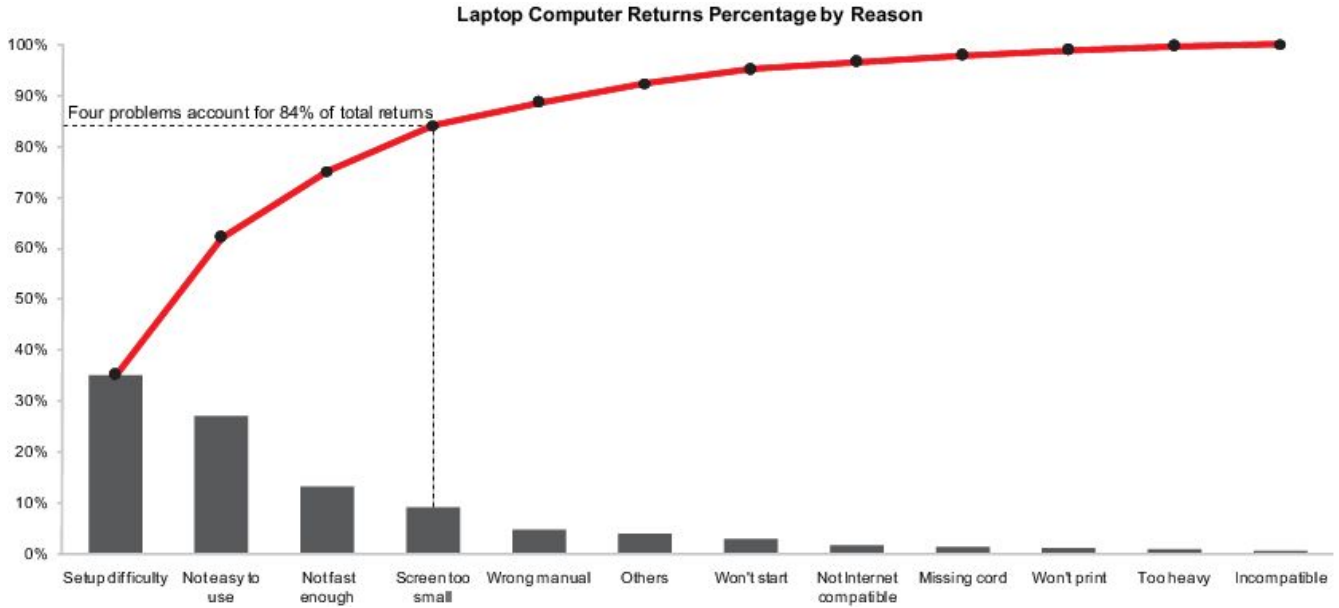
To make the start, end and duration consistent, we can express the duration as a percentage. So we can compare these asynchronous time series.

Expressing time as 0-100% to compare asynchronous processes



Part-to-whole and ranking

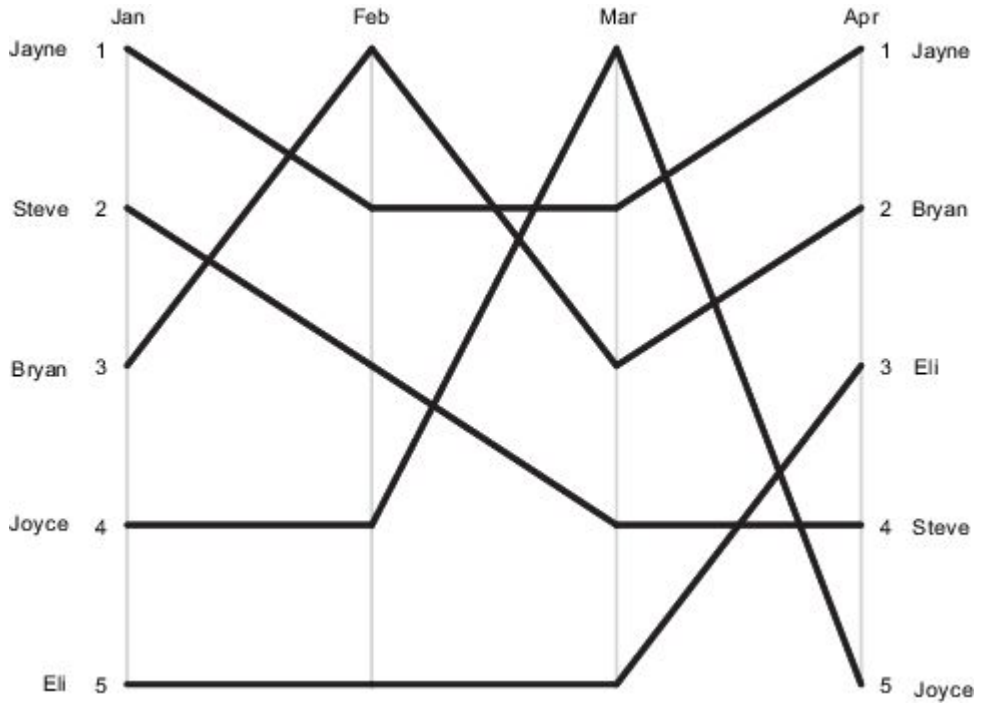
Pareto charts



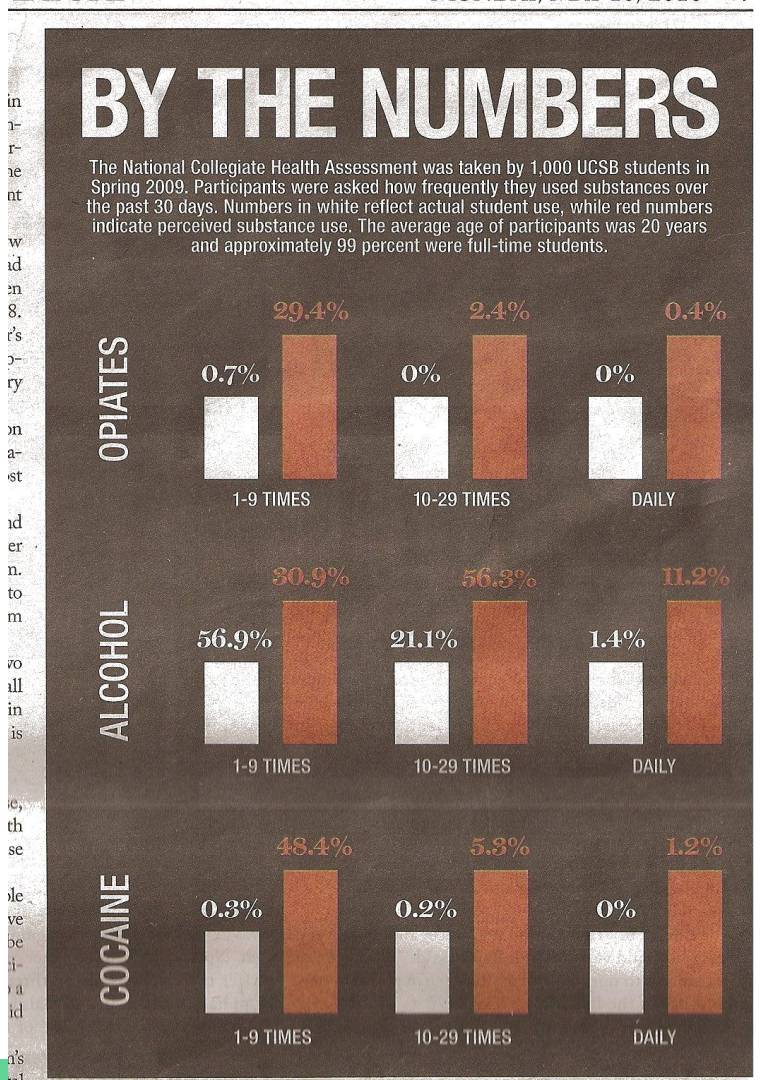
Using line graphs to view ranking changes through time

A line graph can be used to show salespeople sorted by sales performance.

In this case, only ranking changes are clear, not the values.



Exercise

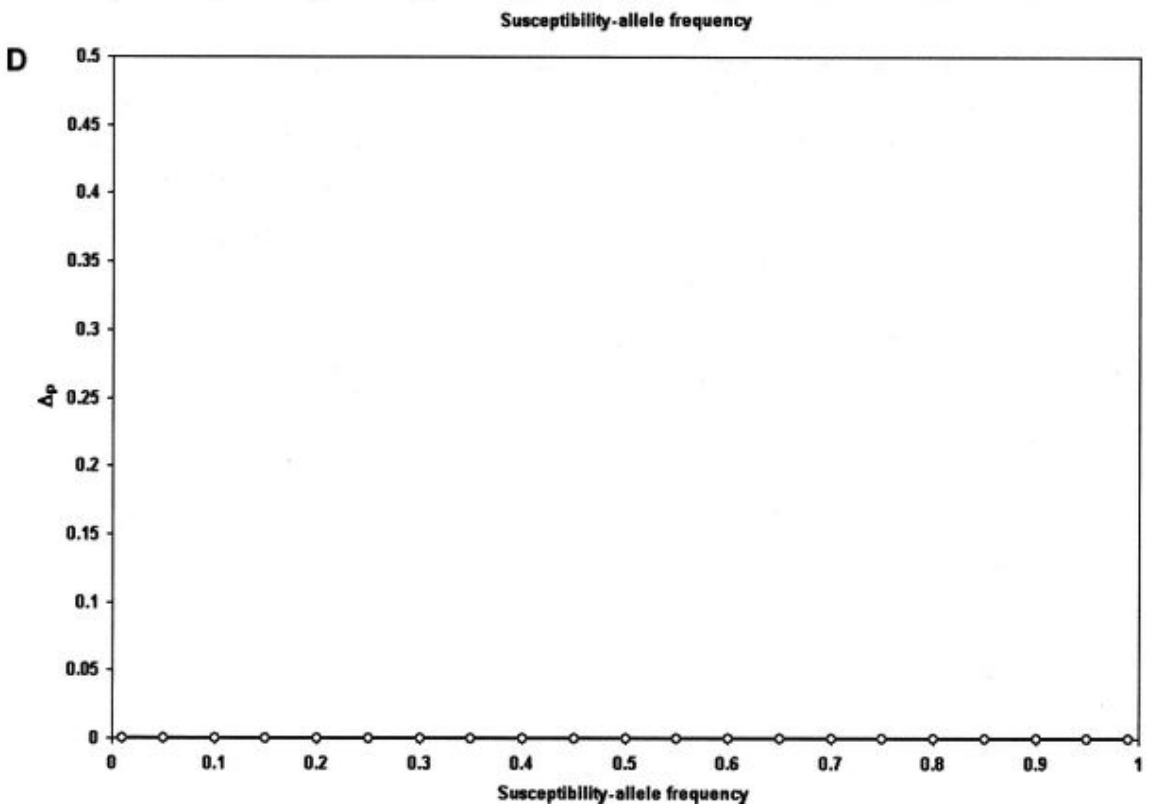


eagerEyes

Exercise

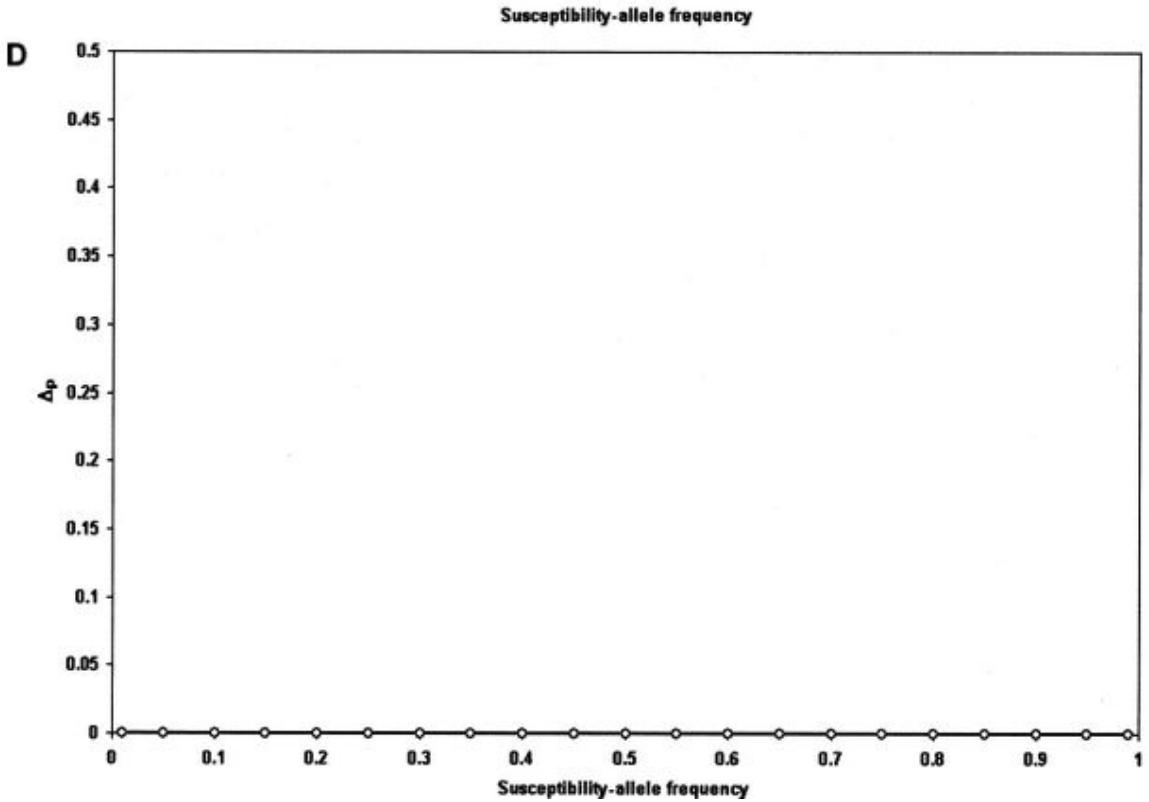
Título do artigo: Rational Inferences about Departures from Hardy-Weinberg Equilibrium

Qual o fator de impacto?



Exercise

Fator de impacto: 9.025



Deviation analysis

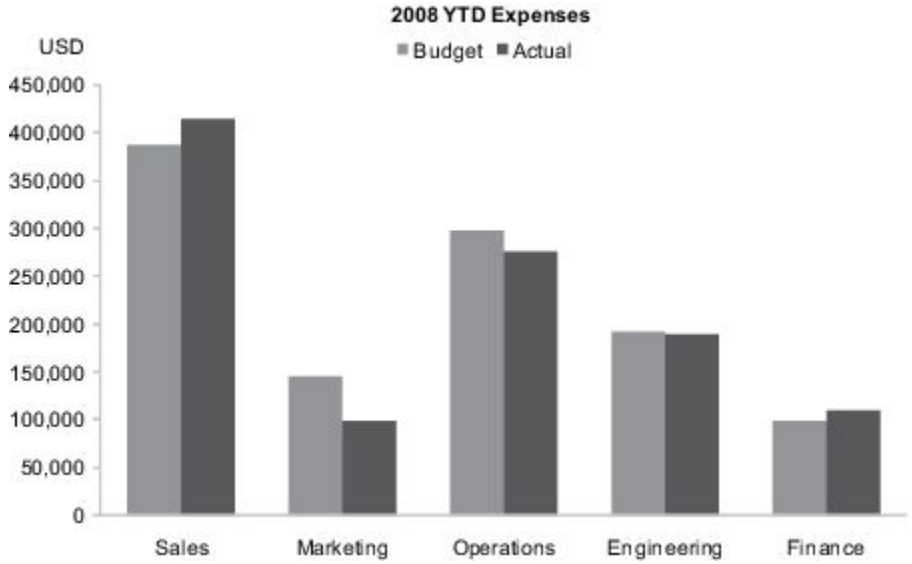
Deviation analysis

We call deviation analysis the comparison of a set of values with respect to a reference, which can be, for example, average, budget, a value in a past moment.

Deviation analysis

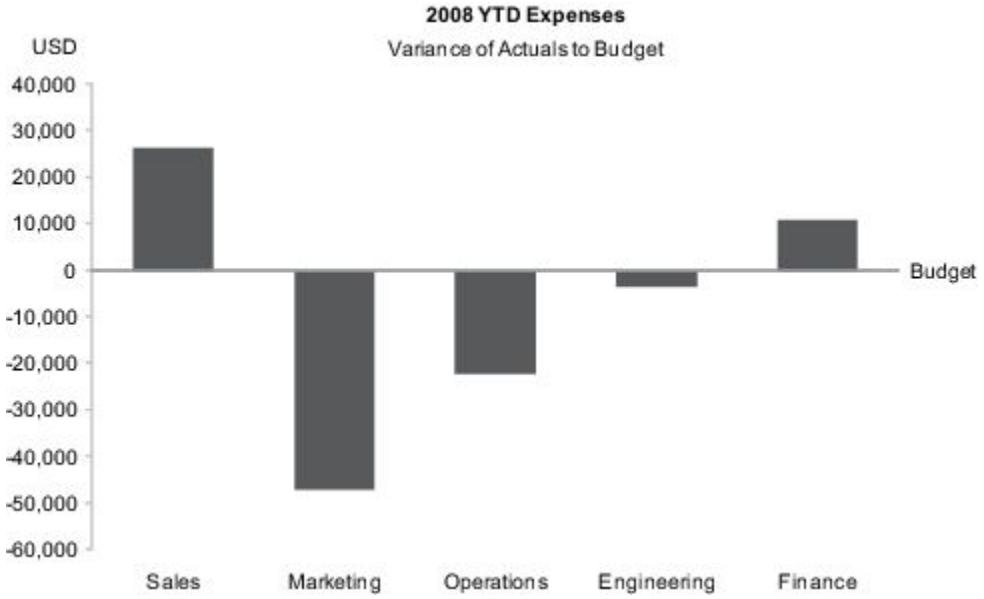
Let's look at a classic example, which is the comparison of actual expenses with a given budget.

How can we focus on differences?



Deviation analysis

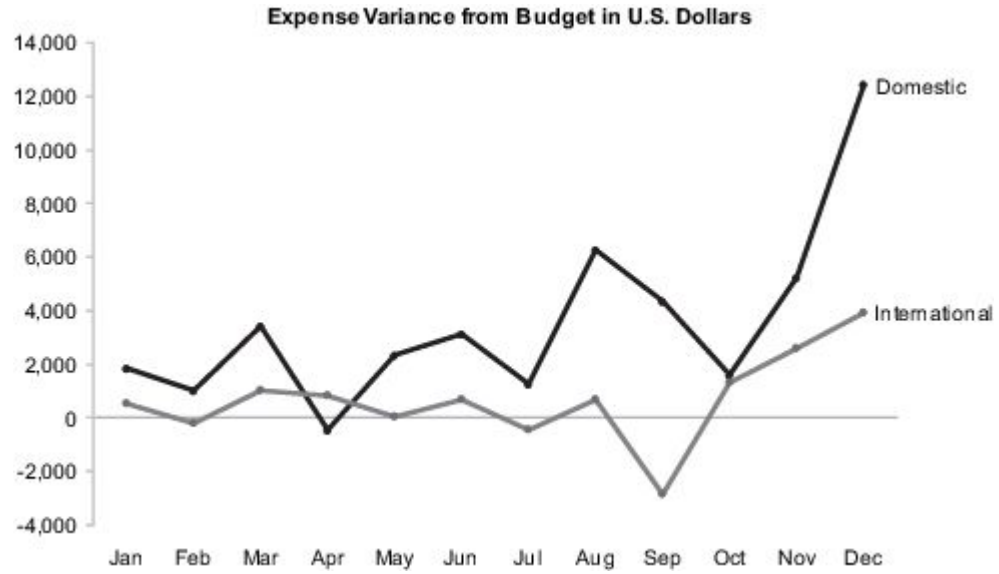
Here we show directly the difference between the real and the budget.



