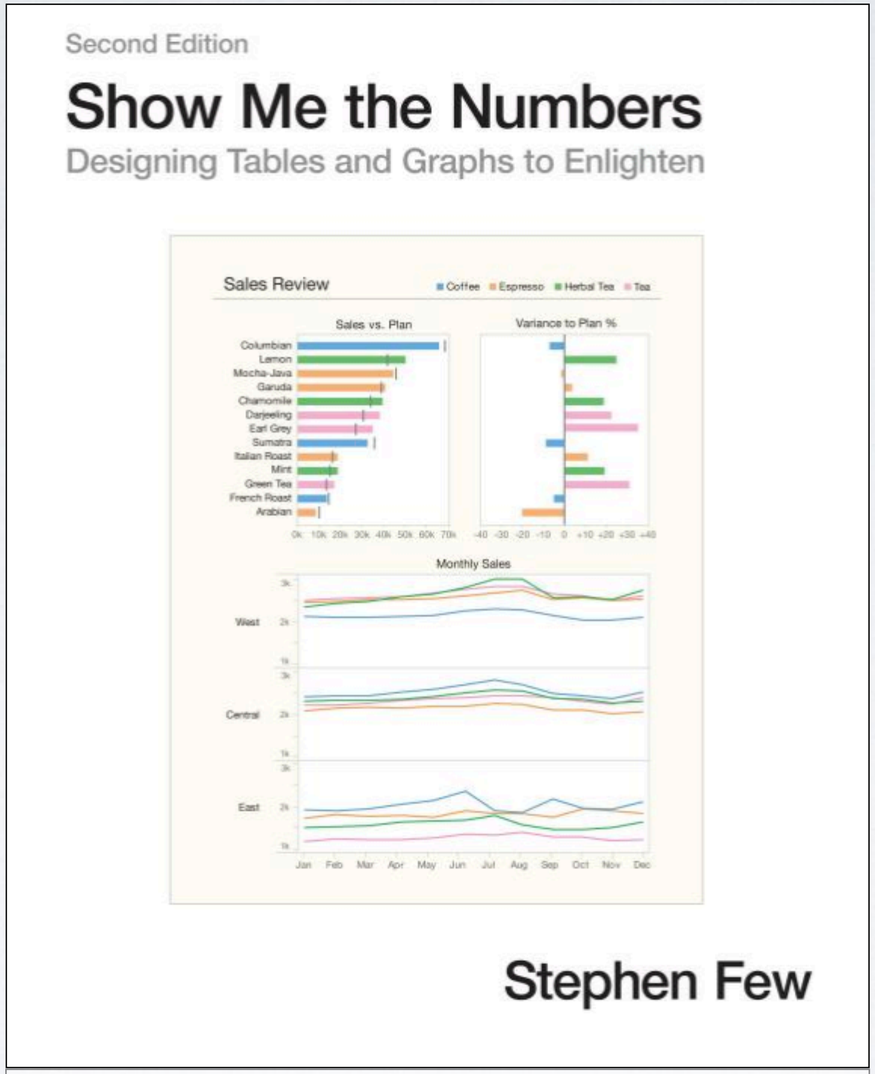
Basics of tables and graphs

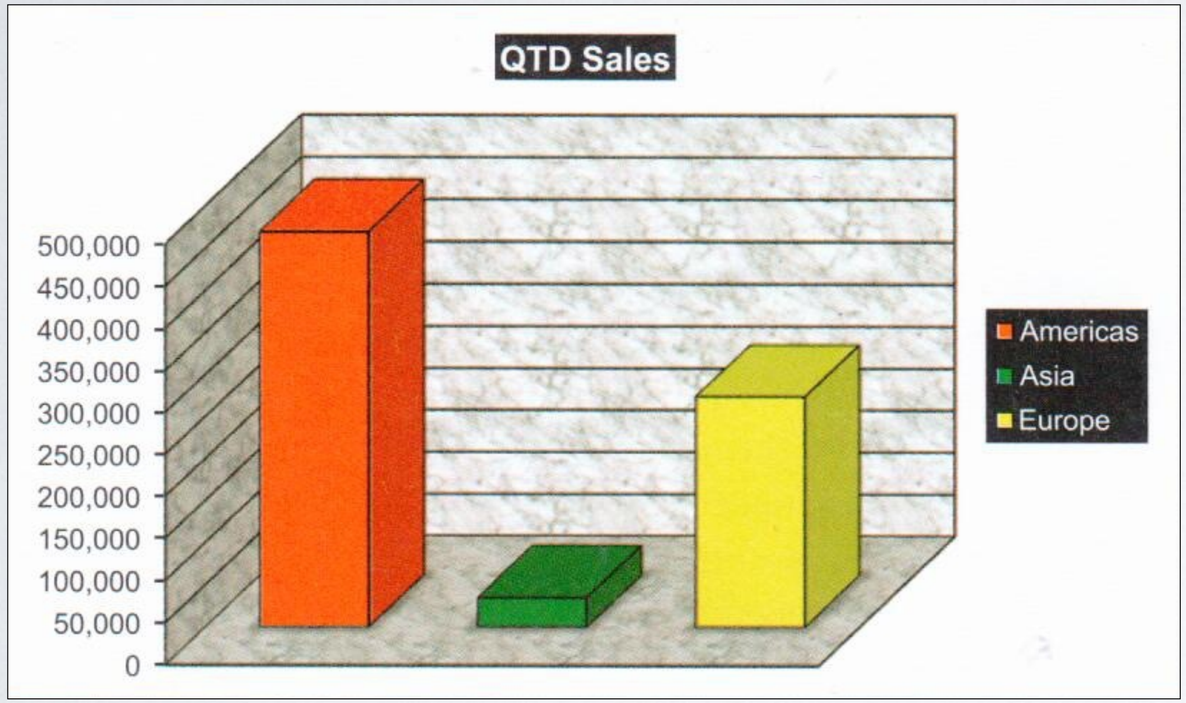
Marco D'Ambros

SHOW ME THE NUMBERS!



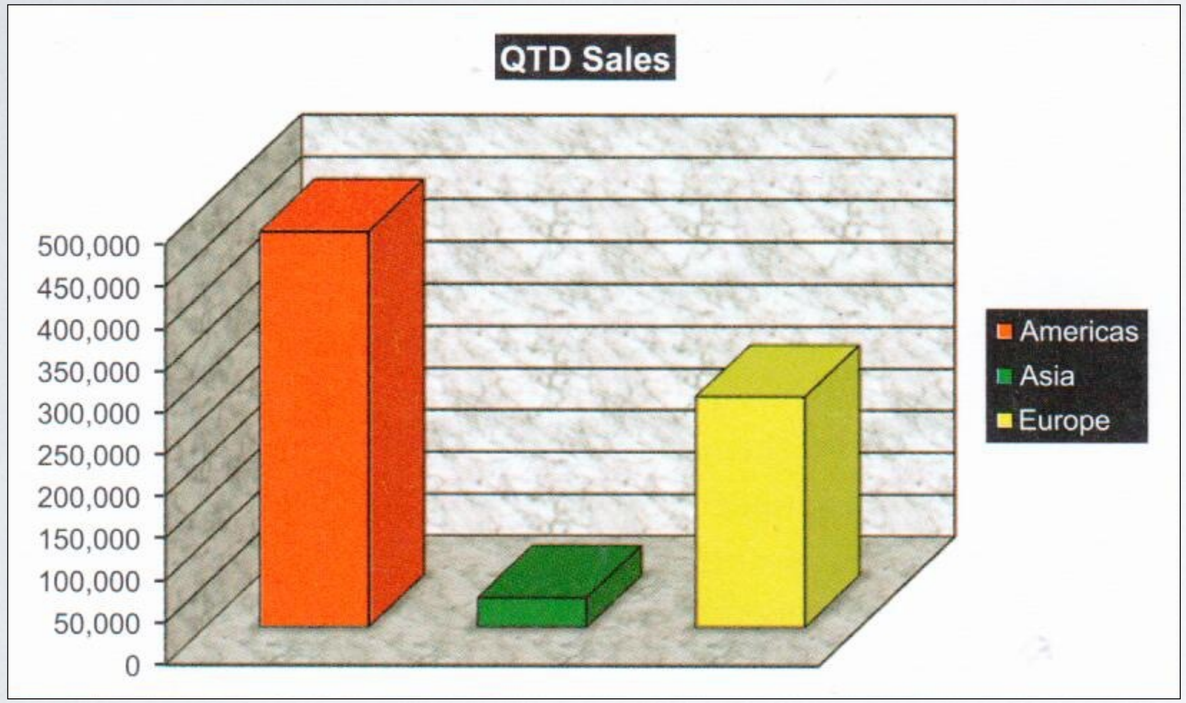
INEFFECTIVE GRAPHS EXAMPLES

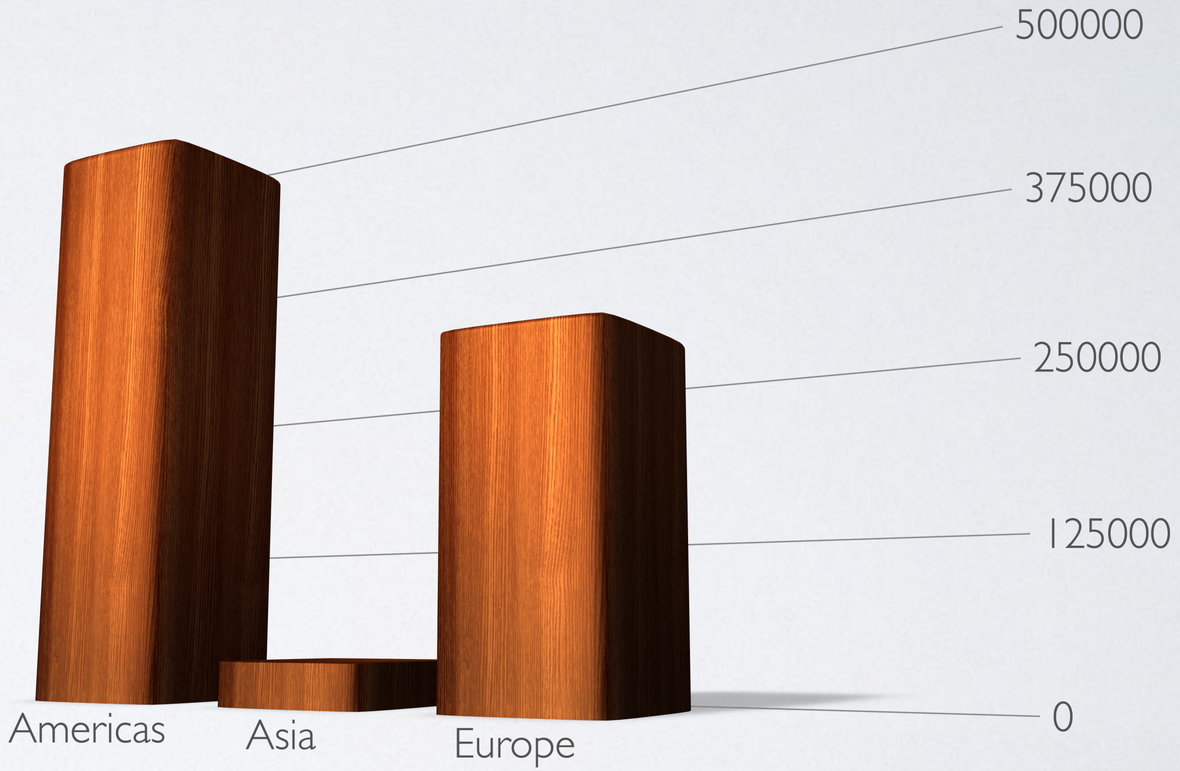
Goal: Inform an executive about current state of quarterly sale, split by regions



INEFFECTIVE GRAPHS EXAMPLES

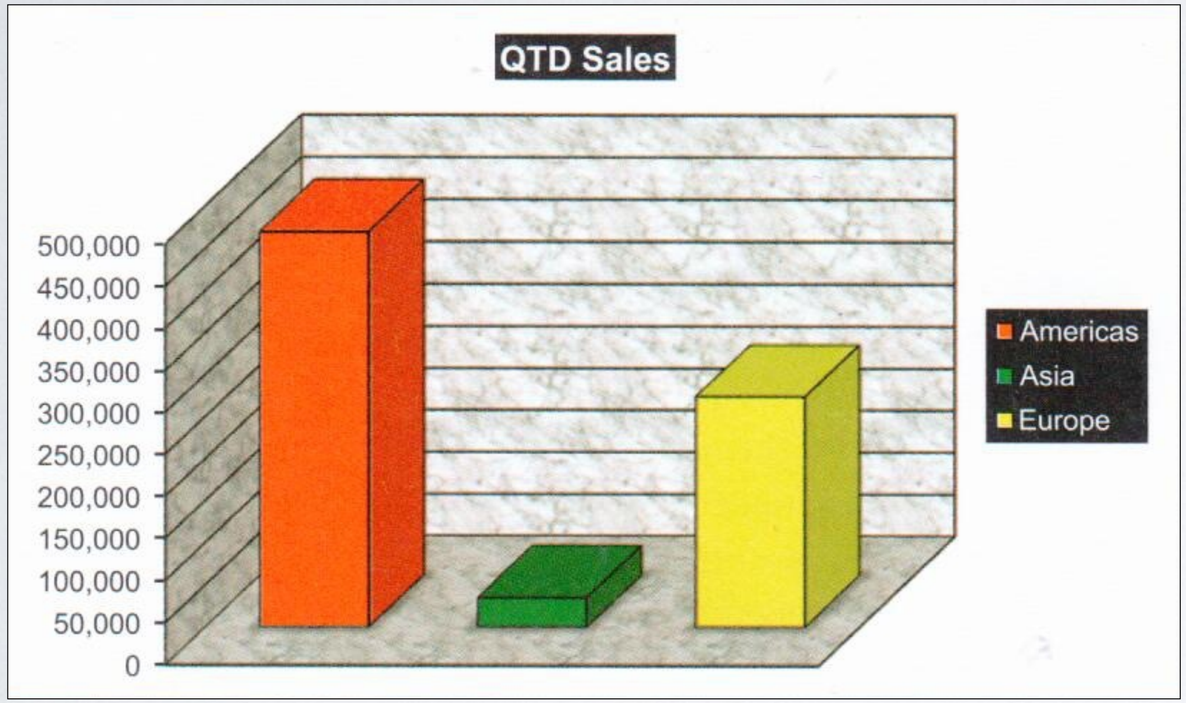
Goal: Inform an executive about current state of quarterly sale, split by regions