Expressing deviations as percentages

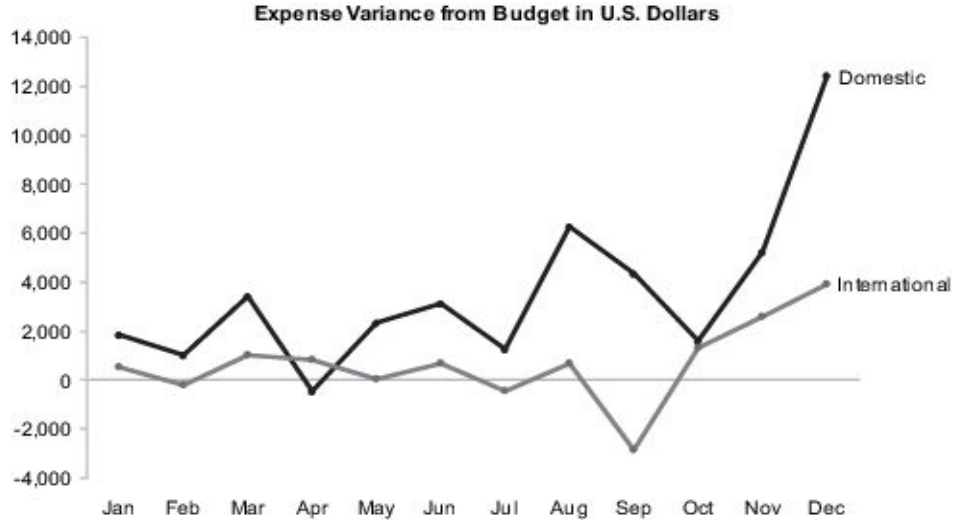
In this example, the deviations between actual expenses and budget are expressed as percentage.

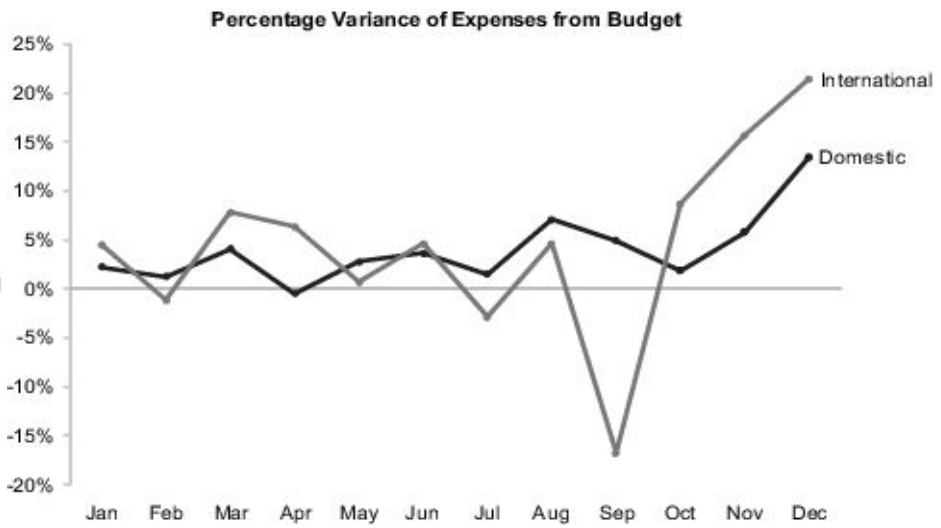
Domestic and international expenses exceed the budget, but domestic is much higher.



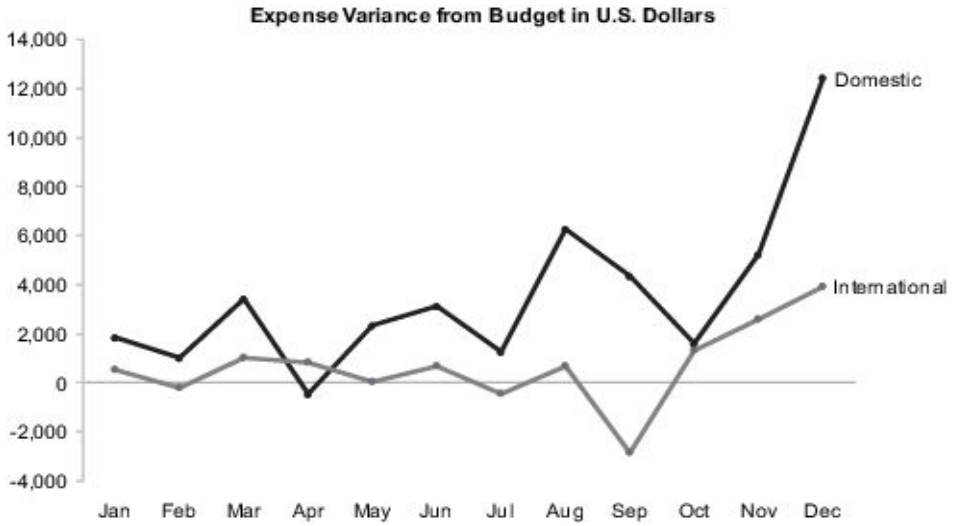
Expressing deviations as percentages

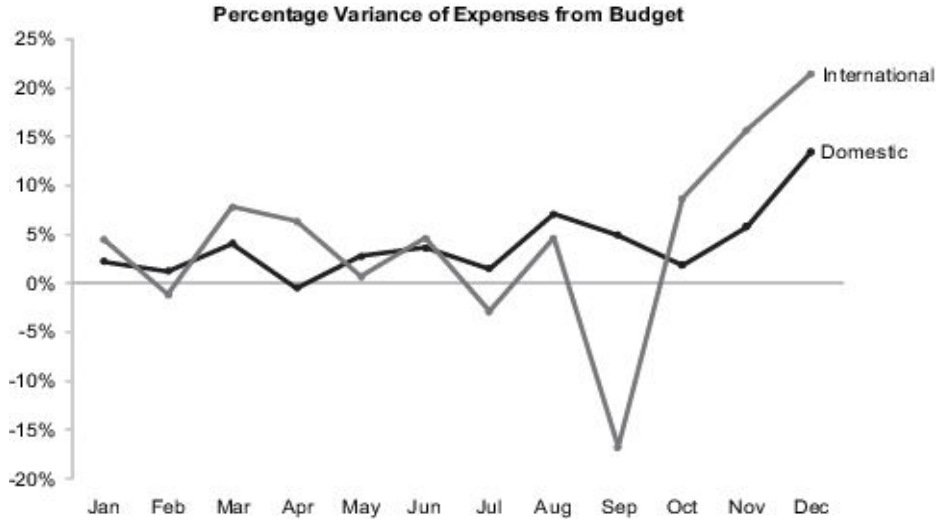
Domestic and international expenses exceed the budget, but domestic is much higher. Why?





Expressing deviations as percentages





Expenses (USD)	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
Domestic Actual	84,853	84,838	88,103	85,072	88,723	90,384	89,374	95,273	94,239	92,394	96,934	105,034
Domestic Budget	83,000	83,830	84,668	85,515	86,370	87,234	88,106	88,987	89,877	90,776	91,684	92,600
International Actual	12,538	12,438	14,934	14,033	13,945	15,938	14,086	15,934	13,945	17,338	19,384	22,394
International Budget	12,000	12,600	13,860	13,200	13,860	15,246	14,520	15,246	16,771	15,972	16,771	18,448

Distribution analysis

Outliers

Regra do dedão para calcular *outliers*:

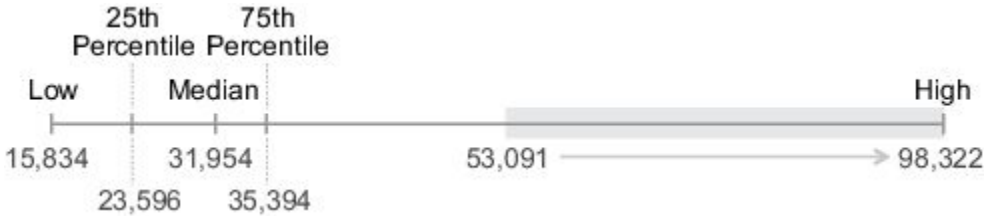
Tomando como exemplo a distribuição de salários utilizada no início da aula.

1) Percentil75 - Percentil25 = 35.394 - 23.596 = 11.798
2) 11.798 * 1.5 = 17.697
3) Upper threshold = Percentil75 + 17.697 = 53.091
4) Lower threshold = Percentil25 - 17.697 = 5.899

Outliers

Regra do dedão para calcular *outliers*:

Tomando como exemplo a distribuição de salários utilizada no início da aula.



1) Percentil75 - Percentil25 = 35.394 - 23.596 = 11.798
2) 11.798 * 1.5 = 17.697
3) Upper threshold = Percentil75 + 17.697 = 53.091
4) Lower threshold = Percentil25 - 17.697 = 5.899

Histograms

Single distribution displays

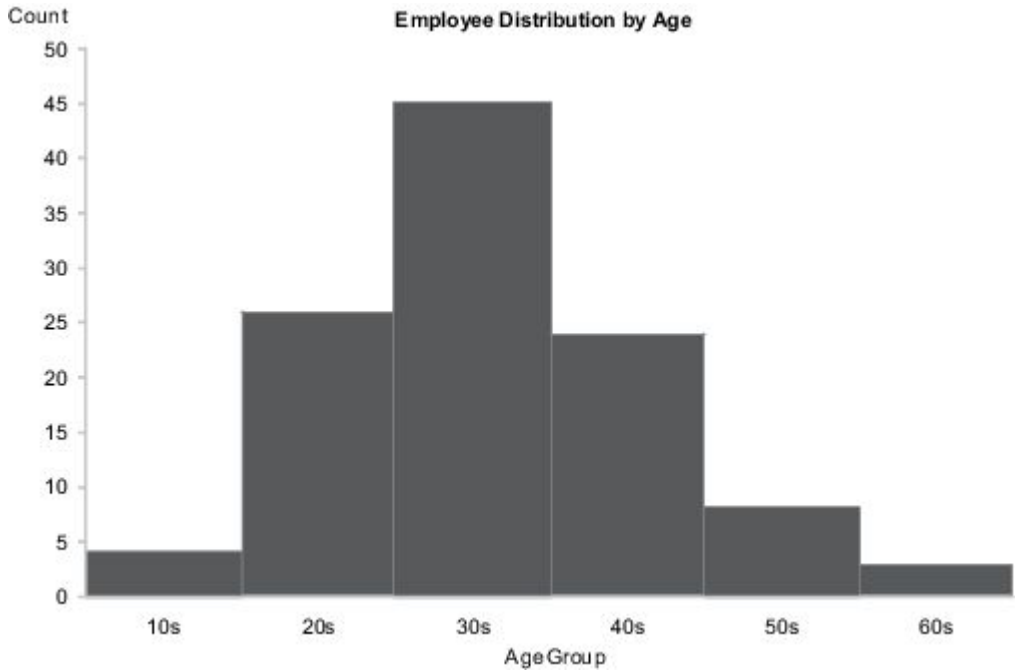
The bars touch one another to give an idea of continuity.

This representation shows:

- Distribution shape;
- It allows to compare the magnitude of each interval.

But we don't see:

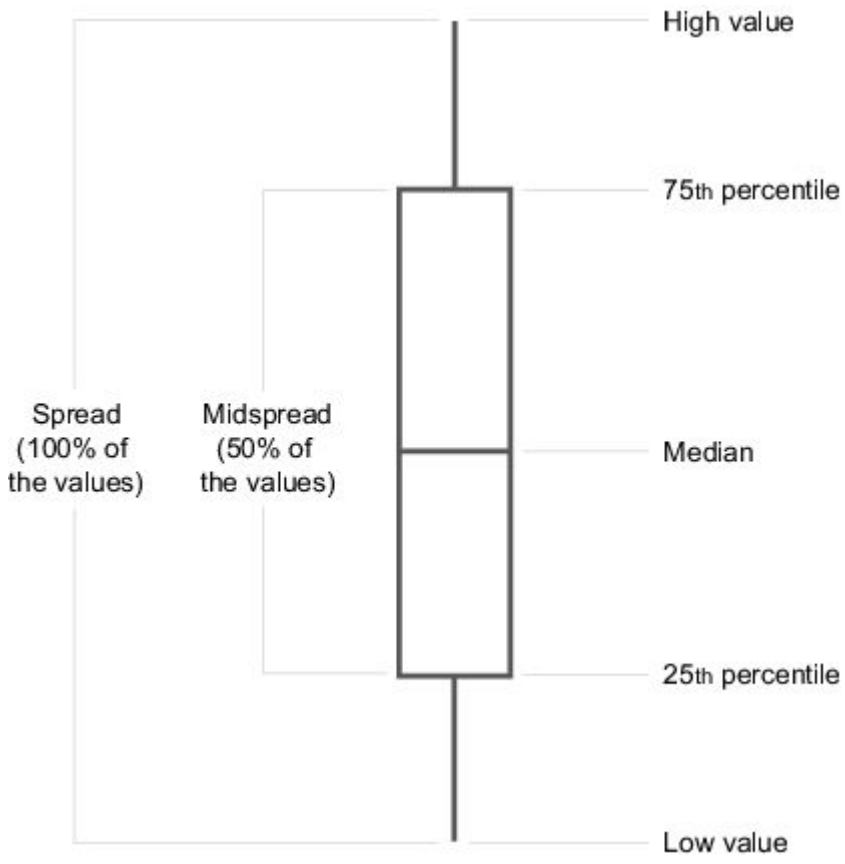
- Center;
- Spread.



Box plots

Multiple distribution displays

This type of graph is generally used when we want to represent many distributions.

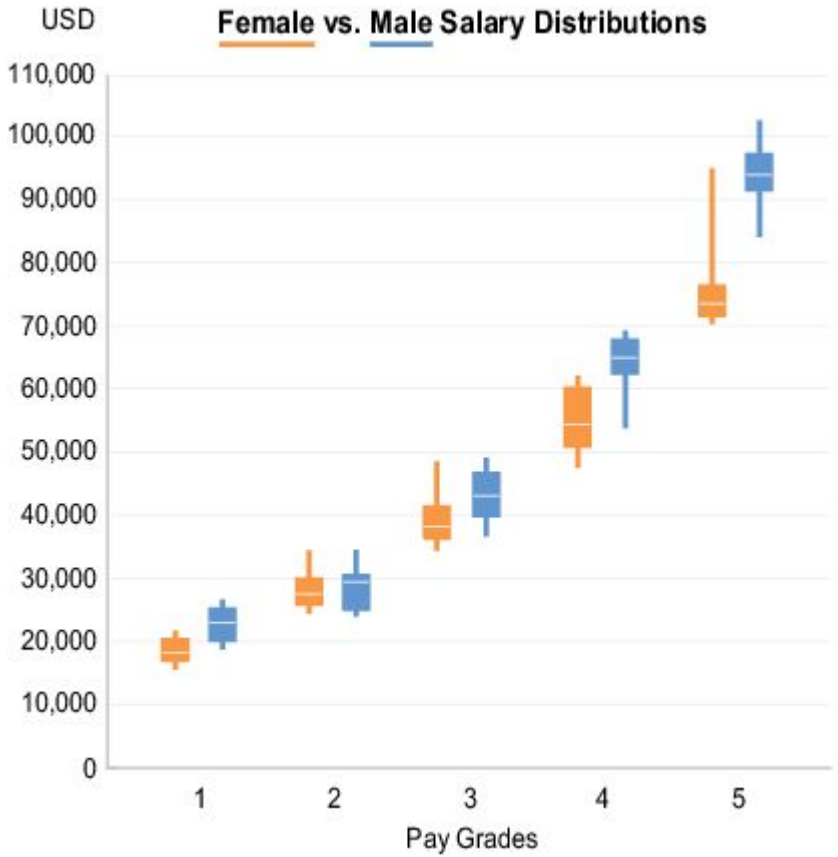


Box plot

Multiple distribution displays

Exercise:

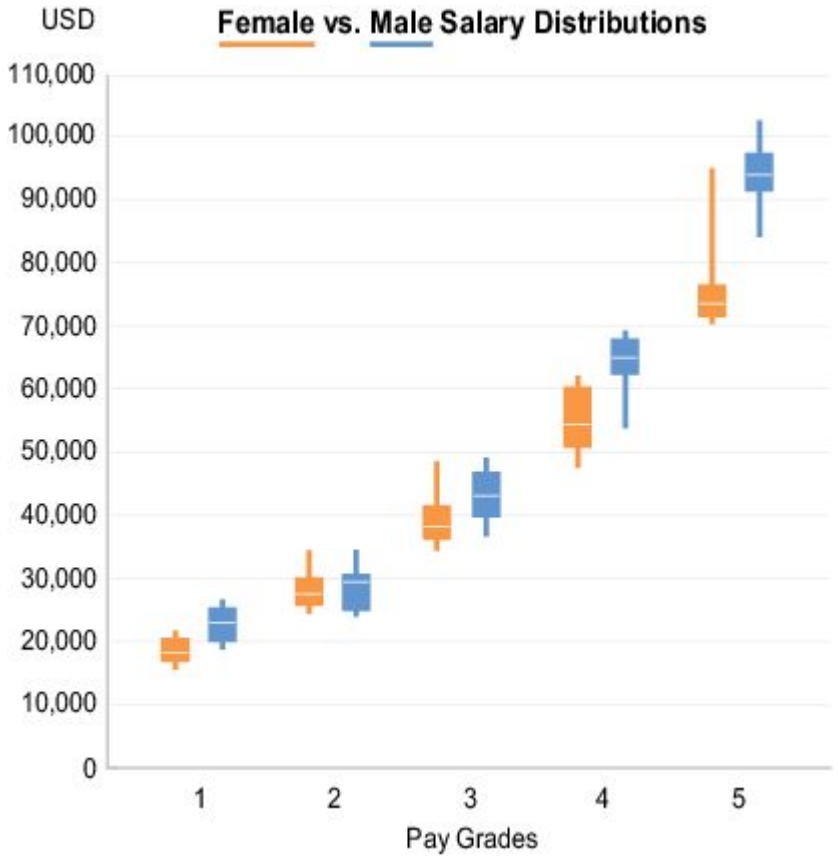
Interpret this graph regarding the salaries of men and women in 5 different ranges.