INEFFECTIVE GRAPHS EXAMPLES

Goal: Inform an executive about current state of quarterly sale, split by regions

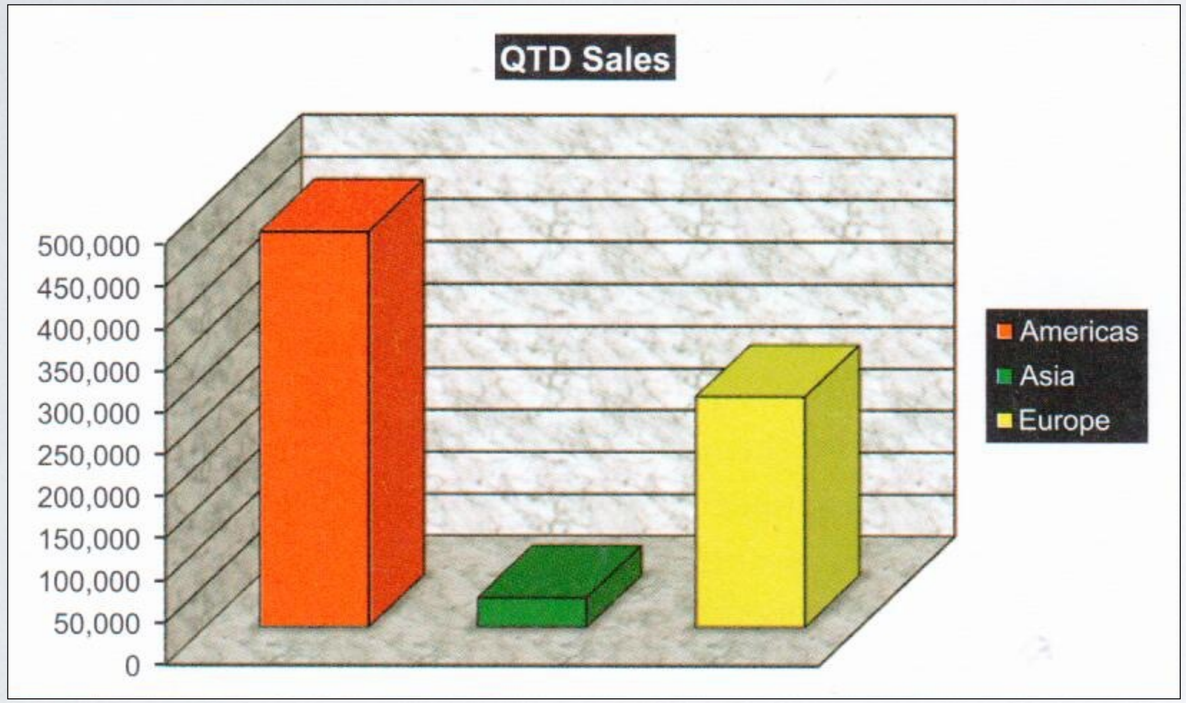


QTD Sales

Americas	$469,384
Europe	$273,854
Asia	$34,847

INEFFECTIVE GRAPHS EXAMPLES

Goal: Inform an executive about current state of quarterly sale, split by regions



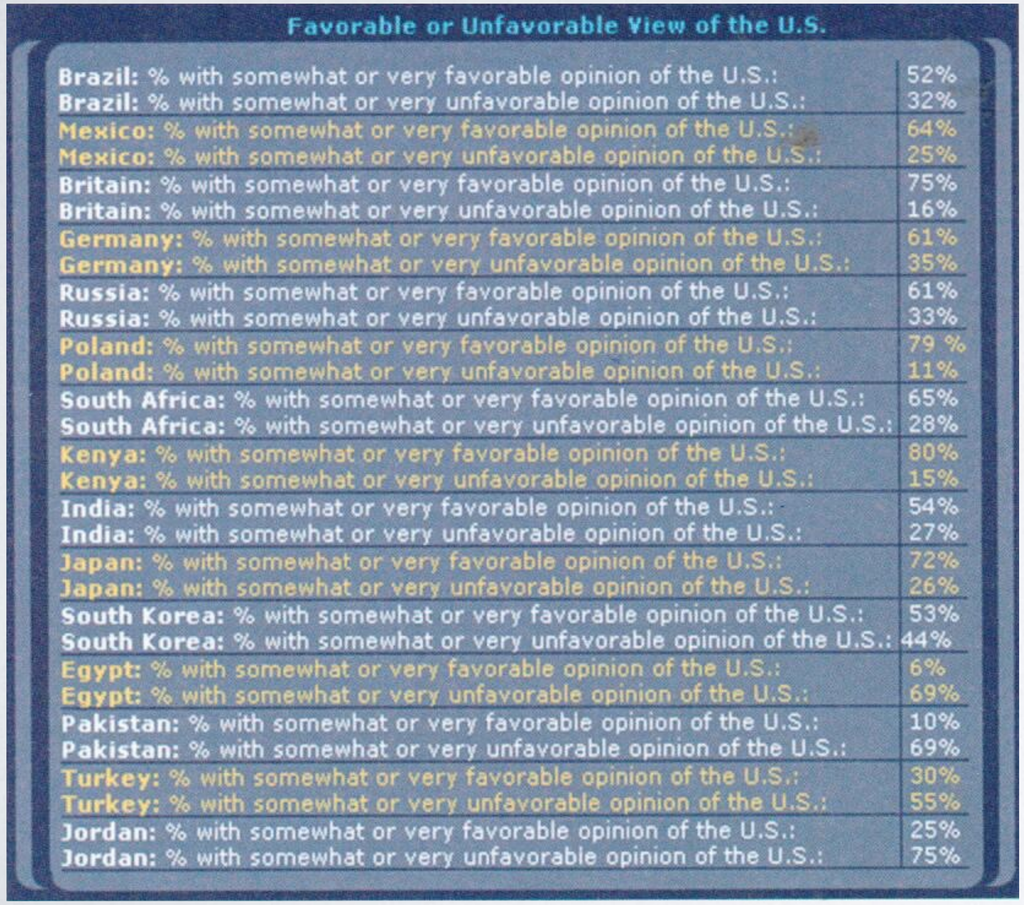
2011 Q1-to-Date Regional Sales
As of March 15, 2011

	Current			Qtr End	
	Sales (U.S. $)	Percent of Total Sales	Percent of Qtr Plan	Projected Sales (U.S. $)	Projected Percent of Qtr Plan
Americas	469,384	60%	85%	586,730	107%
Europe	273,854	35%	91%	353,272	118%
Asia	34,847	5%	50%	43,210	62%
	$778,085	100%	85%	$983,212	107%

Note: To date, 82% of the quarter has elapsed.

INEFFECTIVE GRAPHS EXAMPLES

Goal: which countries are favourable vs unfavourable to the U.S.?



Favorable or Unfavorable View of the U.S.

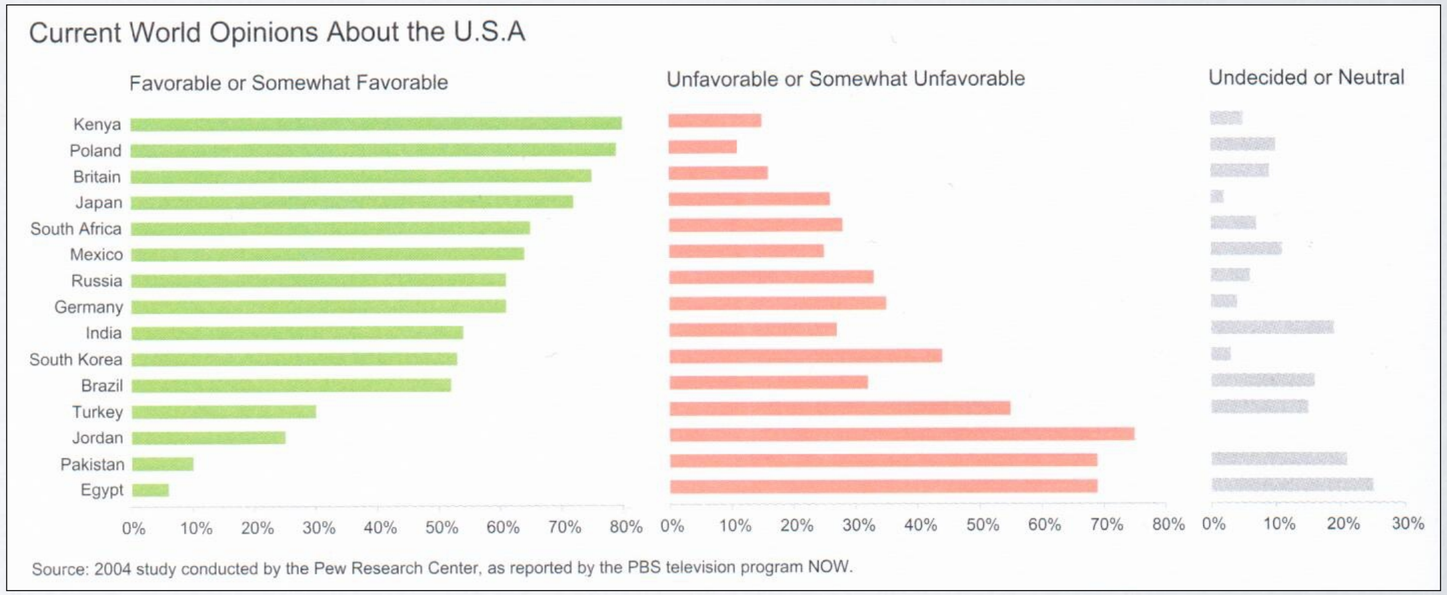
Brazil: % with somewhat or very favorable opinion of the U.S.:	52%
Brazil: % with somewhat or very unfavorable opinion of the U.S.:	32%
Mexico: % with somewhat or very favorable opinion of the U.S.:	64%
Mexico: % with somewhat or very unfavorable opinion of the U.S.:	25%
Britain: % with somewhat or very favorable opinion of the U.S.:	75%
Britain: % with somewhat or very unfavorable opinion of the U.S.:	16%
Germany: % with somewhat or very favorable opinion of the U.S.:	61%
Germany: % with somewhat or very unfavorable opinion of the U.S.:	35%
Russia: % with somewhat or very favorable opinion of the U.S.:	61%
Russia: % with somewhat or very unfavorable opinion of the U.S.:	33%
Poland: % with somewhat or very favorable opinion of the U.S.:	79 %
Poland: % with somewhat or very unfavorable opinion of the U.S.:	11%
South Africa: % with somewhat or very favorable opinion of the U.S.:	65%
South Africa: % with somewhat or very unfavorable opinion of the U.S.:	28%
Kenya: % with somewhat or very favorable opinion of the U.S.:	80%
Kenya: % with somewhat or very unfavorable opinion of the U.S.:	15%
India: % with somewhat or very favorable opinion of the U.S.:	54%
India: % with somewhat or very unfavorable opinion of the U.S.:	27%
Japan: % with somewhat or very favorable opinion of the U.S.:	72%
Japan: % with somewhat or very unfavorable opinion of the U.S.:	26%
South Korea: % with somewhat or very favorable opinion of the U.S.:	53%
South Korea: % with somewhat or very unfavorable opinion of the U.S.:	44%
Egypt: % with somewhat or very favorable opinion of the U.S.:	6%
Egypt: % with somewhat or very unfavorable opinion of the U.S.:	69%
Pakistan: % with somewhat or very favorable opinion of the U.S.:	10%
Pakistan: % with somewhat or very unfavorable opinion of the U.S.:	69%
Turkey: % with somewhat or very favorable opinion of the U.S.:	30%
Turkey: % with somewhat or very unfavorable opinion of the U.S.:	55%
Jordan: % with somewhat or very favorable opinion of the U.S.:	25%
Jordan: % with somewhat or very unfavorable opinion of the U.S.:	75%

INEFFECTIVE GRAPHS EXAMPLES

Goal: which countries are favourable vs unfavourable to the U.S.?

Favorable or Unfavorable View of the U.S.

Brazil: % with somewhat or very favorable opinion of the U.S.:	52%
Brazil: % with somewhat or very unfavorable opinion of the U.S.:	32%
Mexico: % with somewhat or very favorable opinion of the U.S.:	64%
Mexico: % with somewhat or very unfavorable opinion of the U.S.:	25%
Britain: % with somewhat or very favorable opinion of the U.S.:	75%
Britain: % with somewhat or very unfavorable opinion of the U.S.:	16%
Germany: % with somewhat or very favorable opinion of the U.S.:	61%
Germany: % with somewhat or very unfavorable opinion of the U.S.:	35%
Russia: % with somewhat or very favorable opinion of the U.S.:	61%
Russia: % with somewhat or very unfavorable opinion of the U.S.:	33%
Poland: % with somewhat or very favorable opinion of the U.S.:	79 %
Poland: % with somewhat or very unfavorable opinion of the U.S.:	11%
South Africa: % with somewhat or very favorable opinion of the U.S.:	65%
South Africa: % with somewhat or very unfavorable opinion of the U.S.:	28%
Kenya: % with somewhat or very favorable opinion of the U.S.:	80%
Kenya: % with somewhat or very unfavorable opinion of the U.S.:	15%
India: % with somewhat or very favorable opinion of the U.S.:	54%
India: % with somewhat or very unfavorable opinion of the U.S.:	27%
Japan: % with somewhat or very favorable opinion of the U.S.:	72%
Japan: % with somewhat or very unfavorable opinion of the U.S.:	26%
South Korea: % with somewhat or very favorable opinion of the U.S.:	53%
South Korea: % with somewhat or very unfavorable opinion of the U.S.:	44%
Egypt: % with somewhat or very favorable opinion of the U.S.:	6%
Egypt: % with somewhat or very unfavorable opinion of the U.S.:	69%
Pakistan: % with somewhat or very favorable opinion of the U.S.:	10%
Pakistan: % with somewhat or very unfavorable opinion of the U.S.:	69%
Turkey: % with somewhat or very favorable opinion of the U.S.:	30%
Turkey: % with somewhat or very unfavorable opinion of the U.S.:	55%
Jordan: % with somewhat or very favorable opinion of the U.S.:	25%
Jordan: % with somewhat or very unfavorable opinion of the U.S.:	75%