Box plot

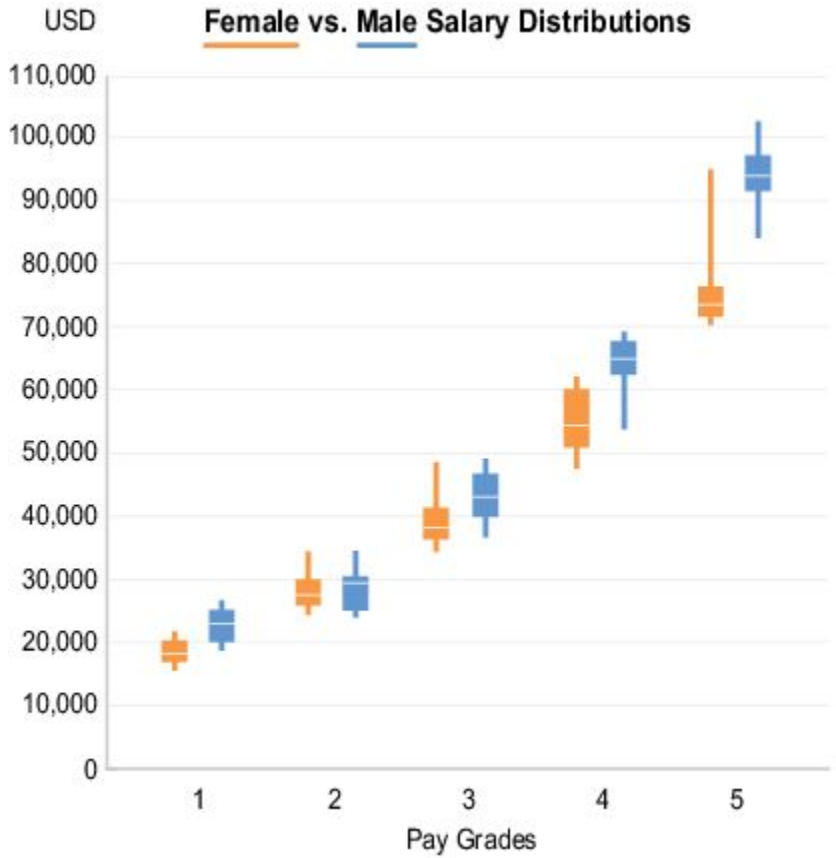
Multiple distribution displays

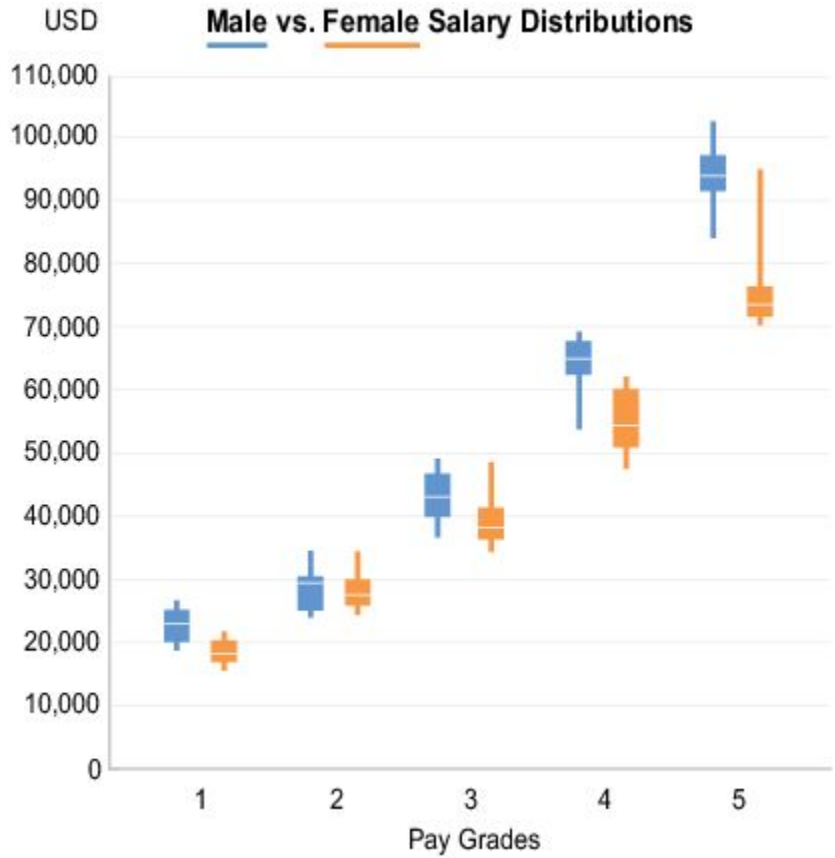
- Women are paid less than men in all salary ranges;
- The disparity in salaries for men and women becomes increasingly greater as salaries increase;
- Salaries vary the most for women in the higher salary grades.



Box plot

Multiple distribution displays



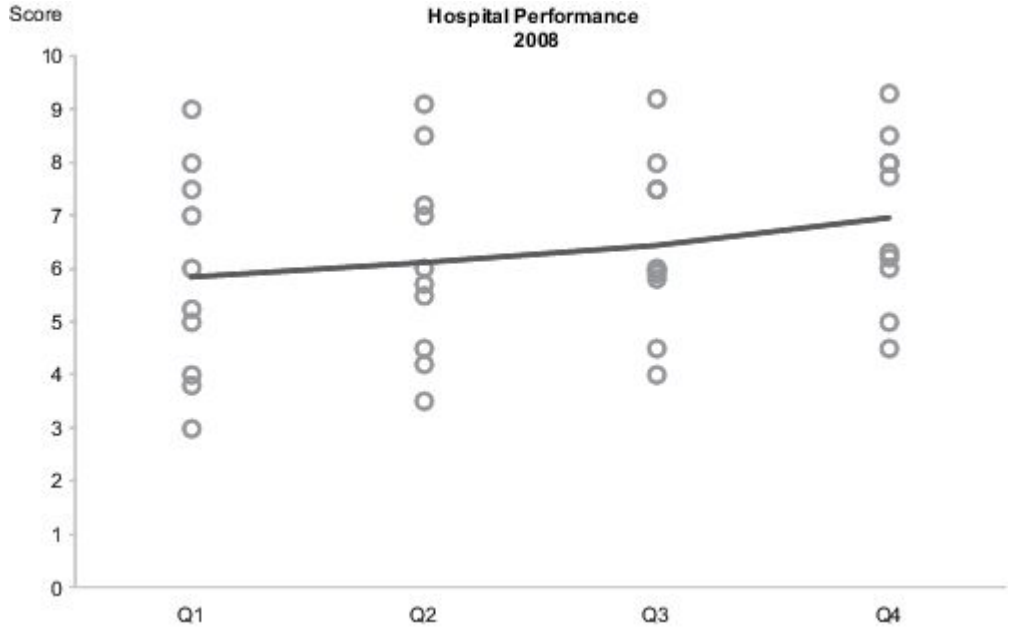


Multiple strip plots

Multiple distribution displays

In this example, the performance of 10 hospitals in health care is presented. The average of each distribution is connected by a line.

The distribution spread is smaller in Q4 than in Q1.

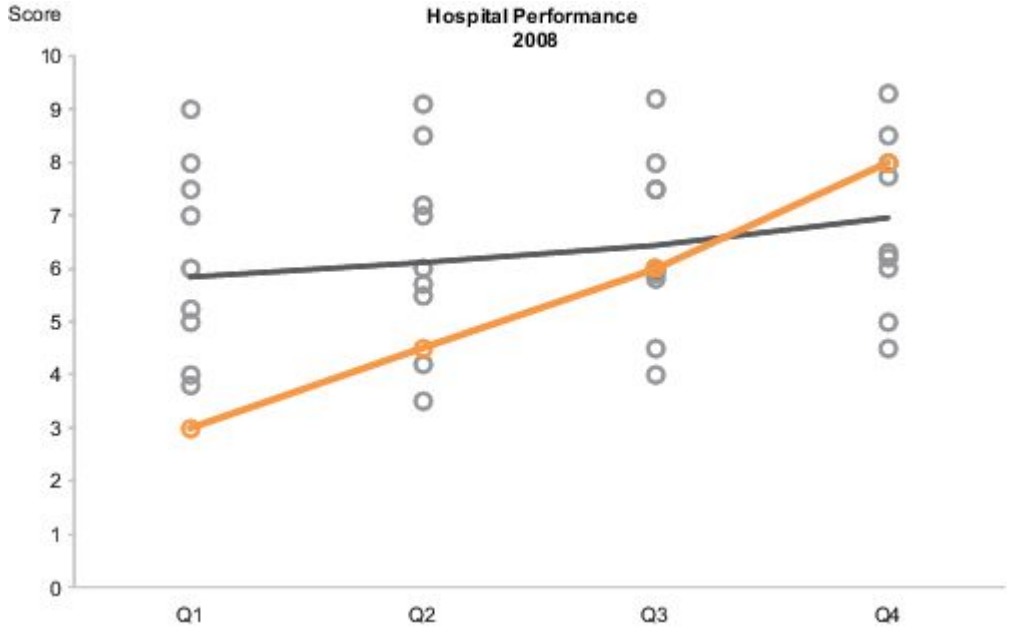


Multiple strip plots

Multiple distribution displays

The hospitals were not connected to avoid cluttering the graph. It would be interesting to offer this feature in an interactive way.

This type of representation is feasible for small sets of values.



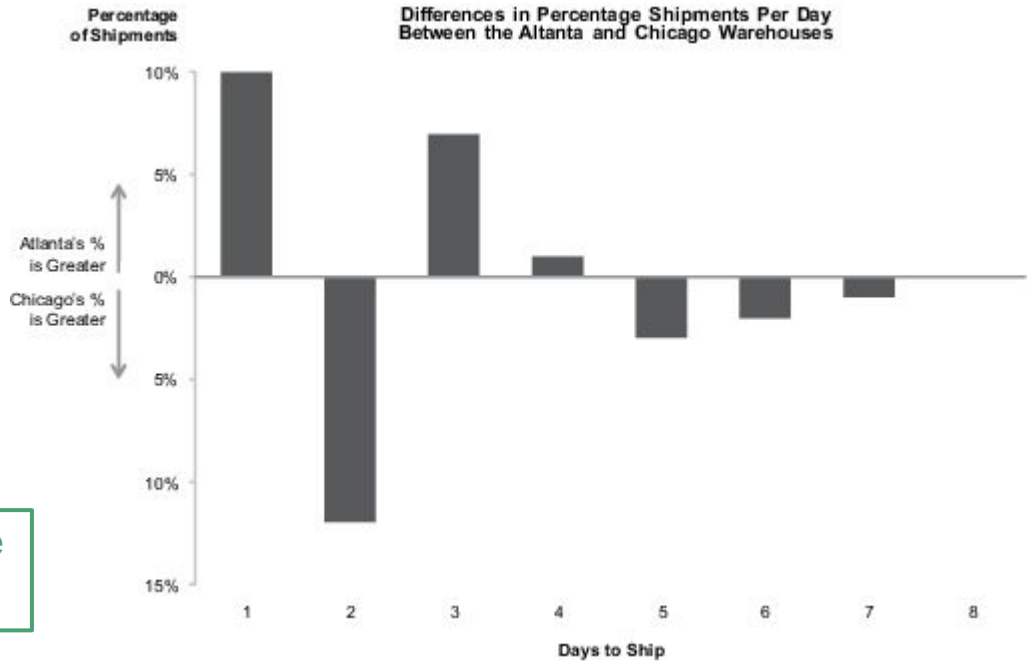
Distribution deviation graphs

Multiple distribution displays

When we want to focus on how 2 distributions are different, we can directly represent the differences on a deviation graph.

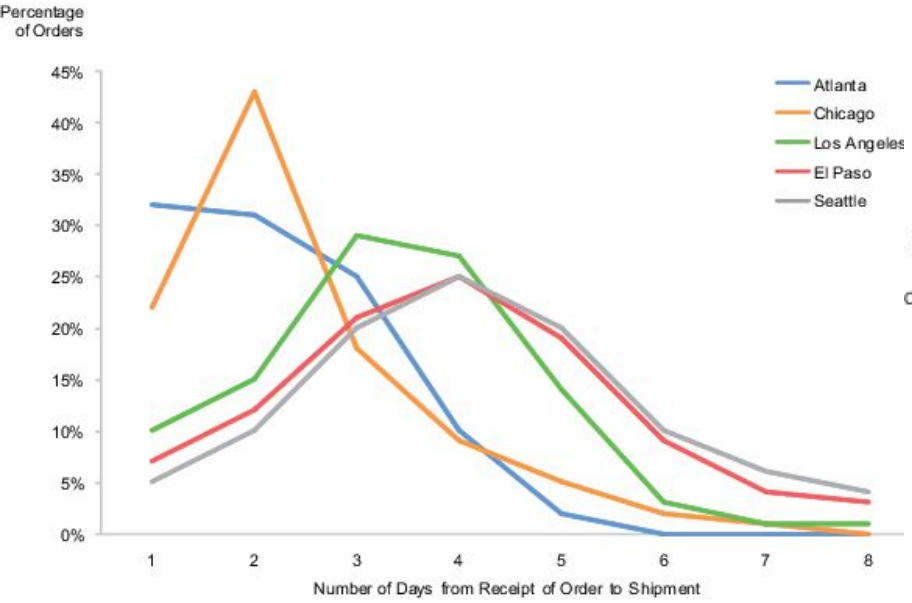
In this example, we see the differences, in percentage, of deliveries per day.

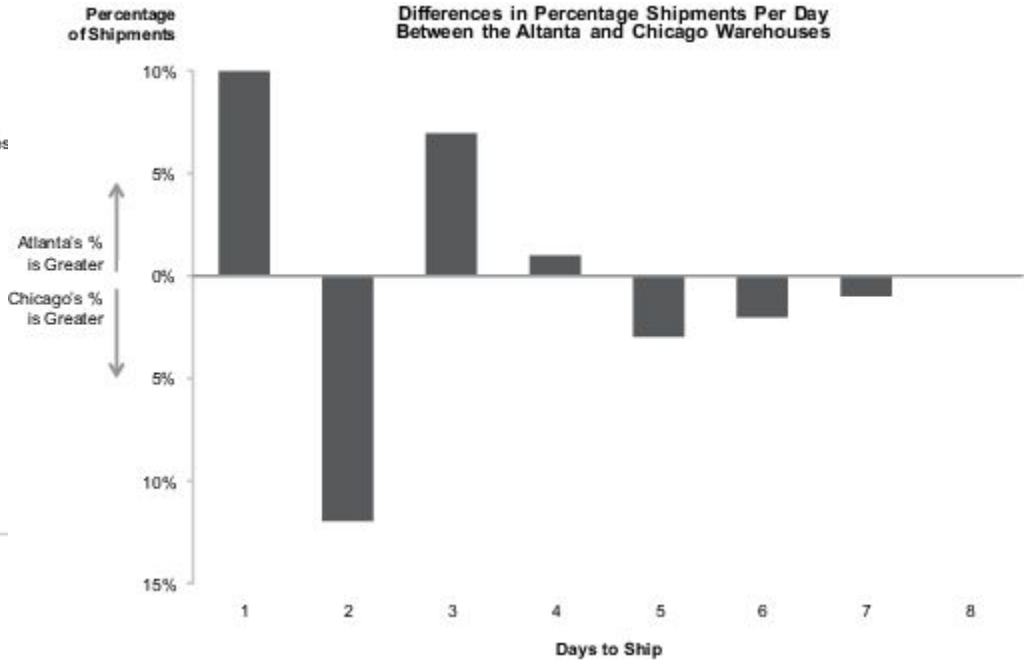
We are not interested in the shape of the distributions, only in the differences.



Distribution deviation graphs

Multiple distribution displays

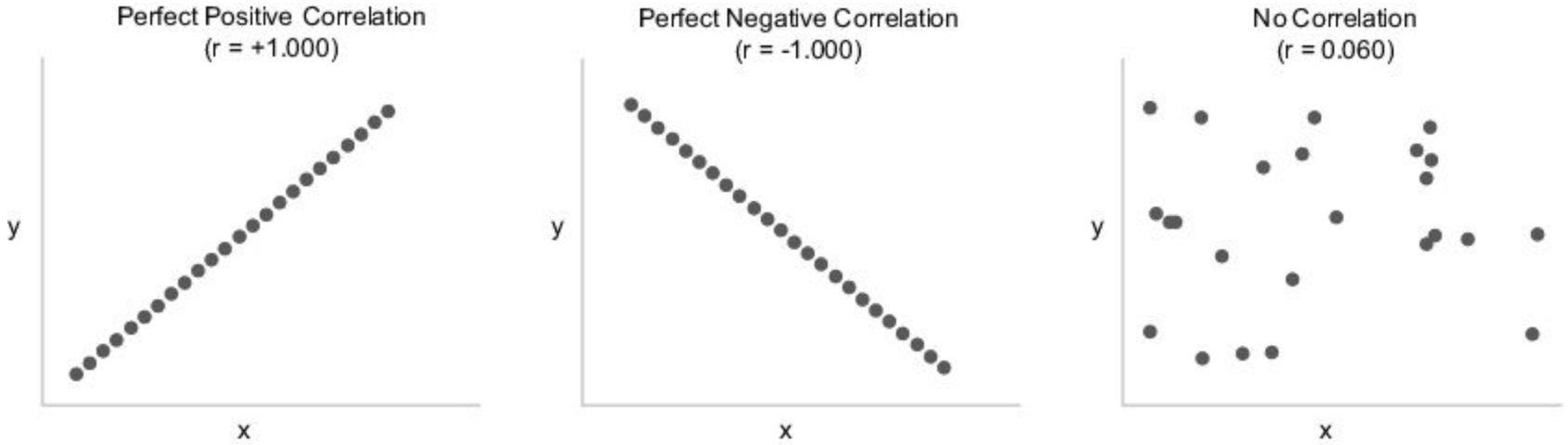




Frequency polygons

Correlation analysis

Statistical summaries of correlations



There is no general way to determine whether a value of r is considered strong or weak. It depends on the data set and the purpose of the analysis.

Resumos estatísticos das correlações

O coeficiente de determinação (r*r) descreve a força da correlação mas não sua direção.

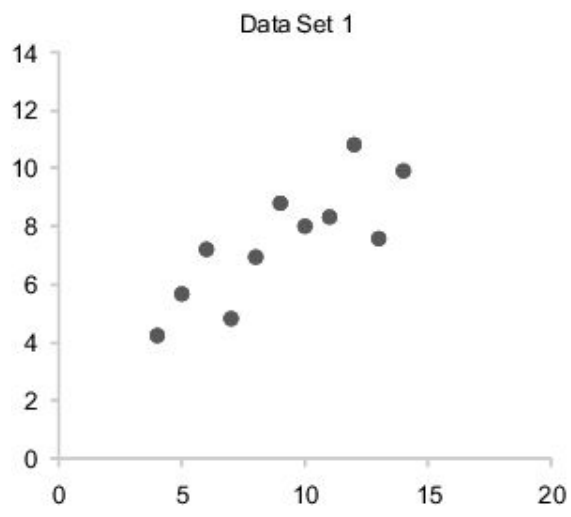
- Coeficiente de correlação linear: +0,993
- Coeficiente de determinação: 0,993 * 0,993 = 0,986049 = 0,986

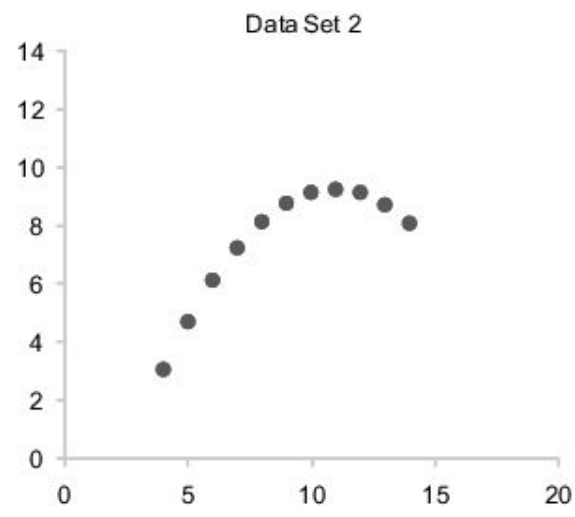
Pode ser expresso como percentual. No exemplo acima o coeficiente de determinação indica que 98,6% das mudanças na variável dependente (peso) pode ser determinado pelo valor da variável independente (altura).

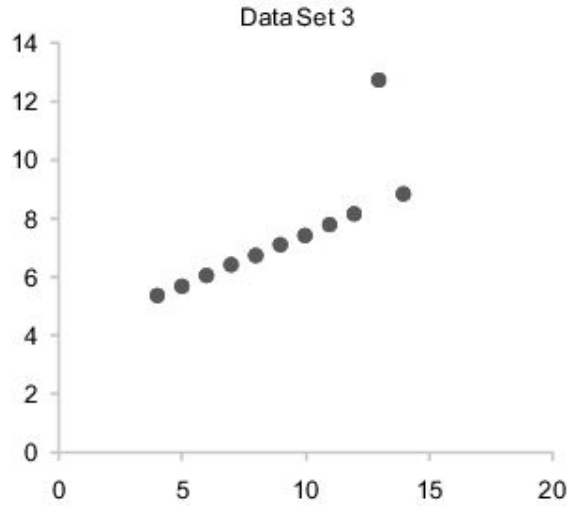
Statistical summaries of correlations

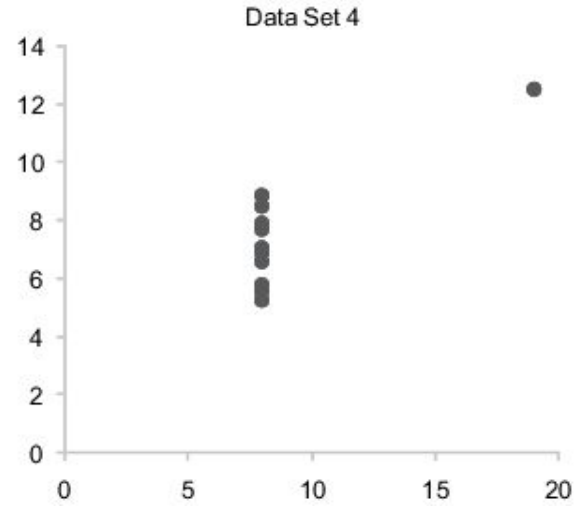
We cannot rely solely on these measures to describe the correlations.

Let's look at an example of the Anscombe Dataset, which consists of 4 sets of values, containing 11 paired values.



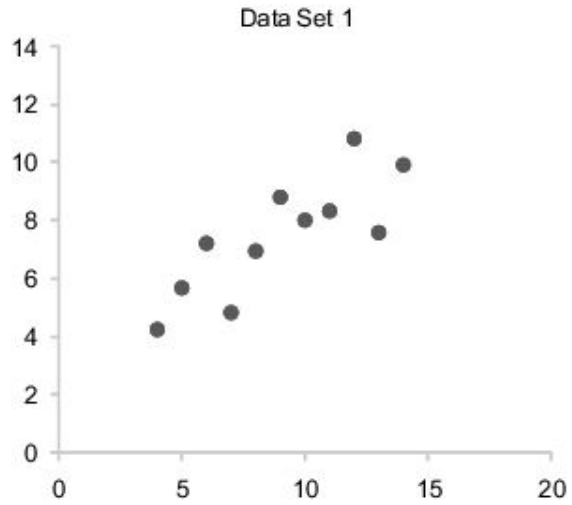


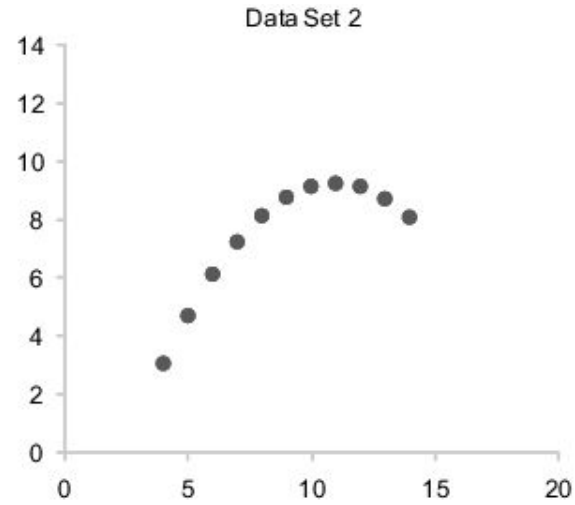


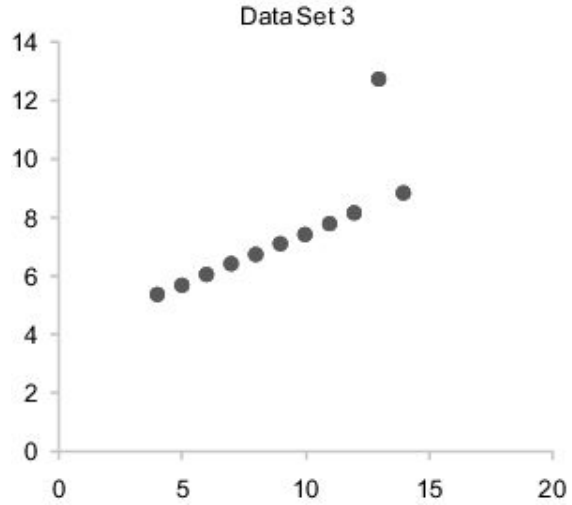


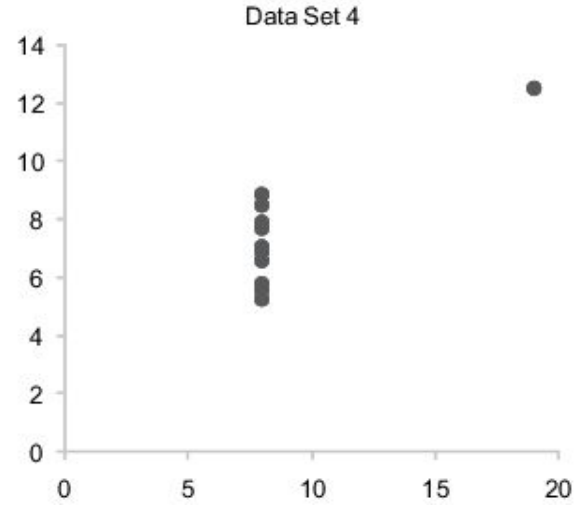
Statistical summaries of correlations

Despite having very different shapes, see what happens with the measures that summarize this data.









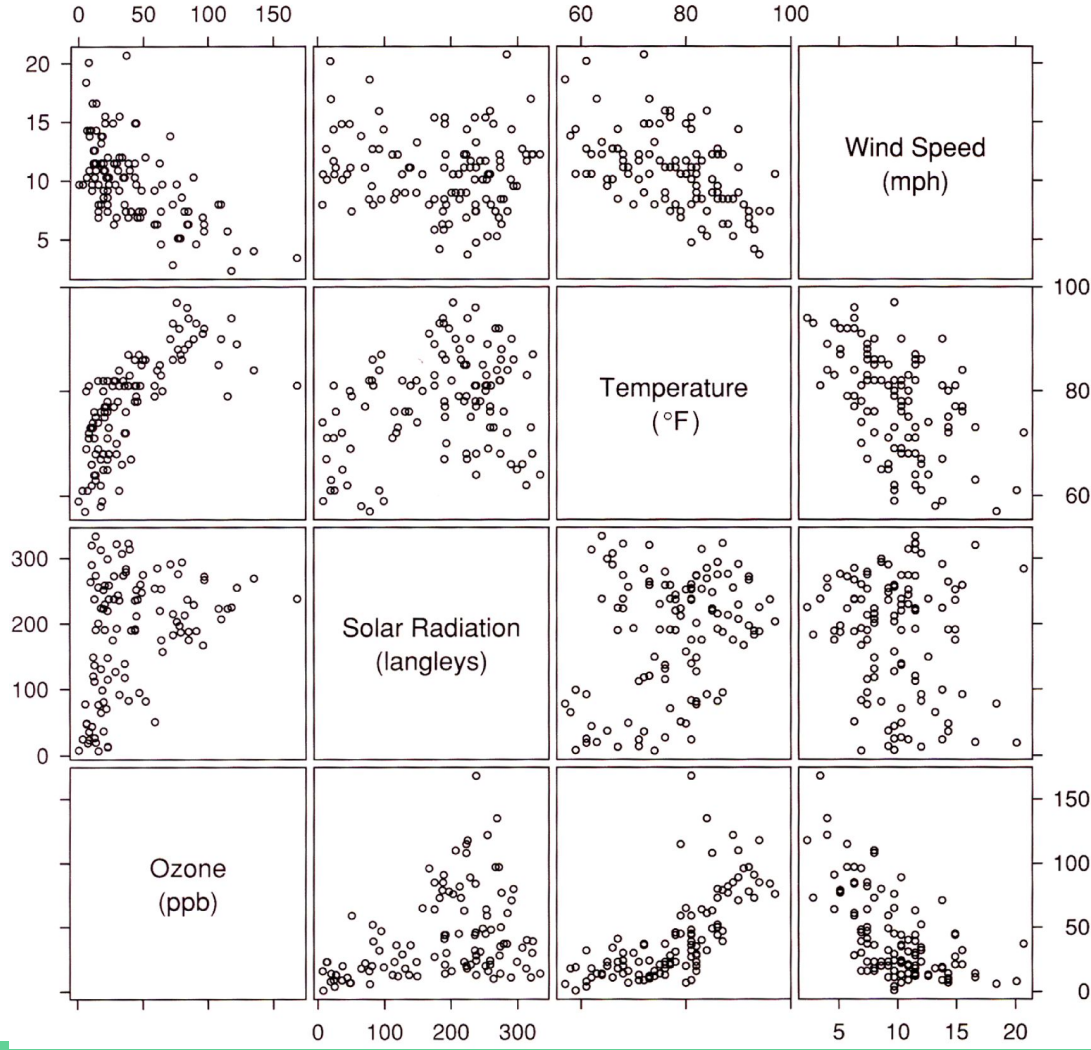
Statistical summaries of correlations

N (número de valores)	11
Média dos valores no eixo x	9,0
Média dos valores no eixo y	7,5
r	0,82
r*r	0,67
Trend line equation	y = 0.5x + 3

Scatterplot matrices

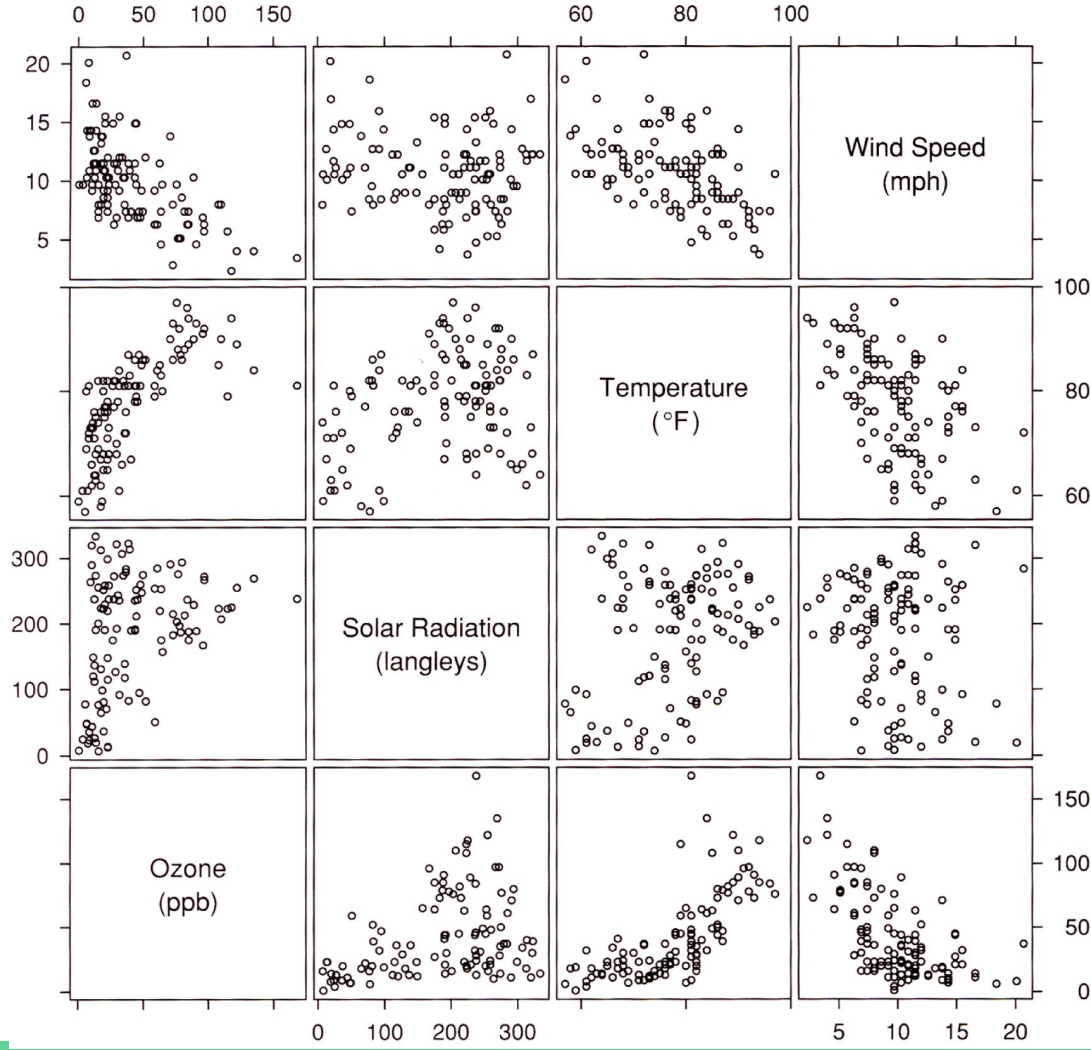
Scatterplot matrices are used to compare multiple pairs of variables. It is an idea similar to the small multiples or trellises applied to the scatterplot.

But be careful with the scales of the axes.



Scatterplot matrices

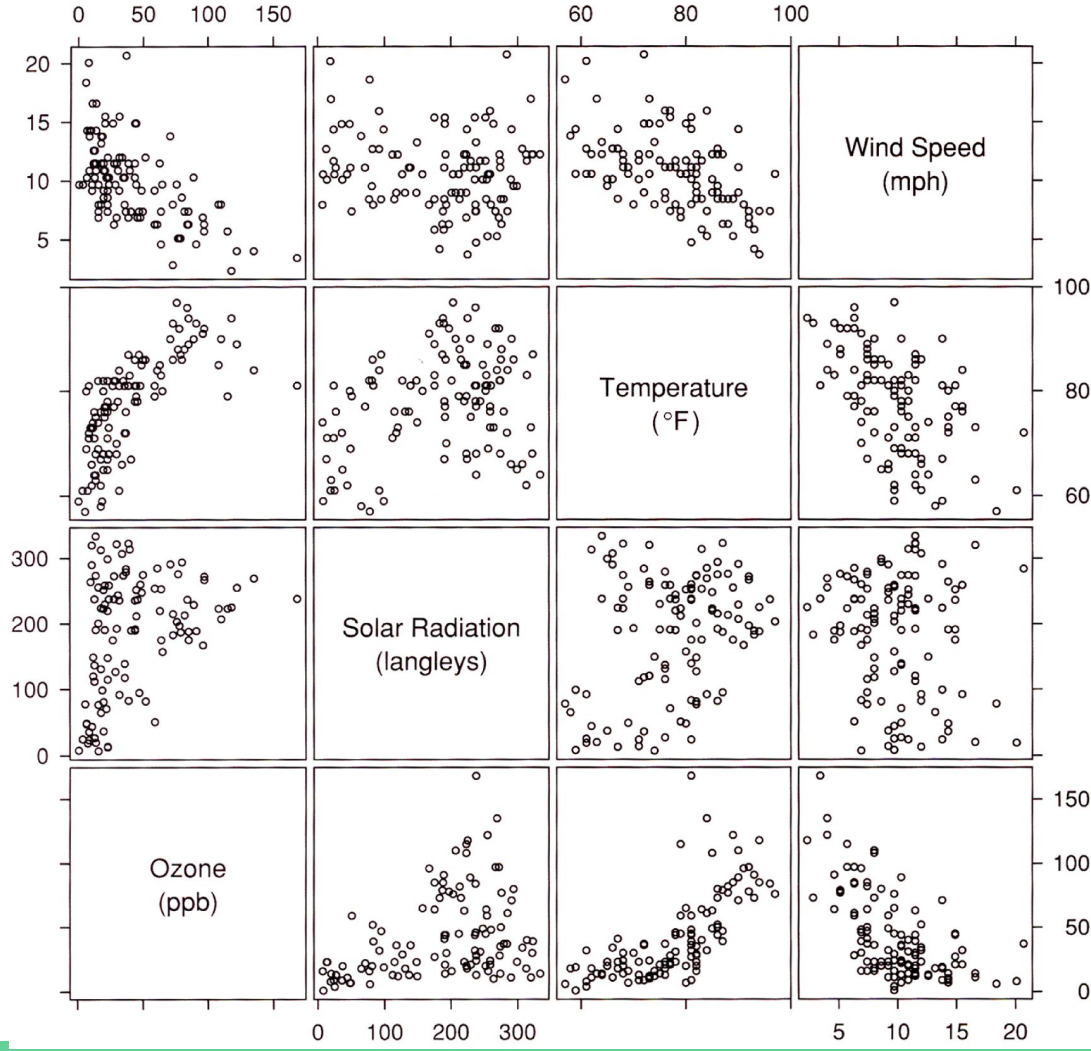
It allows analyzing the relationship between 4 atmospheric variables.



Scatterplot matrices

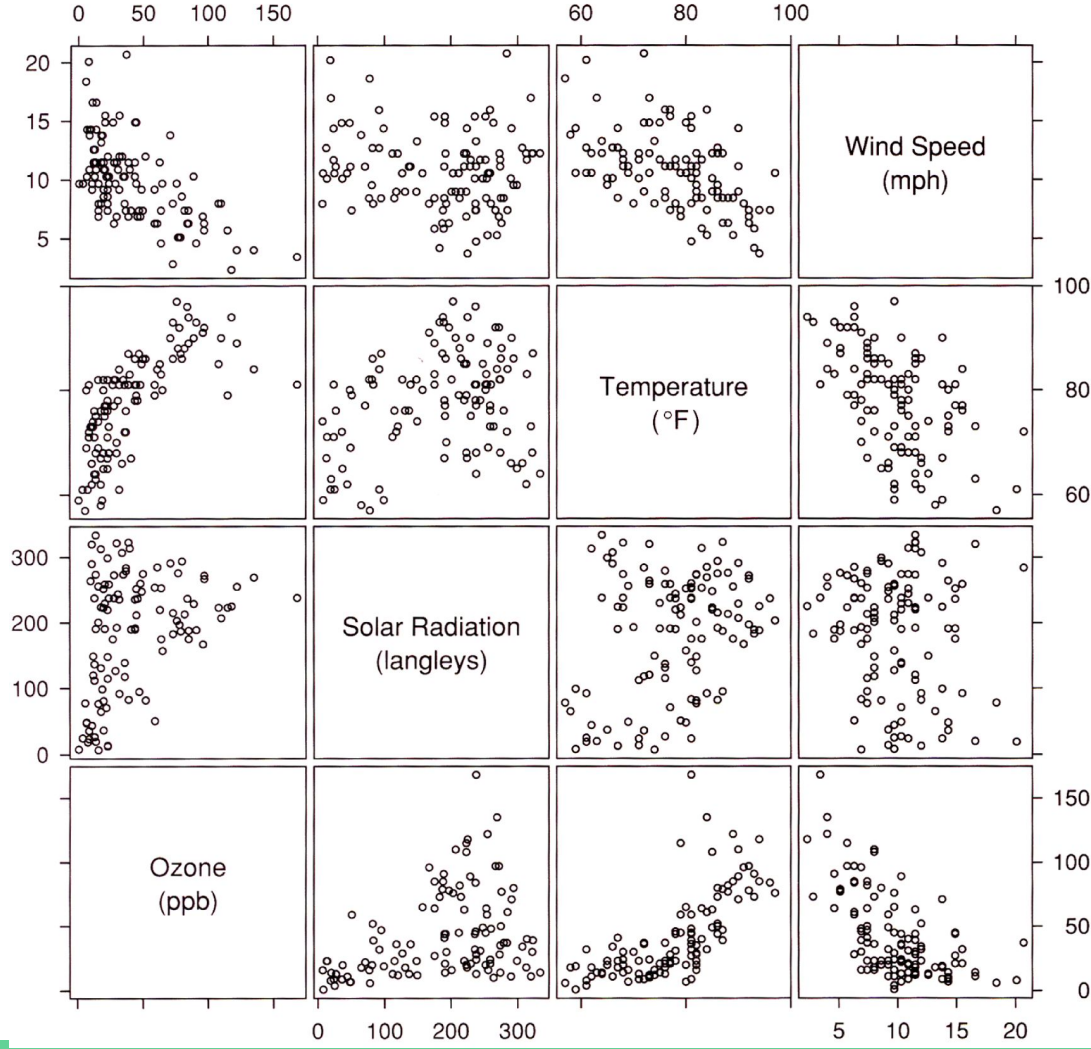
The variable name serves as a label for row and column.