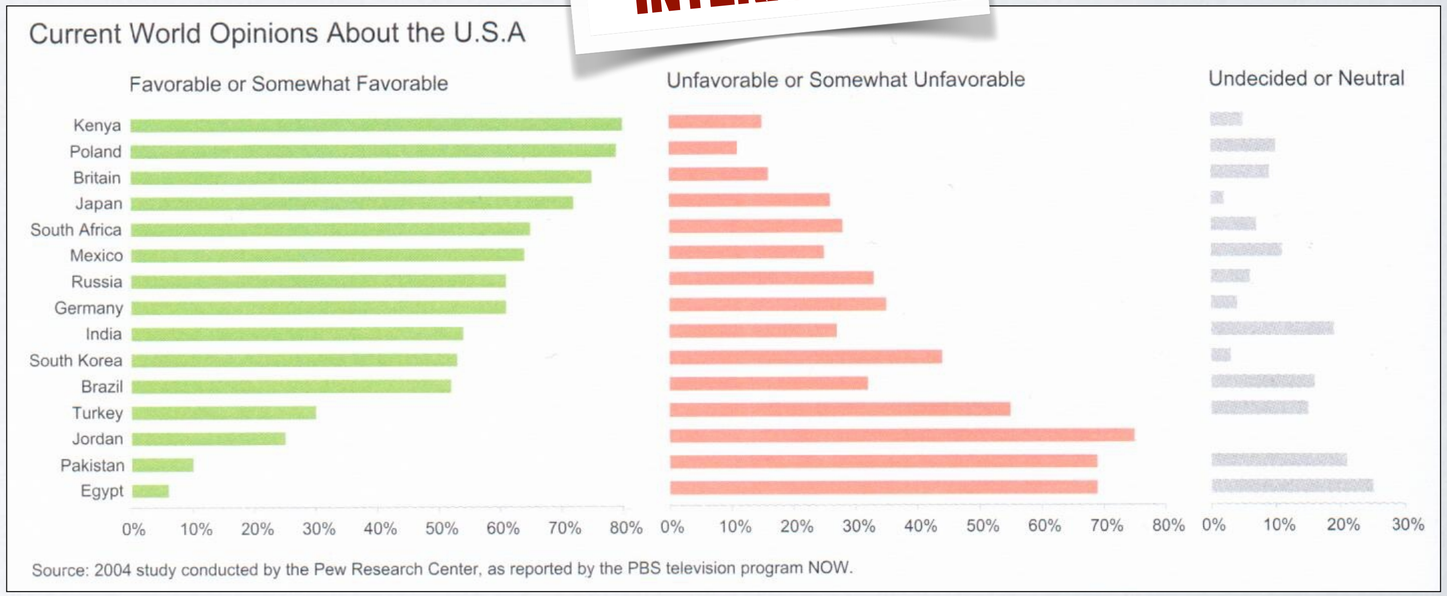
INEFFECTIVE GRAPHS EXAMPLES

Goal: which countries are favourable vs unfavourable to the U.S.?

Favorable or Unfavorable View of the U.S.

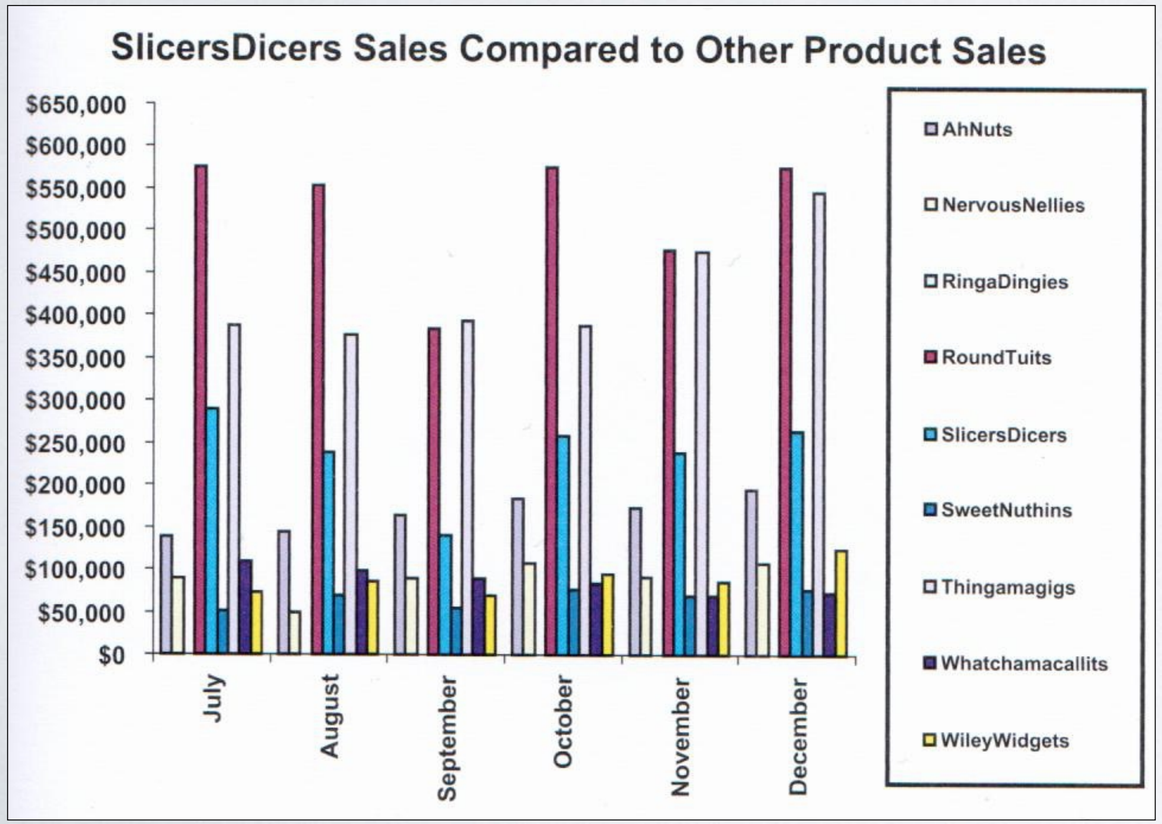
Brazil: % with somewhat or very favorable opinion of the U.S.:	52%
Brazil: % with somewhat or very unfavorable opinion of the U.S.:	32%
Mexico: % with somewhat or very favorable opinion of the U.S.:	64%
Mexico: % with somewhat or very unfavorable opinion of the U.S.:	25%
Britain: % with somewhat or very favorable opinion of the U.S.:	75%
Britain: % with somewhat or very unfavorable opinion of the U.S.:	16%
Germany: % with somewhat or very favorable opinion of the U.S.:	61%
Germany: % with somewhat or very unfavorable opinion of the U.S.:	35%
Russia: % with somewhat or very favorable opinion of the U.S.:	61%
Russia: % with somewhat or very unfavorable opinion of the U.S.:	33%
Poland: % with somewhat or very favorable opinion of the U.S.:	79 %
Poland: % with somewhat or very unfavorable opinion of the U.S.:	11%
South Africa: % with somewhat or very favorable opinion of the U.S.:	65%
South Africa: % with somewhat or very unfavorable opinion of the U.S.:	28%
Kenya: % with somewhat or very favorable opinion of the U.S.:	80%
Kenya: % with somewhat or very unfavorable opinion of the U.S.:	15%
India: % with somewhat or very favorable opinion of the U.S.:	54%
India: % with somewhat or very unfavorable opinion of the U.S.:	27%
Japan: % with somewhat or very favorable opinion of the U.S.:	72%
Japan: % with somewhat or very unfavorable opinion of the U.S.:	26%
South Korea: % with somewhat or very favorable opinion of the U.S.:	53%
South Korea: % with somewhat or very unfavorable opinion of the U.S.:	44%
Egypt: % with somewhat or very favorable opinion of the U.S.:	6%
Egypt: % with somewhat or very unfavorable opinion of the U.S.:	69%
Pakistan: % with somewhat or very favorable opinion of the U.S.:	10%
Pakistan: % with somewhat or very unfavorable opinion of the U.S.:	69%
Turkey: % with somewhat or very favorable opinion of the U.S.:	30%
Turkey: % with somewhat or very unfavorable opinion of the U.S.:	55%
Jordan: % with somewhat or very favorable opinion of the U.S.:	25%
Jordan: % with somewhat or very unfavorable opinion of the U.S.:	75%