To compare 2 variables, look at the intersection of one row with the other column.



Scatterplot matrices

Is there a correlation between ozone and wind speed?



Scatterplot matrices

Going through the scatterplots, we can understand the “story they tell”

Ozone negatively correlates with wind speed.

Ozone correlates positively with temperature.

Ozone apparently correlates positively with solar radiation, but in a weak way.

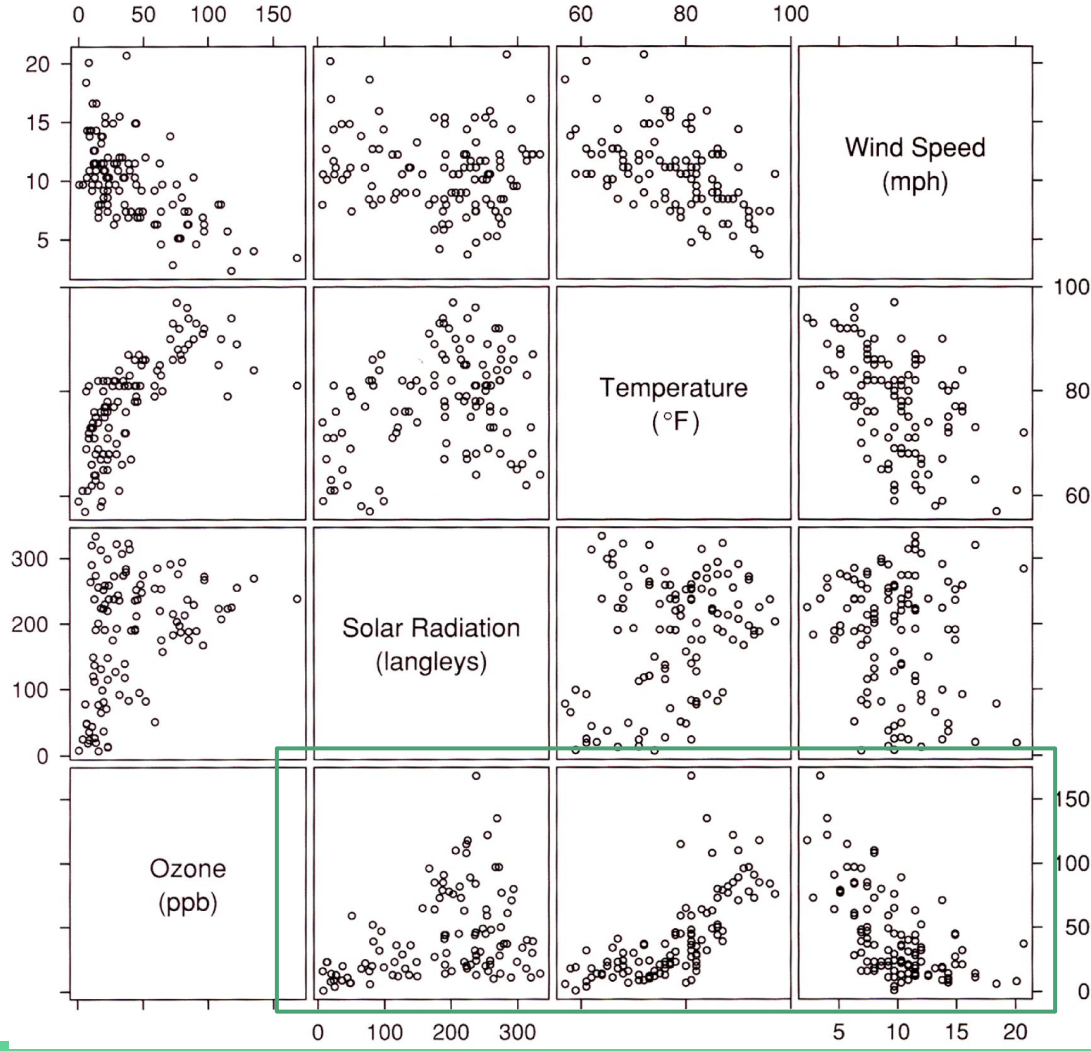


Table lenses

It was originally created by Ramana Rao.

This example displays baseball statistics from 1987 for 323 baseball players.

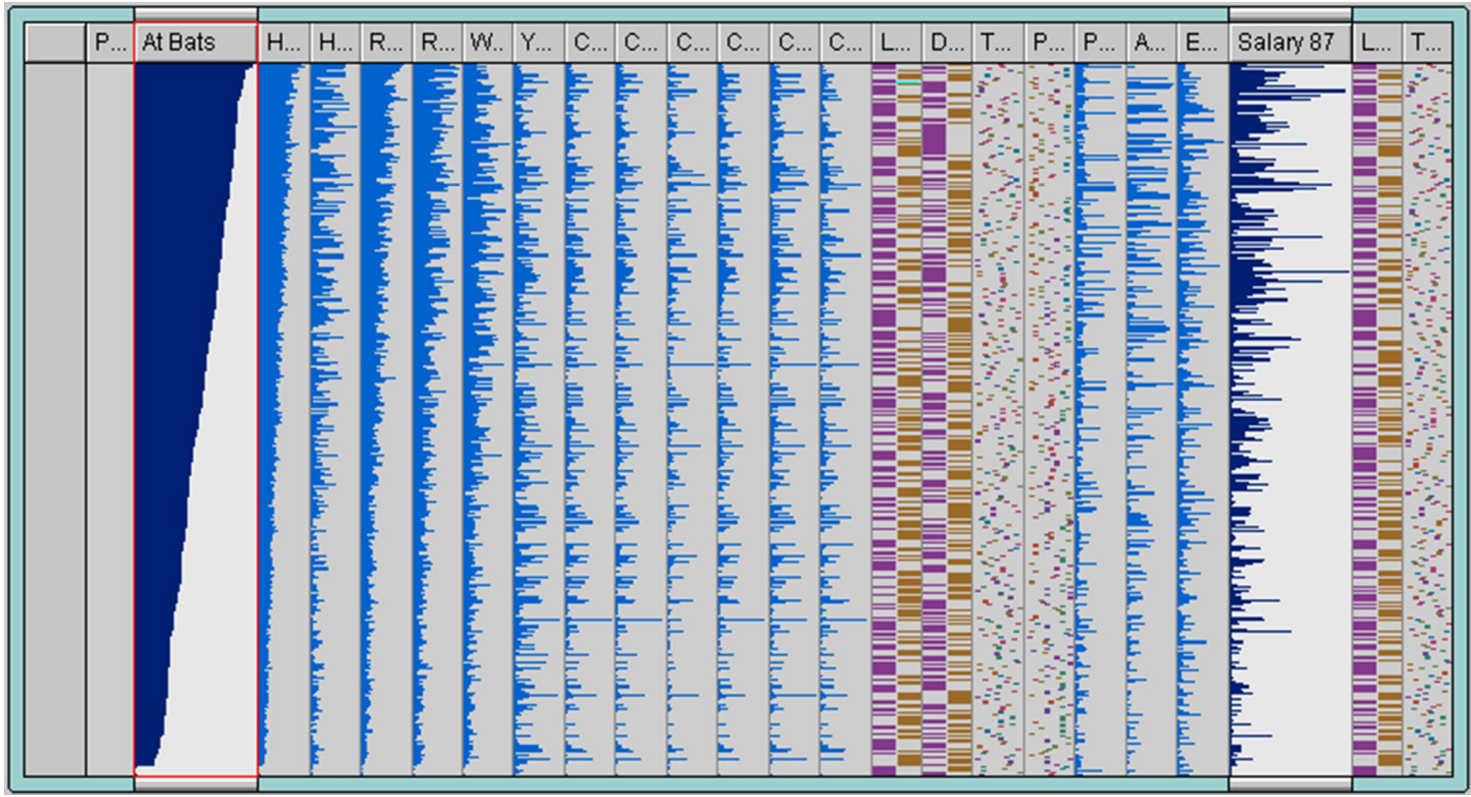
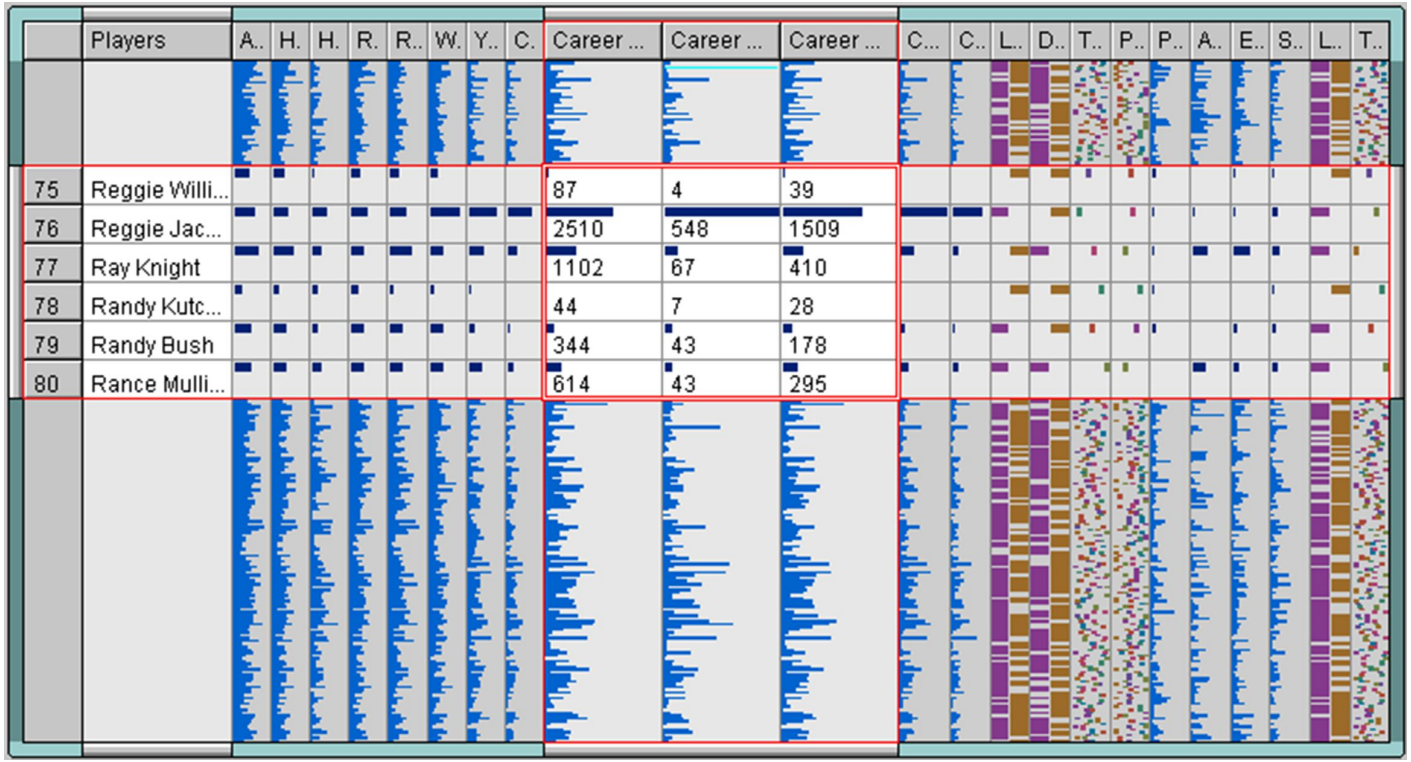


Table lenses

Details on demand.



Multivariate analysis

Multivariate analysis

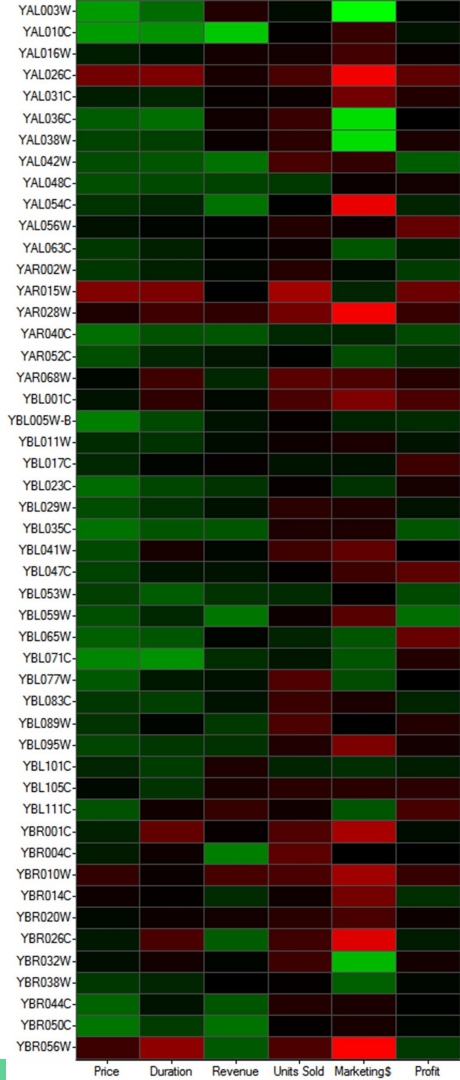
Multivariate analysis compares multiple instances of several variables at once.

The values of these variables for each object must be combined to form the object's multivariate profile.

Multivariate heatmaps

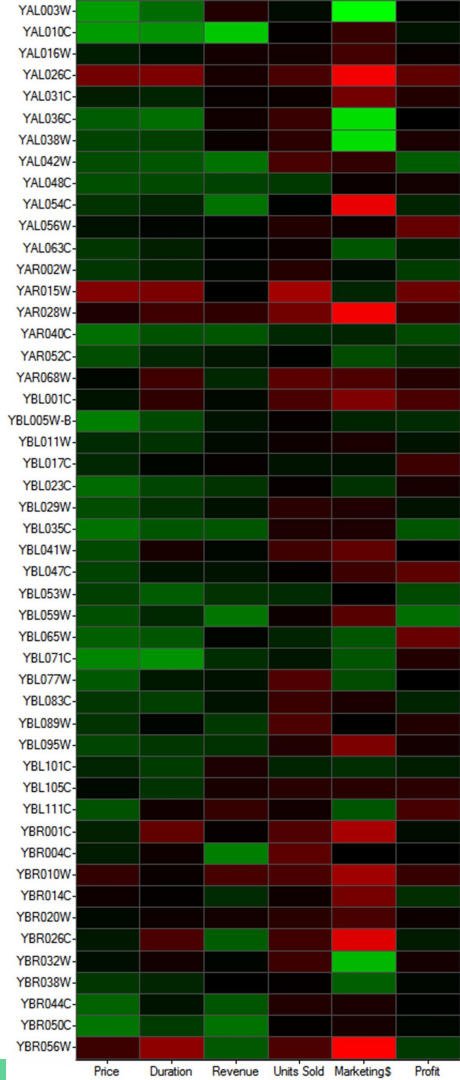
Columns represent:

- Price;
- Duration (length of time on the market);
- Revenue;
- Units sold;
- Marketing expenses;
- Profit.



Multivariate heatmaps

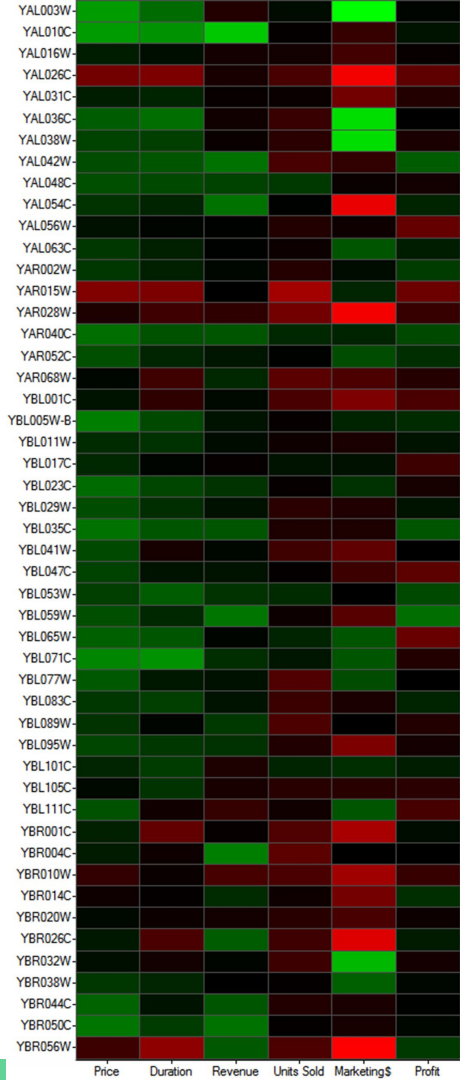
- Higher-than-average are green (the darker the higher)
- Near-average are black
- Lower-than-average appear as red (the darker the lower)



Multivariate heatmaps

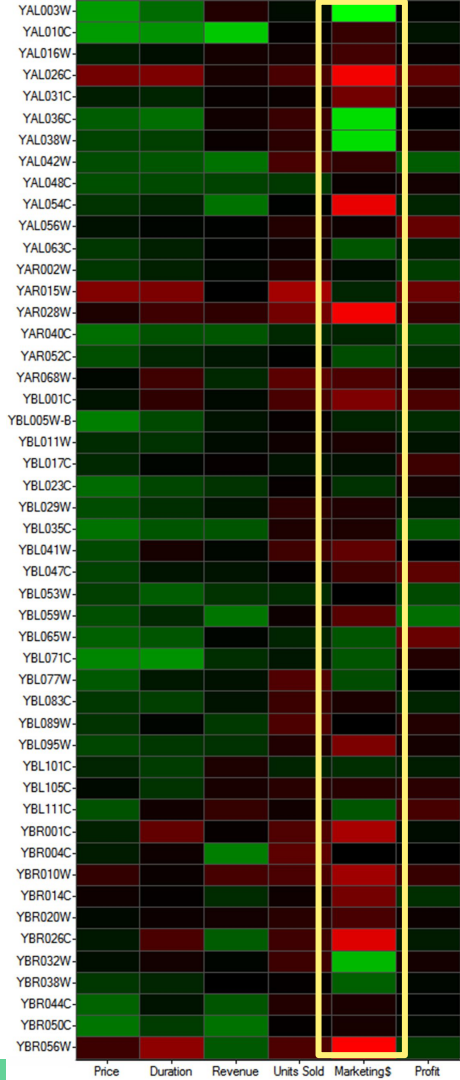
Are you able to find patterns in this heatmap?