INTERACTIVE



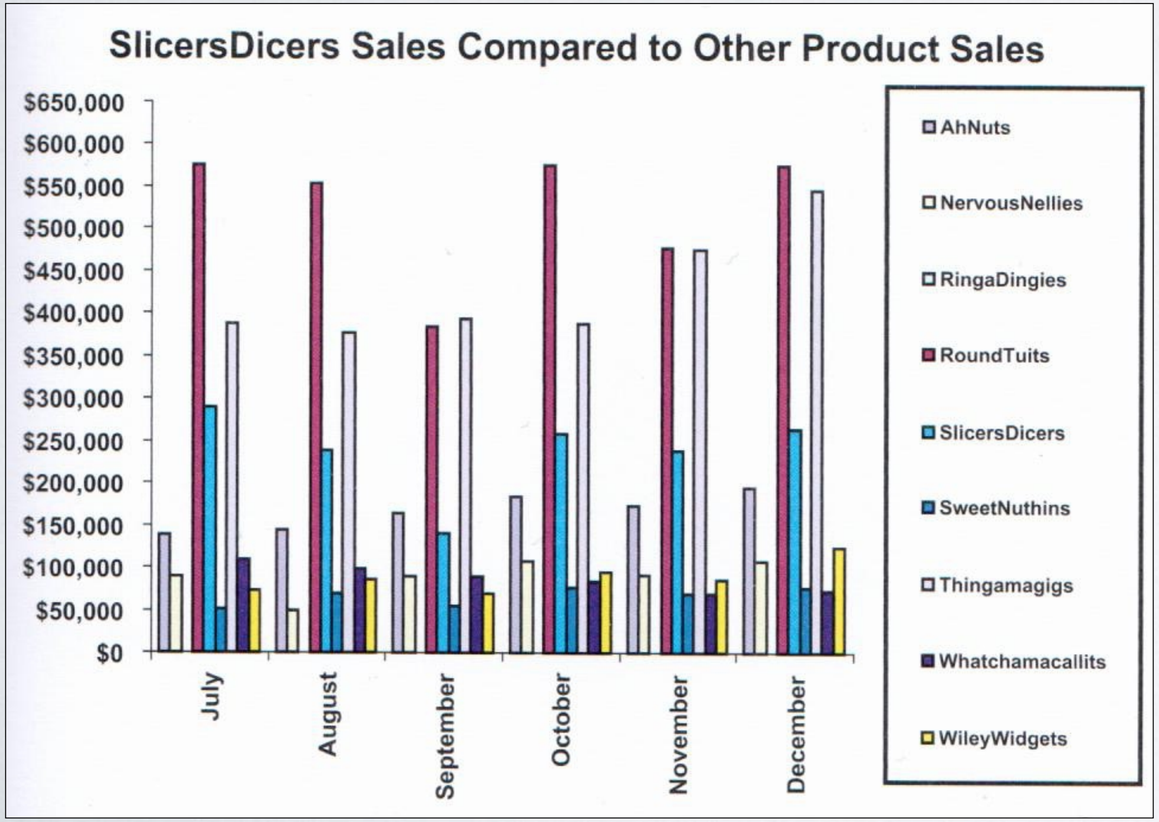
INEFFECTIVE GRAPHS EXAMPLES

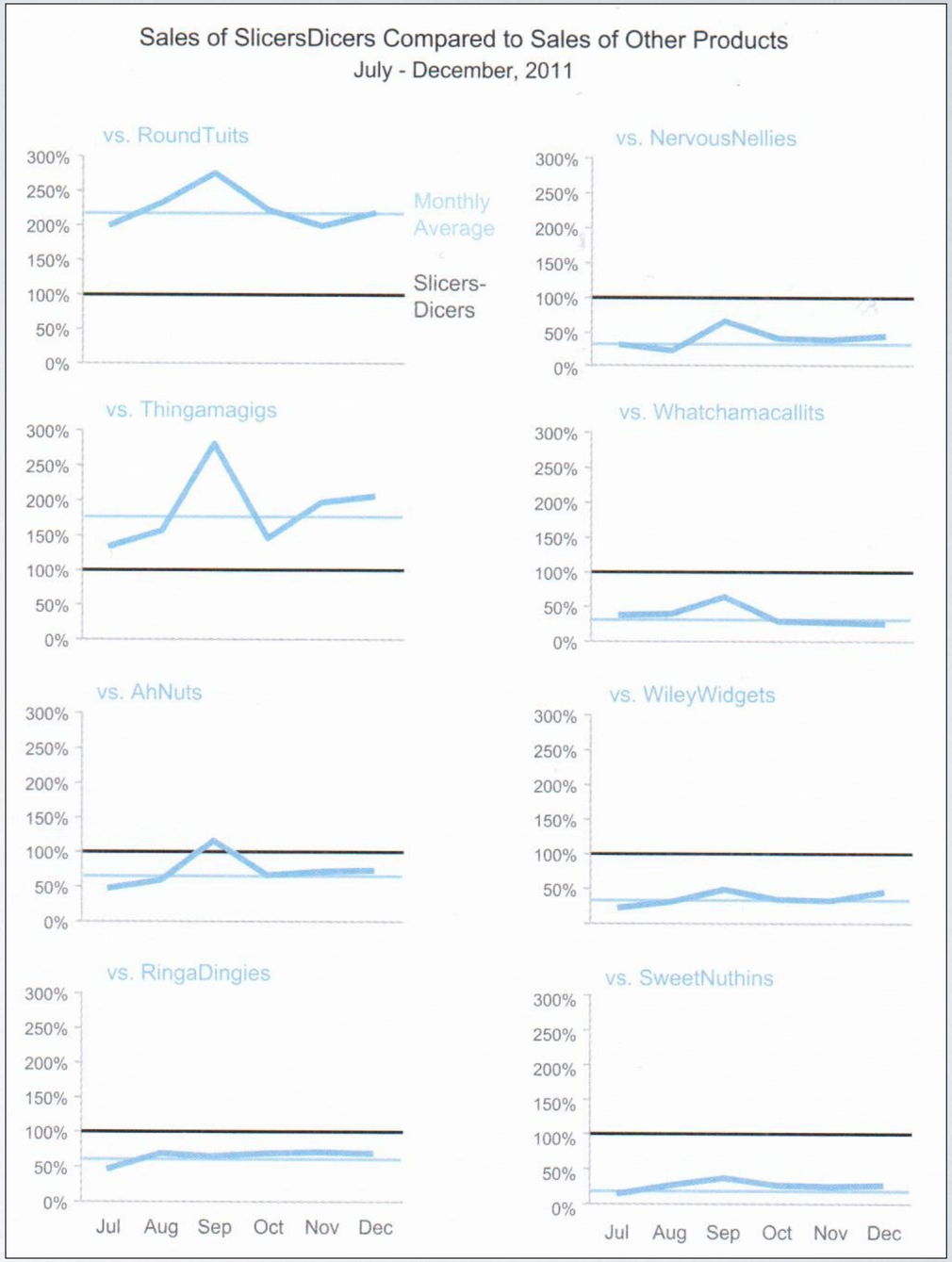
Goal: compare sales of SlicersDicers
with other products