And exceptions?



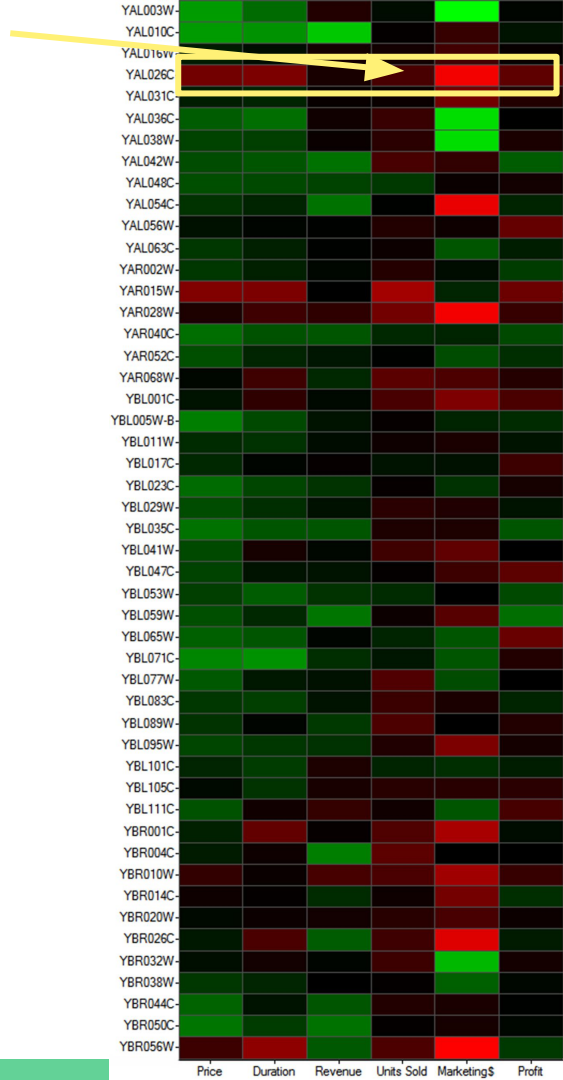
Multivariate heatmaps

Note that bright green and bright red values stand out in the Marketing column.



Multivariate heatmaps

Note that the product YAL026C has lower values than the others (except for Revenue).



Multivariate heatmaps

What do you think about the colors of this heatmap?

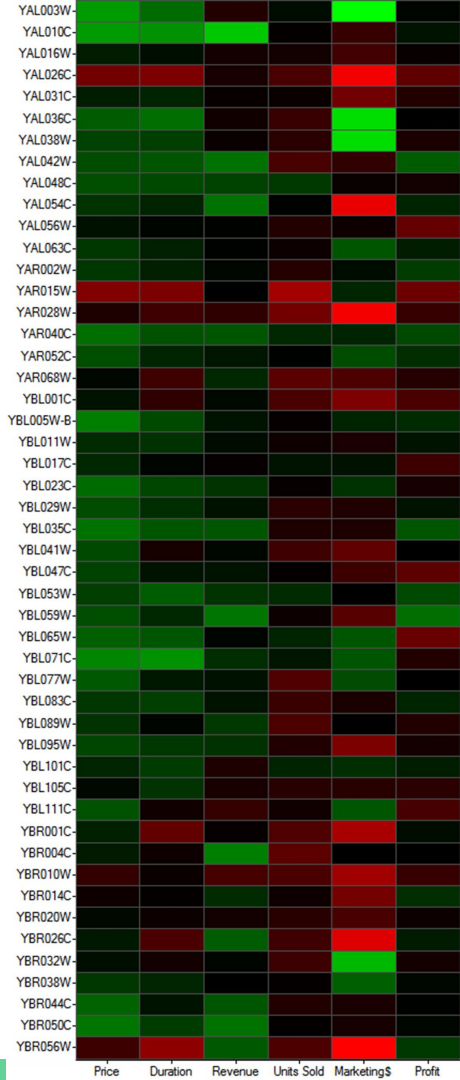
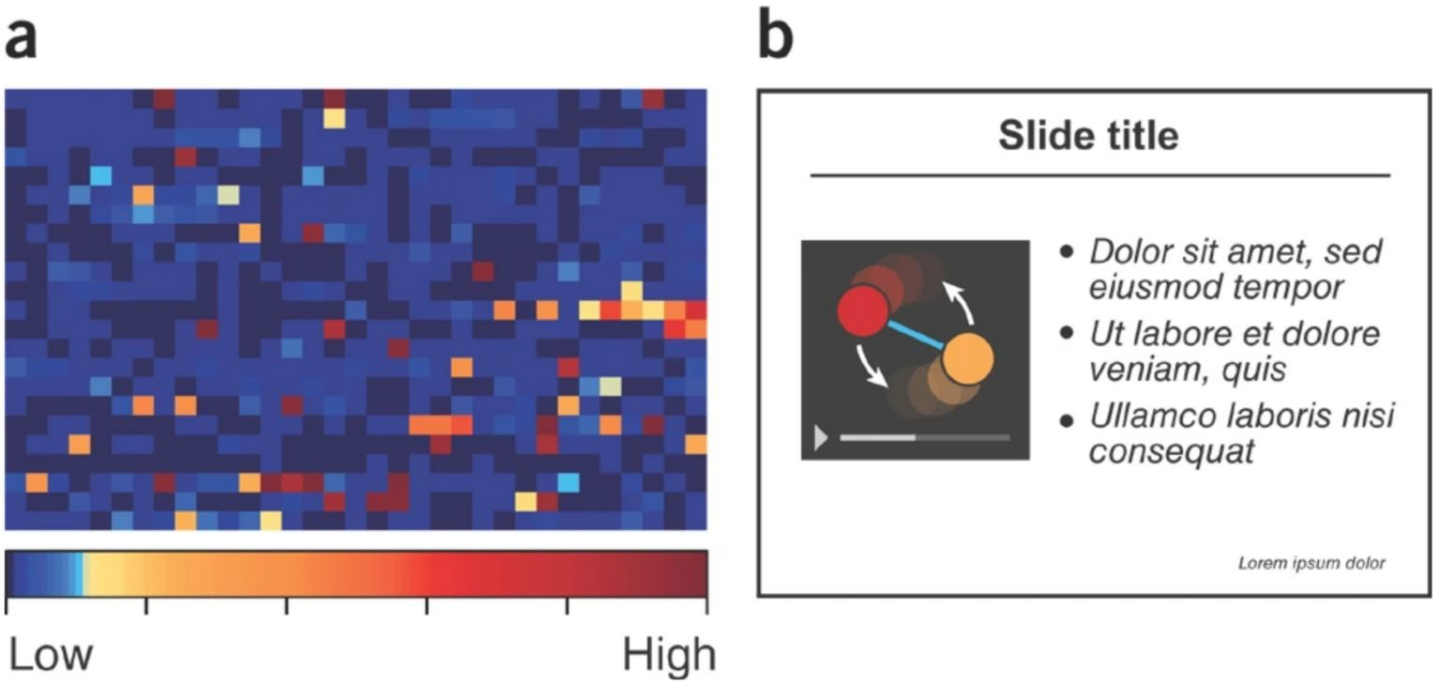


Figure 2: Discordances between salience and relevance can be harmful.

From: Salience to relevance



(**a**) The relative visibility of hues in the color scale is asymmetric, making higher values (represented by deep red) less apparent. (**b**) Continuously moving images can be distracting and can compromise the viewer's ability to concentrate on other content.

https://doi.org/10.1038/nmeth.1762

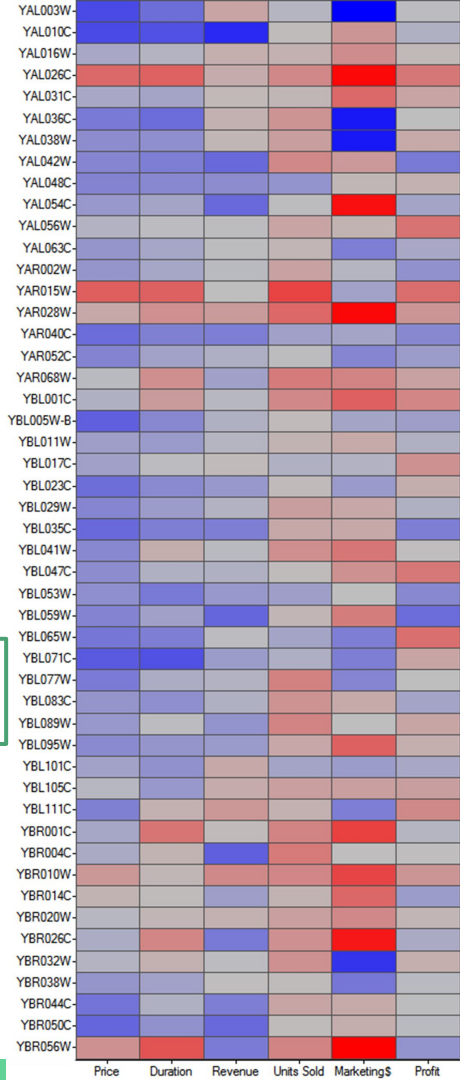
Multivariate heatmaps

Positive values are blue, negative values are red and the values near 0 tend to gray.

These colors can be distinguished even by those who are color blind.

Gray does not attract our attention, giving the false idea of high value like black.

Limited usefulness. It is difficult to see the color combination for a product as a pattern.

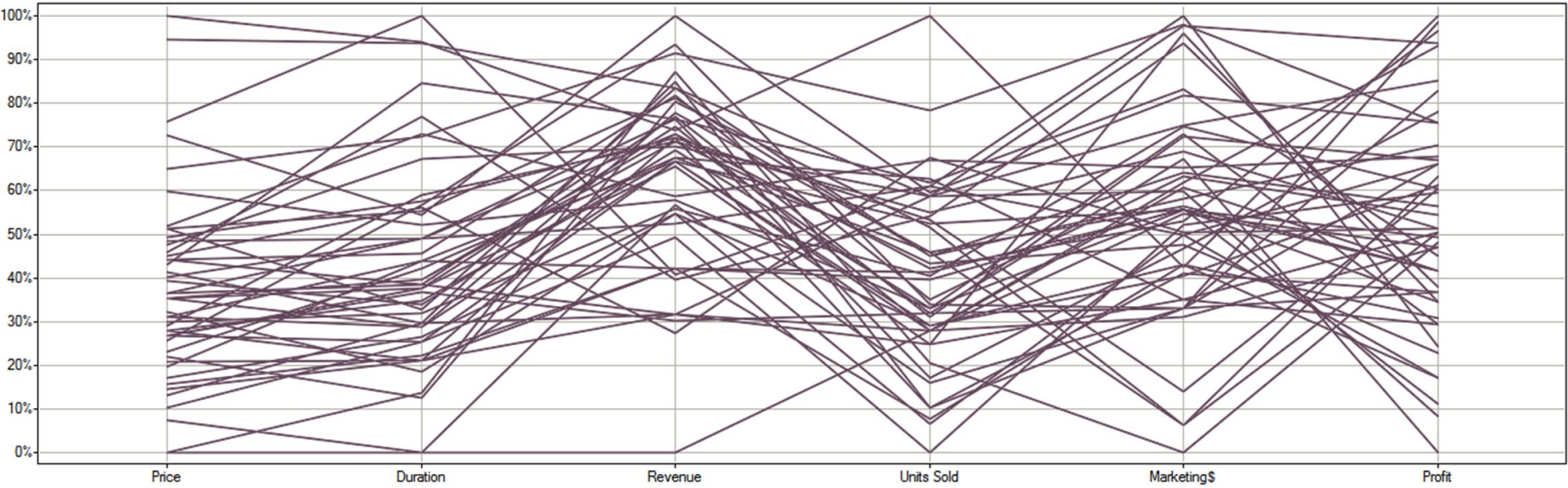


Parallel coordinates plot

- So far, we know line graphs that connect values along a scale like time.
- We saw that the line graphs do not support the visualization of many lines at the same time, it gets cluttered.
- But what about this new type of chart?

Parallel coordinates plot

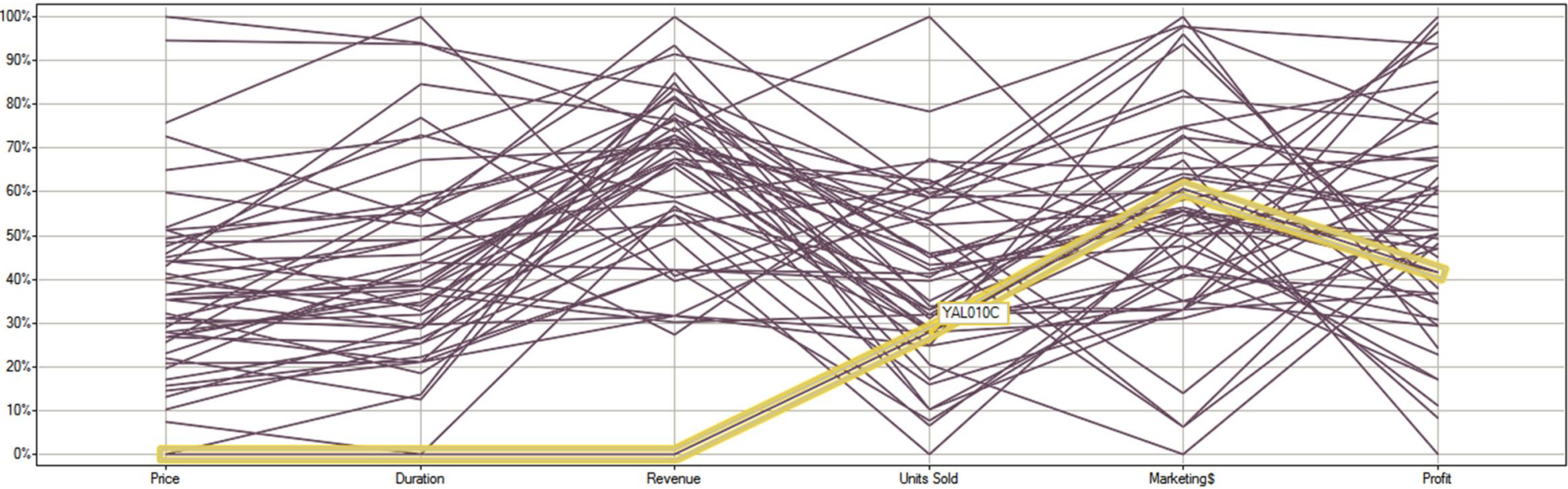
Connects values associated with different variables



Values can be displayed as absolute or percentage scale

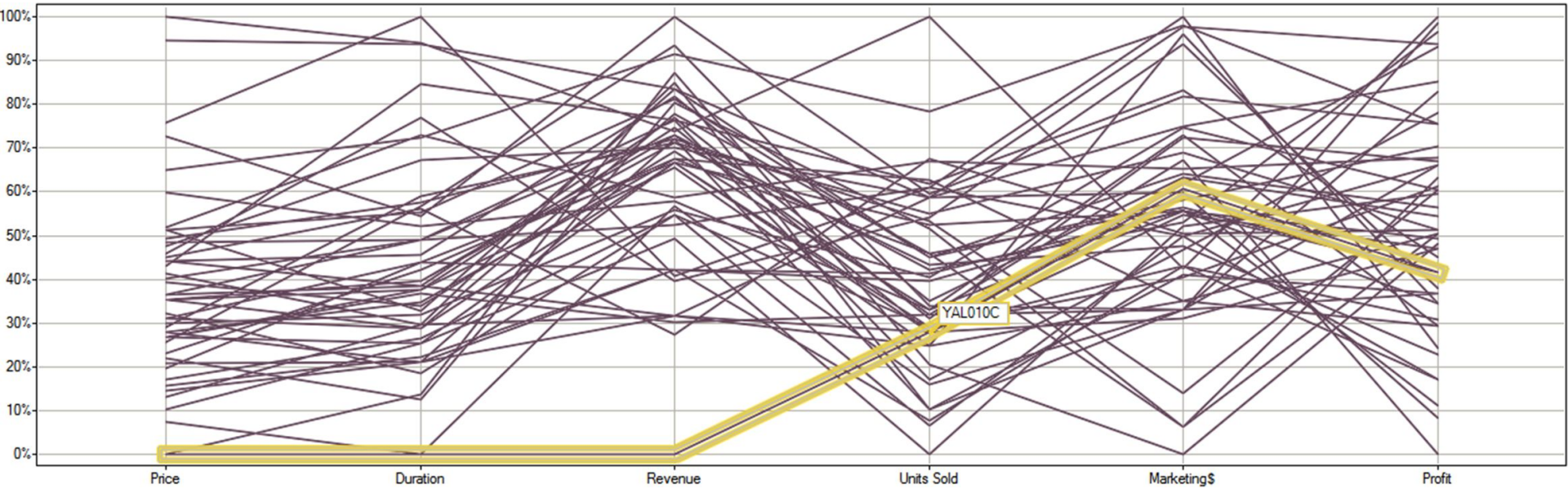
Parallel coordinates plot

Let's look at an example in which it is possible to highlight a line (product).



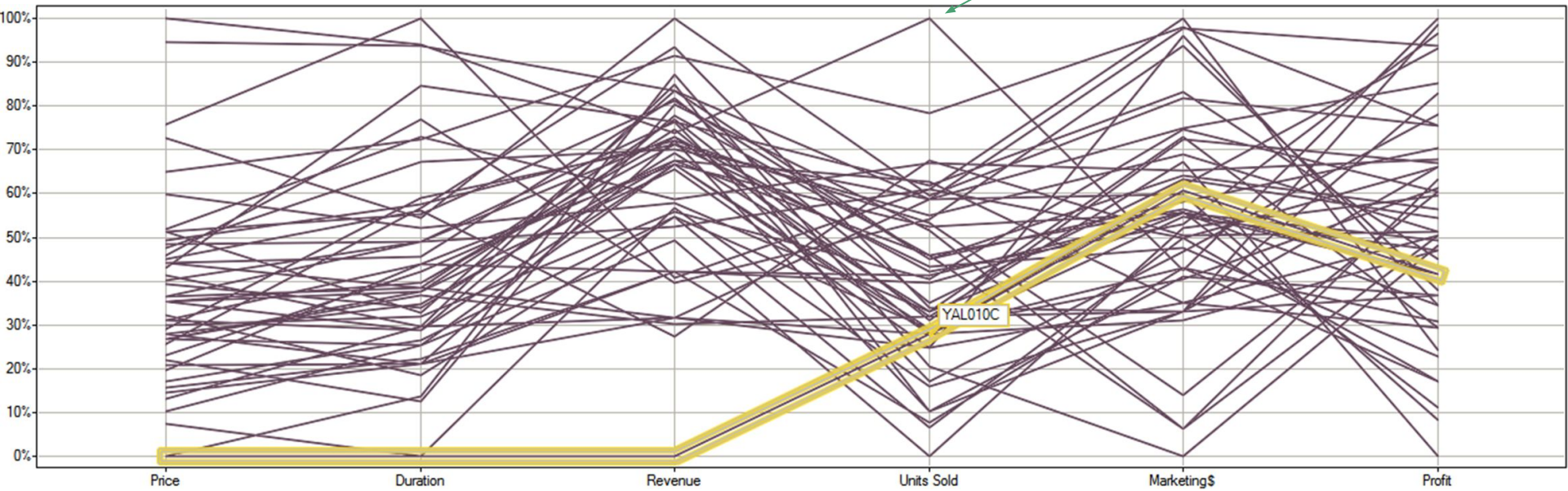
Parallel coordinates plot

The revenue for this product is well below the others, perhaps due to its short life cycle and / or low price.



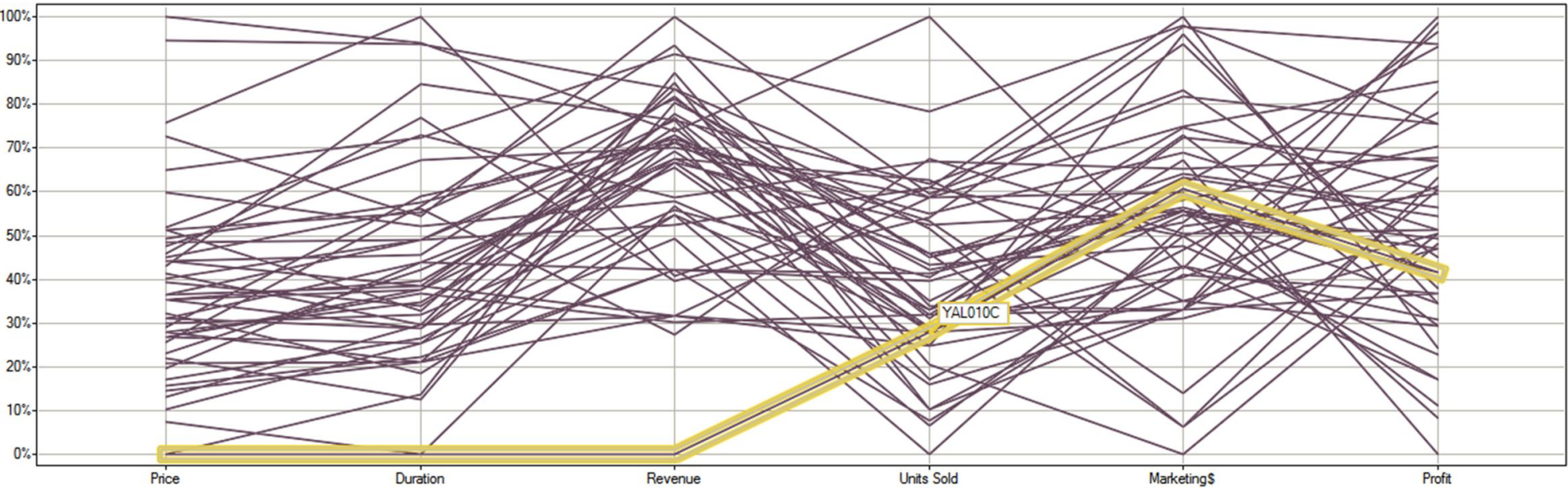
Parallel coordinates plot

A product sold quite a bit more than its closest rival.



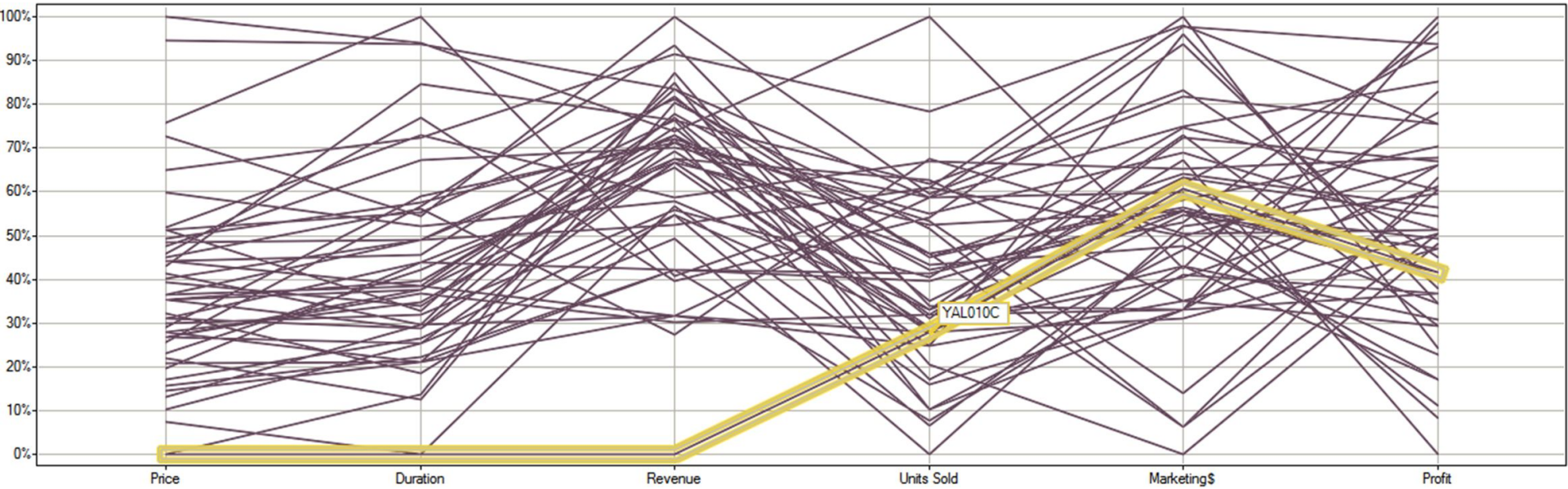
Parallel coordinates plot

Most products have been on the market from between 20% and 60% of the full range of time for all products.



Parallel coordinates plot

There is a concentration of revenues between 65% and 85%.



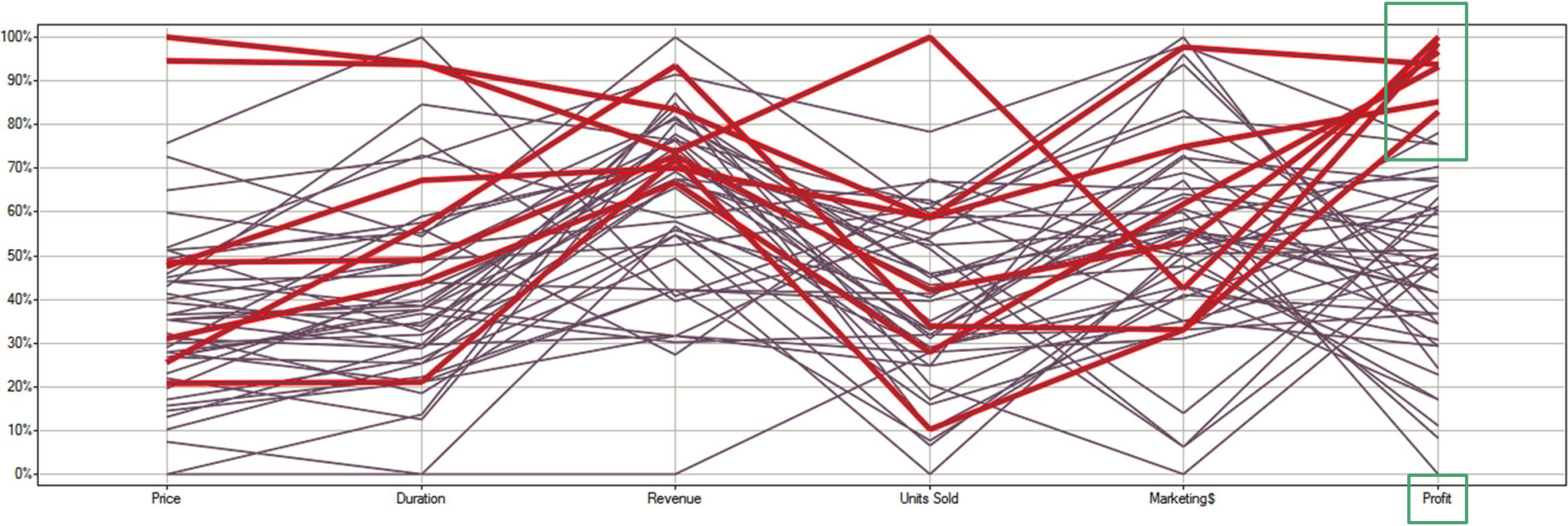
Parallel coordinates plot

A more detailed understanding requires interactive resources. If we want to identify which multivariate profiles correspond to high profits:

- We place the variable of interest last;
- We highlight products with a profit greater than or equal to 80%.

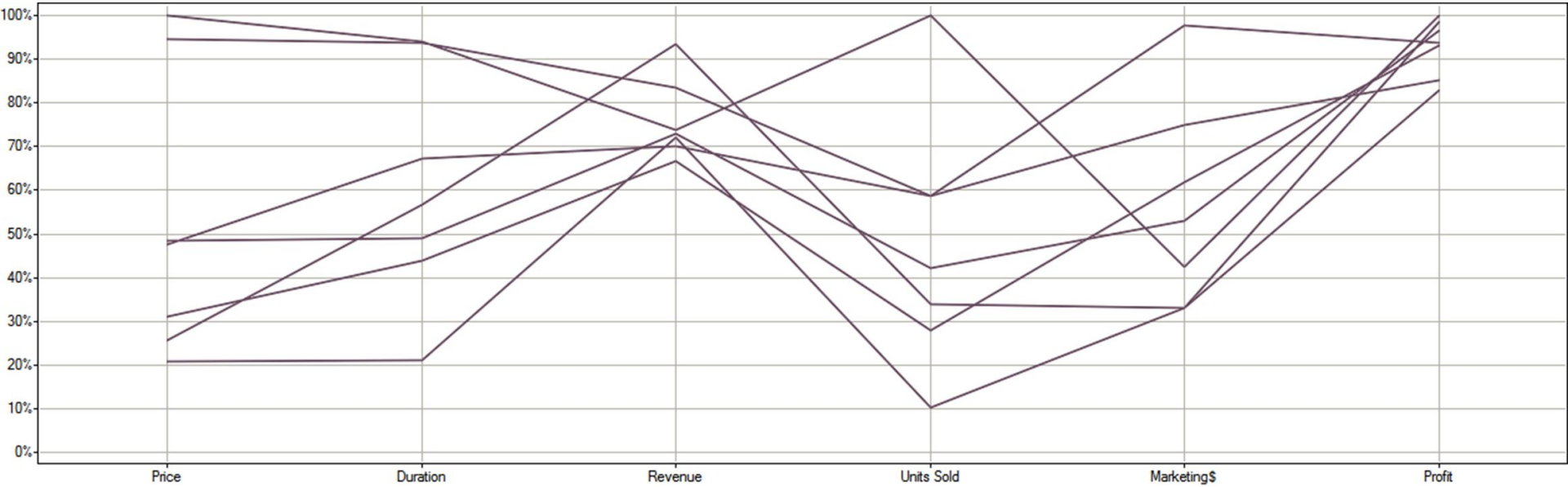
Parallel coordinates plot

A more detailed understanding requires interactive resources.



Parallel coordinates plot

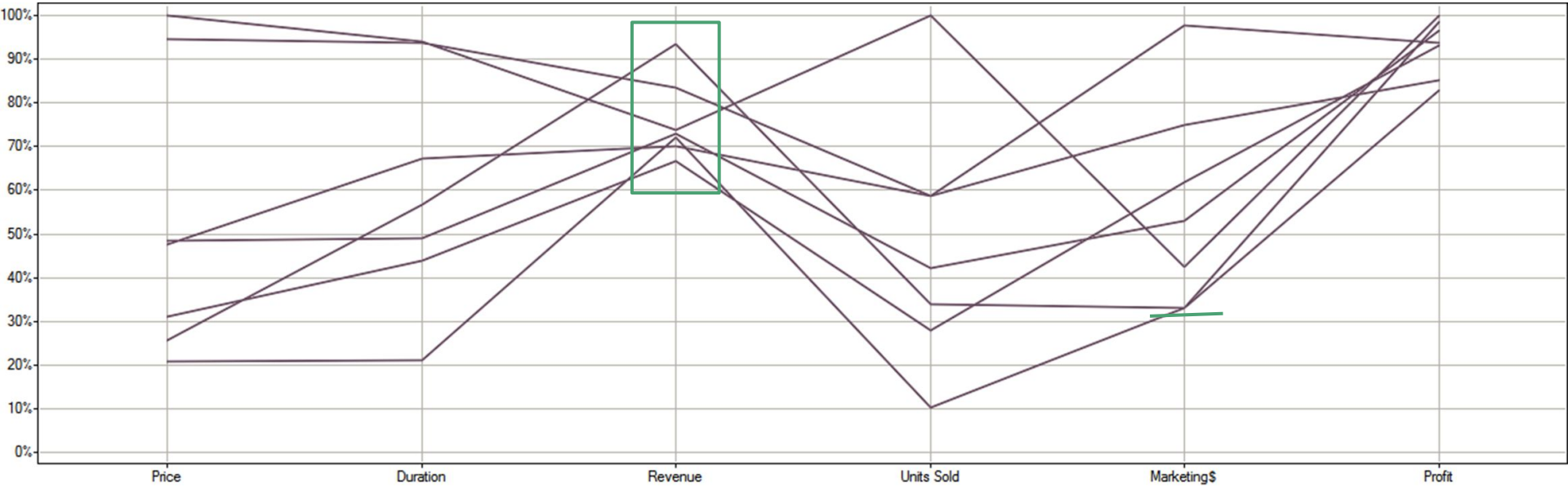
Only products with a profit> = 80% are displayed here.



Parallel coordinates plot

High-profit products have high revenue

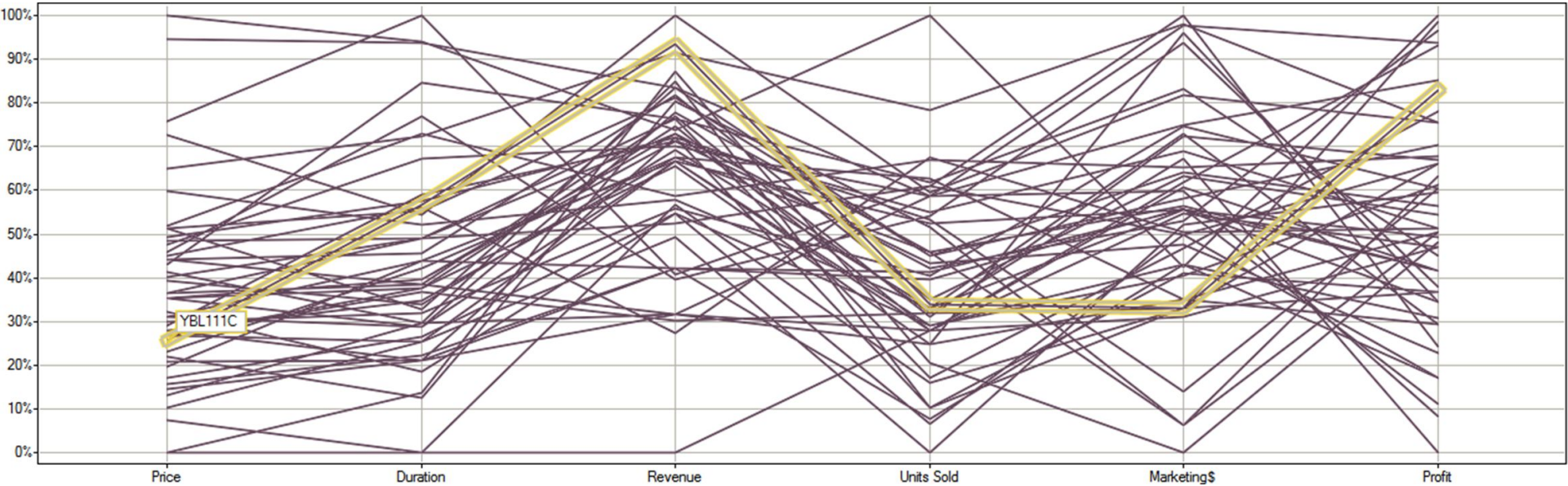
Marketing expenses are over 30%



Multivariate analysis techniques and best practices

Ranking by similarity

In the previous example we could be interested in products with relatively low prices, that are on average in terms of time in the market, that have high revenue, with the sale of few units, marketing expenses below average and high profit.

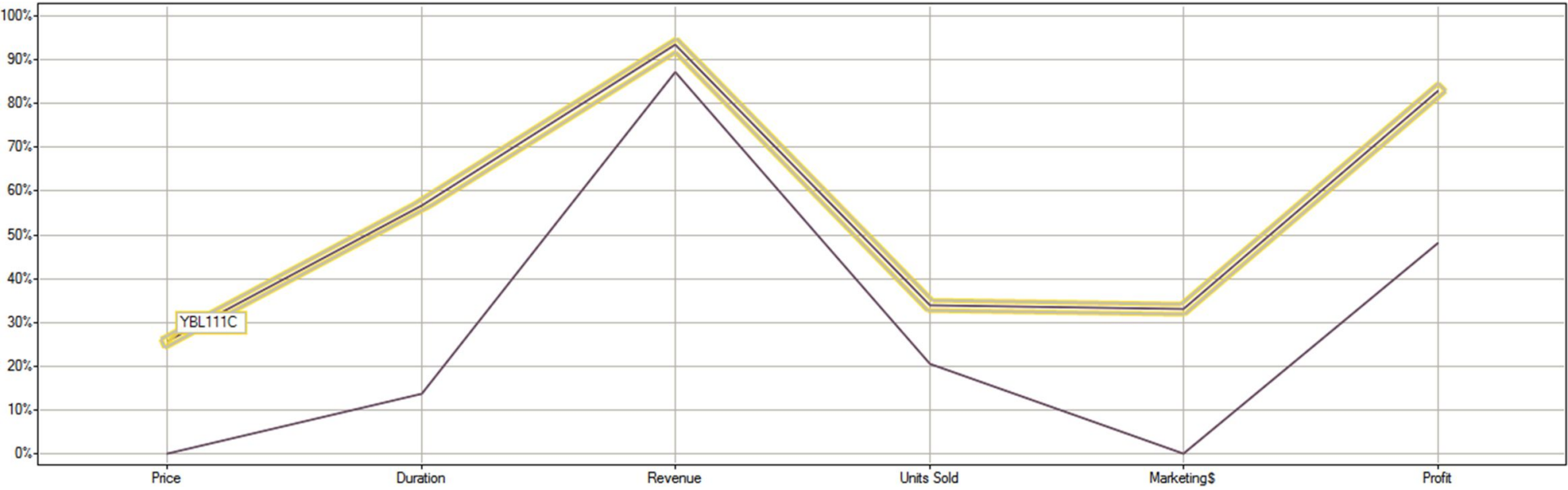


Multivariate analysis techniques and best practices

Ranking by similarity

How to search for products with these characteristics?