INEFFECTIVE GRAPHS EXAMPLES

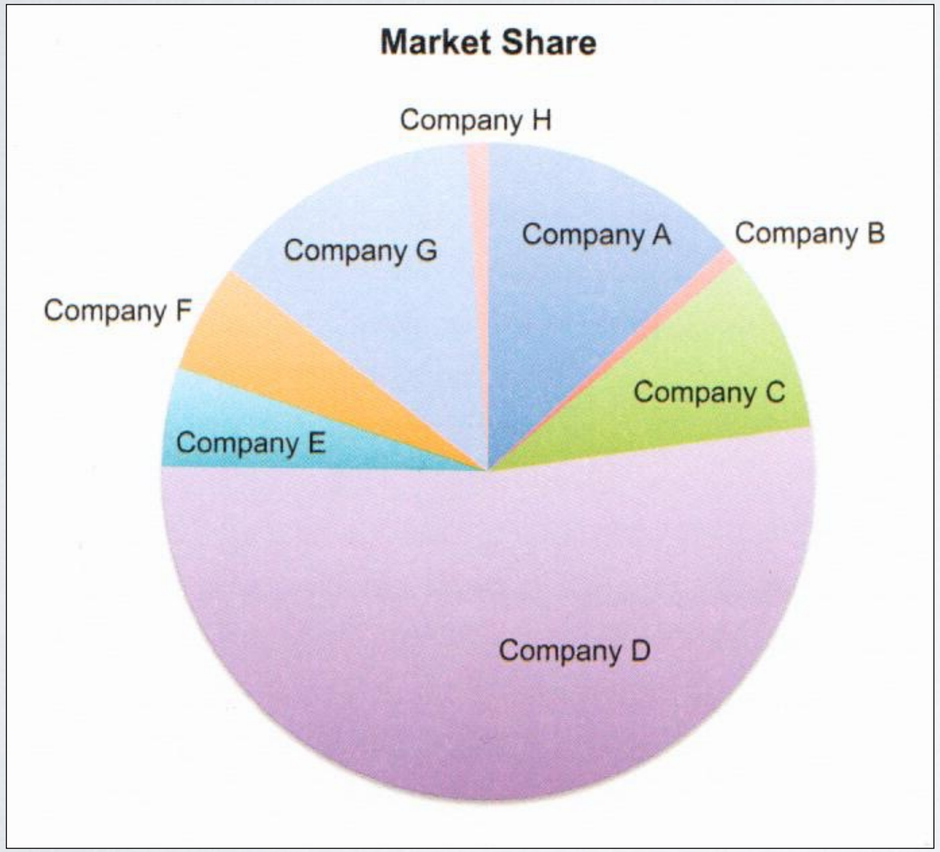
Goal: compare sales of SlicersDicers with other products