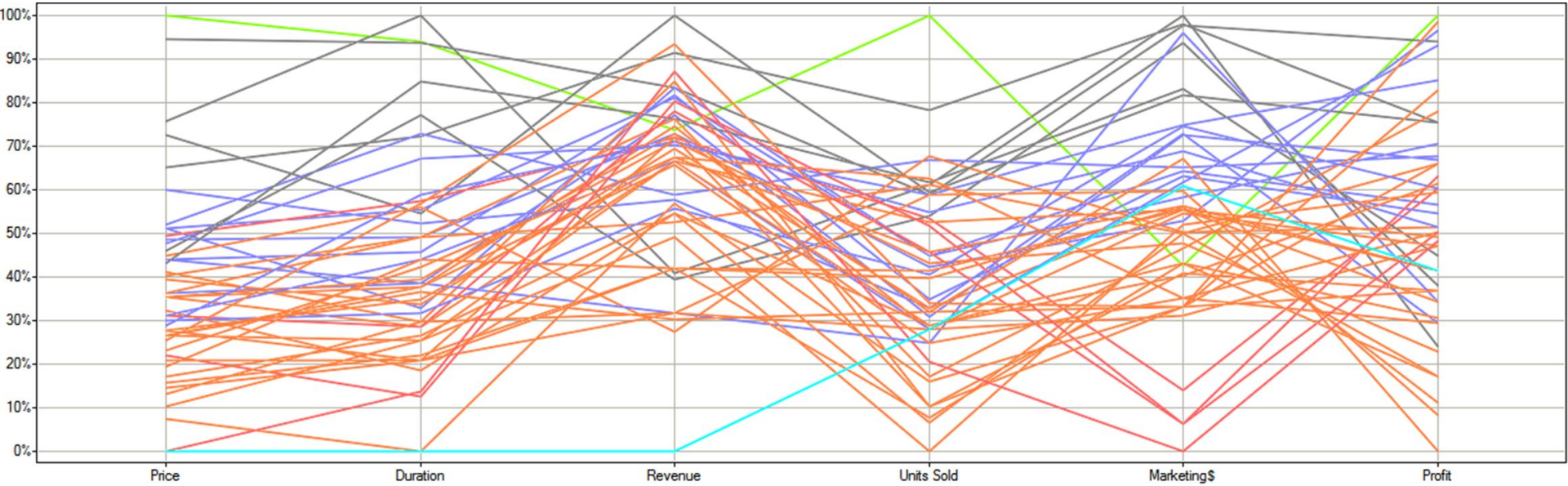
Here the profile most similar to the original was selected.



Multivariate analysis techniques and best practices

Clustering by similarity

Grouping the data from our example, we have the following result:



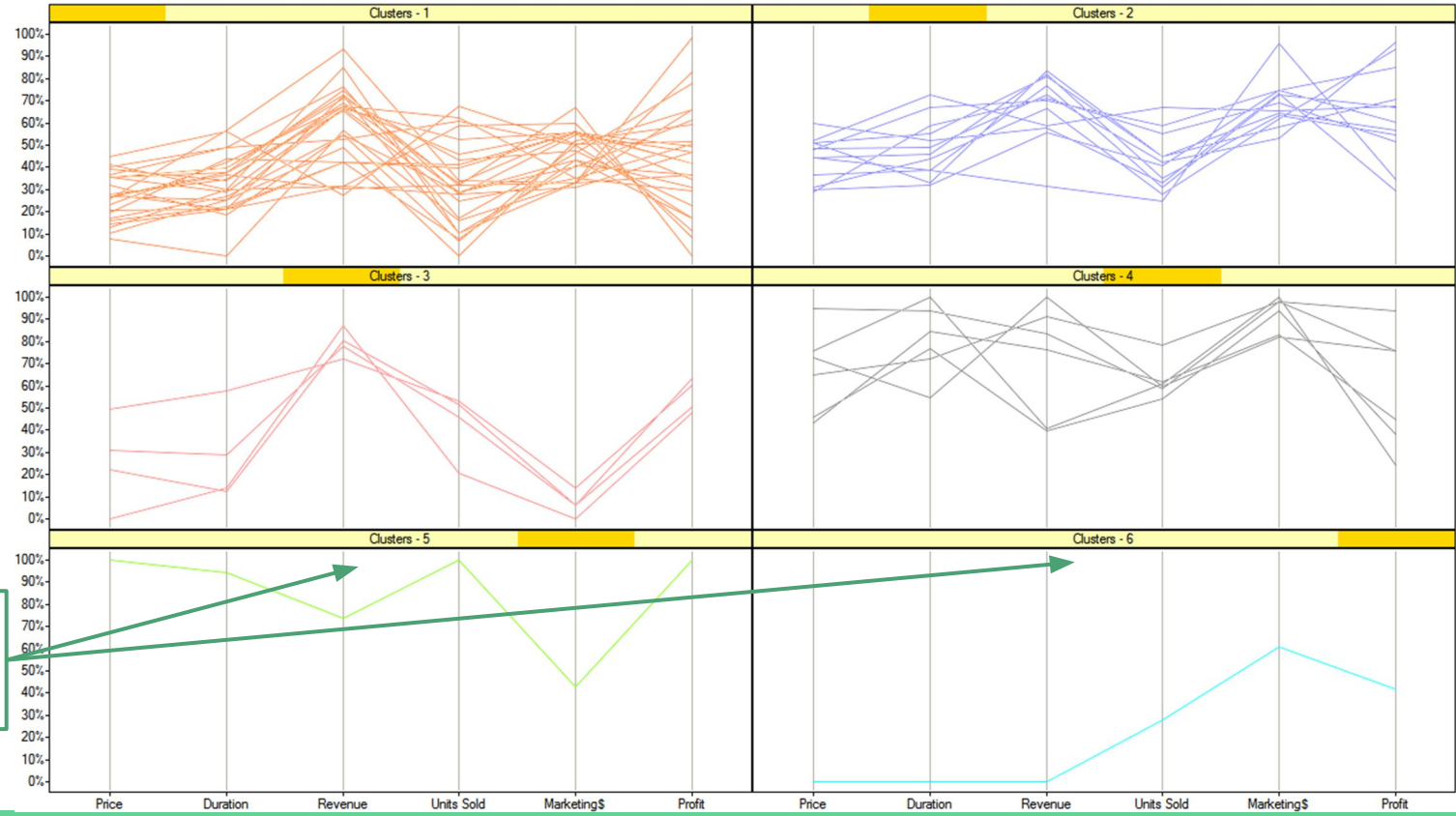
Multivariate analysis techniques and best practices

Clustering by similarity

Separating groups on a Trellis display to improve visualization ...

Multivariate analysis techniques and best practices

Clustering by similarity



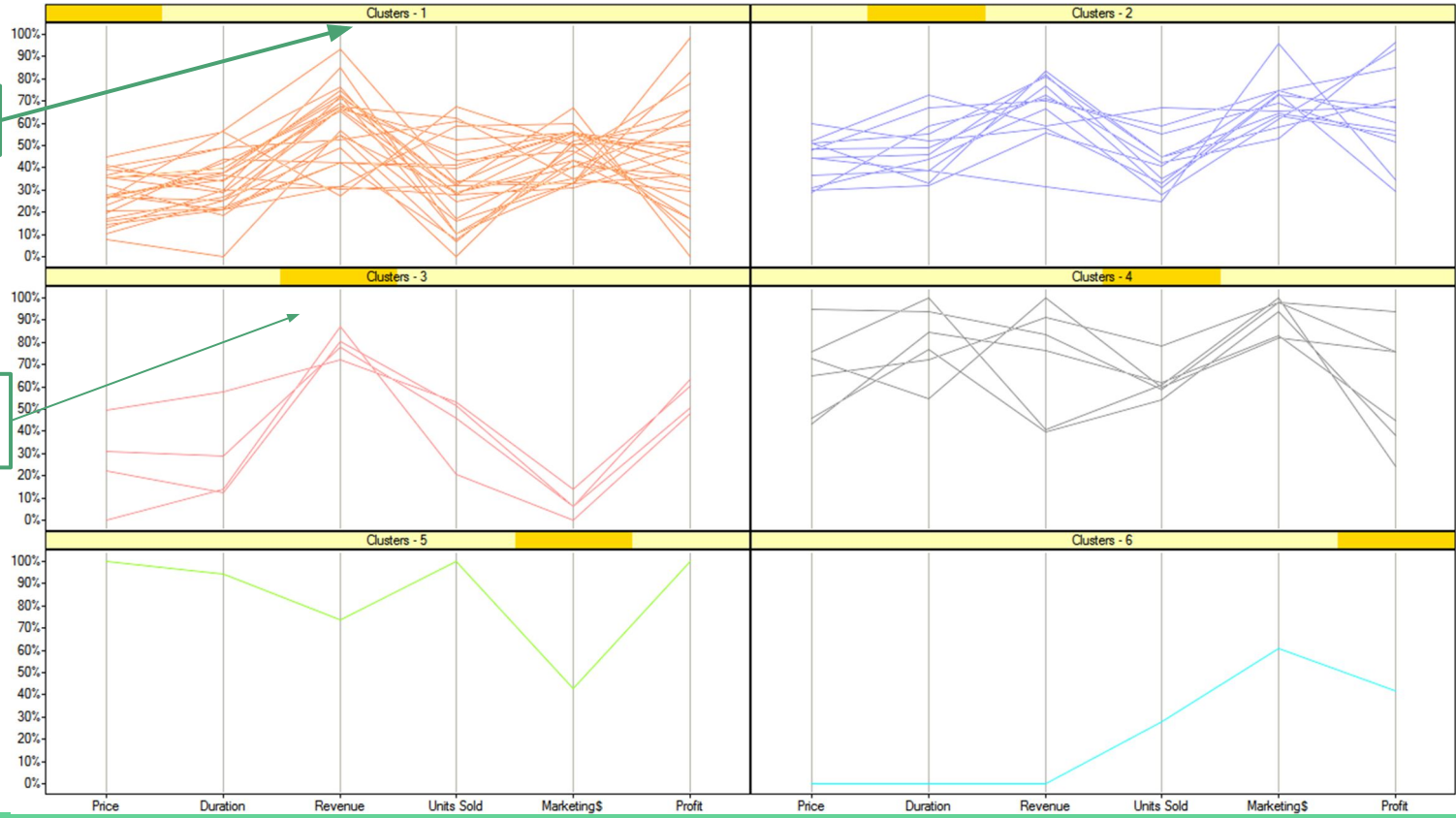
Clusters 5 and 6 have profiles that do not look like the others.

Multivariate analysis techniques and best practices

Clustering by similarity

25 products in cluster 1

Very similar products in group 3



Multivariate analysis techniques and best practices

Clustering by similarity

There is a huge variety of clustering algorithms, such as k-means, k-fearids, DBScan, etc. It is worth reading about each one's strengths and weaknesses to select the best option.

Further reading

Books

Now You See It: Simple Visualization Techniques for Quantitative Analysis

https://www.amazon.com.br/Now-You-See-Visualization-Quantitative/dp/0970601980

The Visual Display of Quantitative Information

https://www.amazon.com.br/Visual-Display-Quantitative-Information/dp/0961392142

Further reading

methagora
a blog from *Nature Methods*

Go
Advanced search

Nature Methods › Blog › Post

Previous post
Promoting shared hardware design

Next post
Importance of data sharing

NATURE METHODS | METHAGORA

Data visualization: A view of every Points of View column

30 Jul 2013 | 8:08 AM | Posted by Daniel Evanko | Category: Featured, Visualization

We've organized all the Points of View columns on data visualization published in *Nature Methods* and provide this as a guide to accessing this trove of practical advice on visualizing scientific data.

As of July 30, 2013 *Nature Methods* has published 35 Points of View columns written by Bang Wong, Martin Krzywinski and their co-authors: Nils Gehlenborg, Cydney Nielsen, Noam Shoresh, Rikke Schmidt Kjærgaard, Erica Savig and Alberto Cairo. As we prepare to launch a new column in our September issue we felt this would be a good time to collect and organize links to all the Points of View articles together in one place to make it easier to navigate this wonderful resource that the authors have provided us. For the month of August we will be making all the columns free to access so everyone can benefit from this practical advice on data visualization.

This should not be the end of the Points of View column though. We will be inviting new visualization experts to author articles on new topics that have not been covered so far or which can be expanded on. This page will be continuously updated whenever a new article is published so stay tuned. If you have a suggestion for a topic you would like to see covered in a future points of view article please comment below.

Update of March 28, 2015: A PDF eBook of the 38 Points of View articles published between August 2010 and February 2015 is now available at the Nature Shop for $7.99 under the title "Visual strategies for biological data: the collected Points of View". The article summaries below provide a nice overview of what is

Current issue
April 2016, Volume 13 No 4

- Journal home
- About
- Current issue
- Subscribe
- Recommend to library

E-alert | RSS | Twitter

nature.com blogs home

Featured posts from this blog

Cell Biology, Editorials, Featured, Genetics & Genomics, Method of the Year
Method of the Year 2016
add a comment

Editorials, Featured, General Interest, Journal happenings, Nature Methods papers
Ten years of Methods
add a comment

Computational, Editorials, Featured, Journal Policy
Guidelines for algorithms and software in Nature Methods
2 Comments

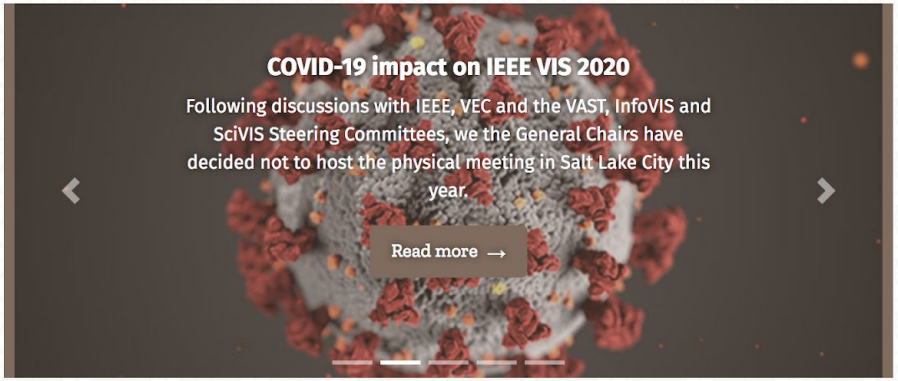
Featured, General Interest, Journal Policy
The dos and don'ts of communicating with editors

http://blogs.nature.com/methagora/2013/07/data-visualization-points-of-view.html

Conferences

VIS 2020

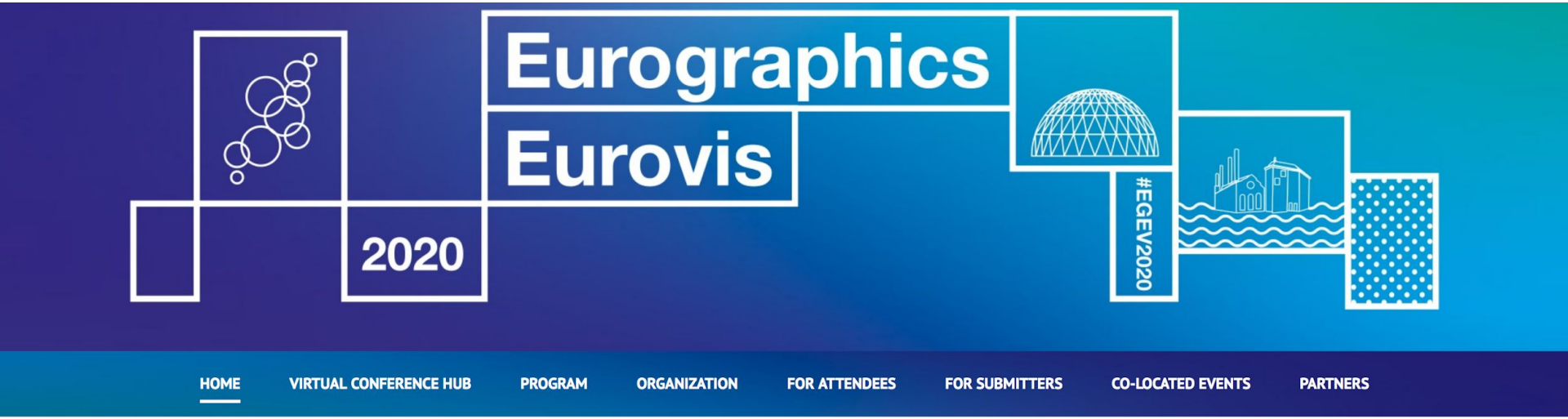
the premier forum for advances in visualization and visual analytics



VIS | VAST • INFOVIS • SCIVIS

25-30 October 2020 in Virtual Conference

Conferences



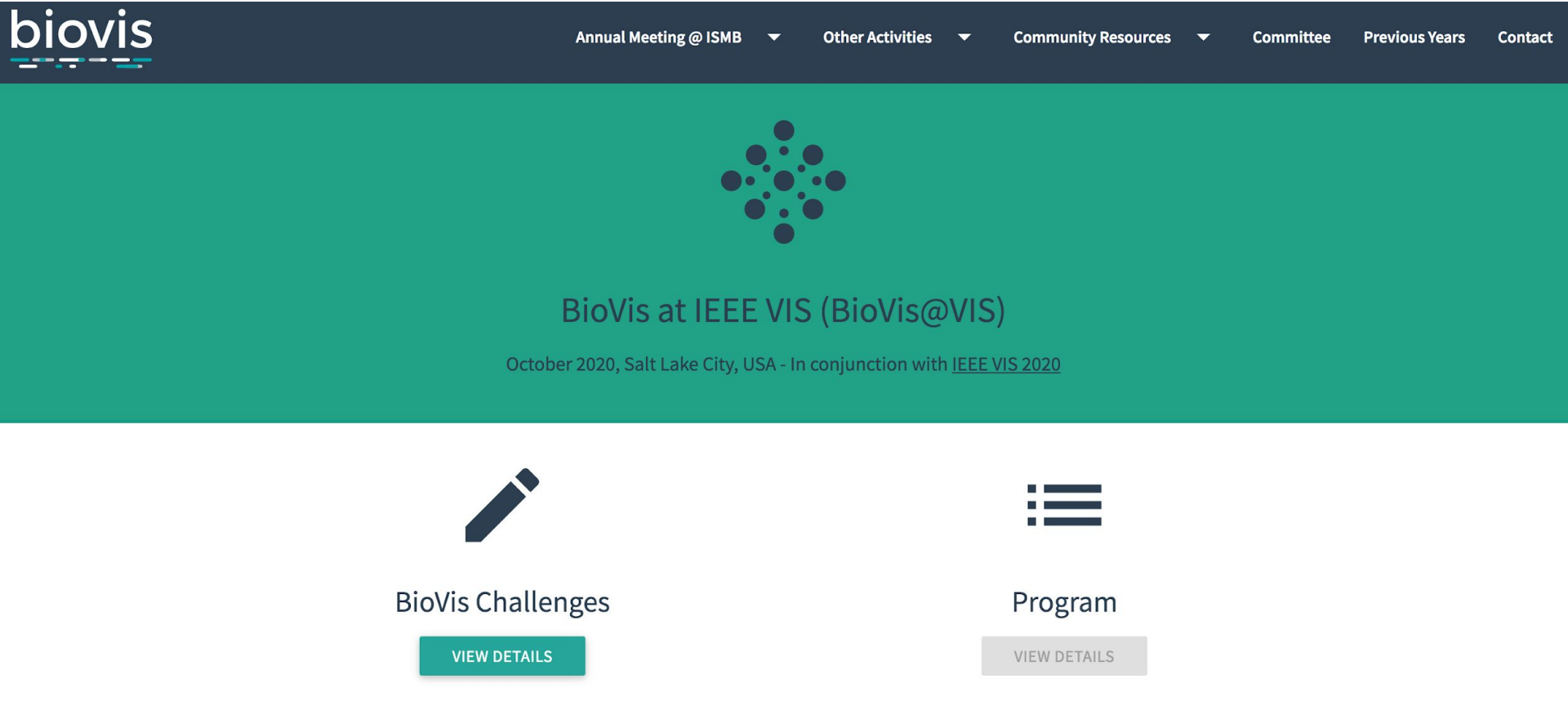
Welcome to the joint conferences, Eurographics & Eurovis 2020

Norrköping, Sweden, May 25-29, 2020

News

- Virtual Posters Event goes live
  May 27, 2020
- Note from the Chairs
  May 24, 2020
- Papers of EGEV2020

Conferences



Further reading

Color blind simulator:

https://www.makeuseof.com/tag/3-easytouse-online-colorblindness-simulators/

https://www.color-blindness.com/coblis-color-blindness-simulator/

https://www.toptal.com/designers/colorfilter/