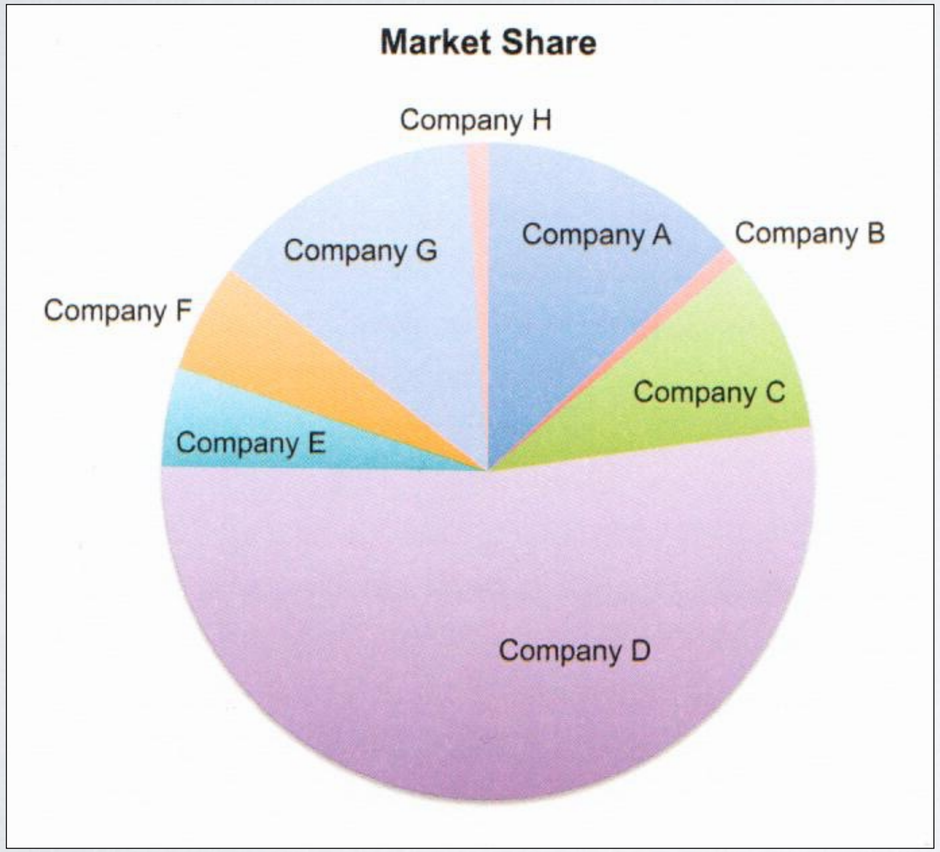
INEFFECTIVE GRAPHS EXAMPLES

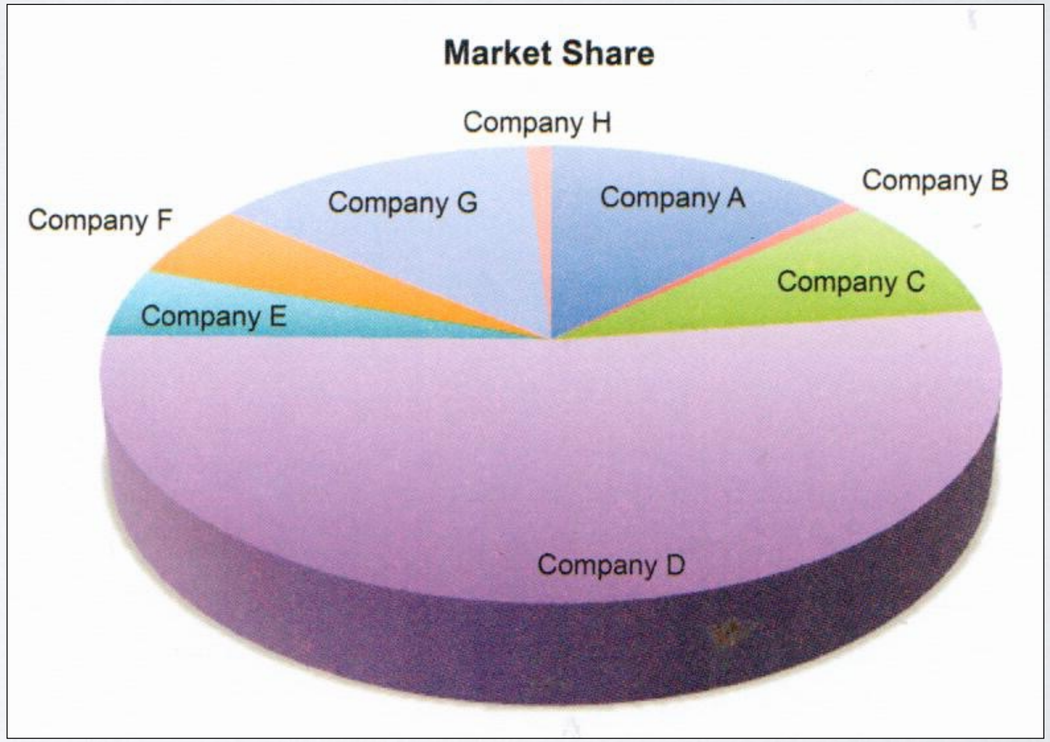
Goal: compare the market share of company G to the shares of its competitors



INEFFECTIVE GRAPHS EXAMPLES

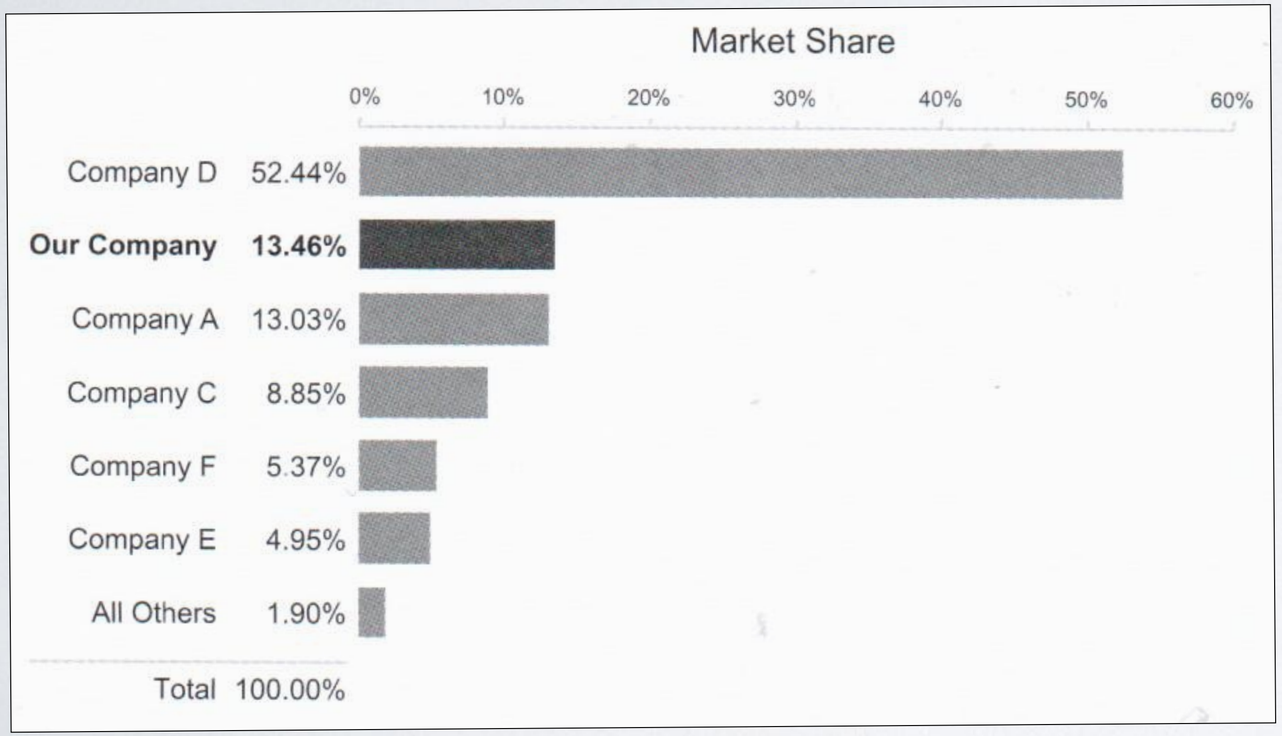
Goal: compare the market share of company G to the shares of its competitors

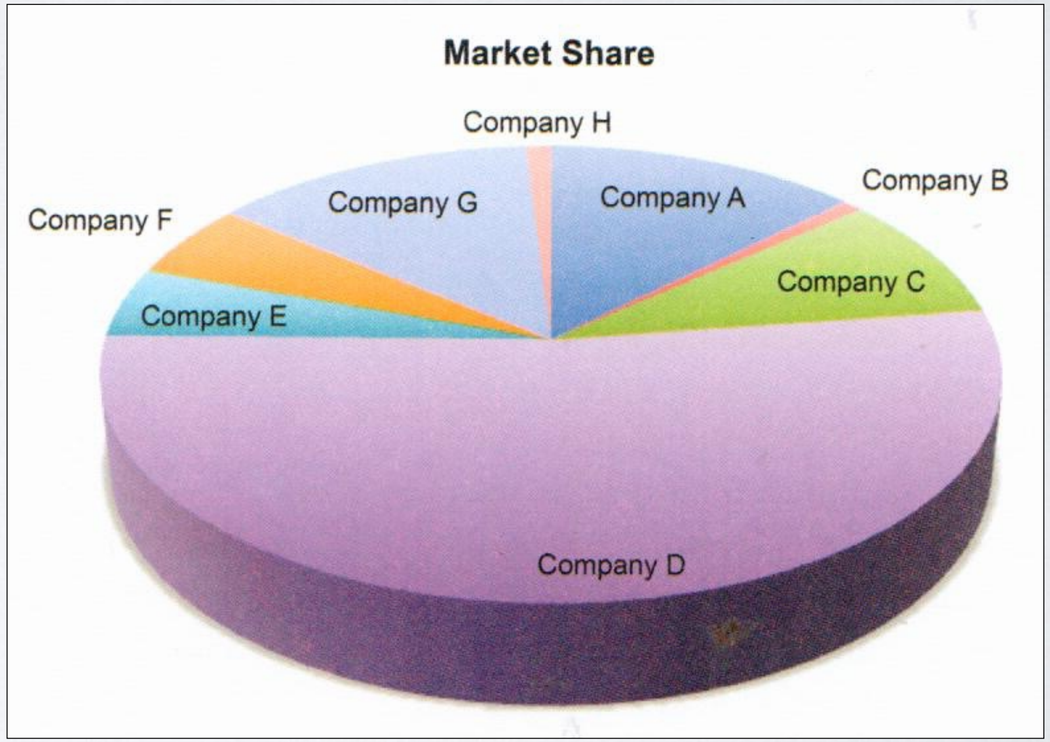




INEFFECTIVE GRAPHS EXAMPLES

Goal: compare the market share of company G to the shares of its competitors



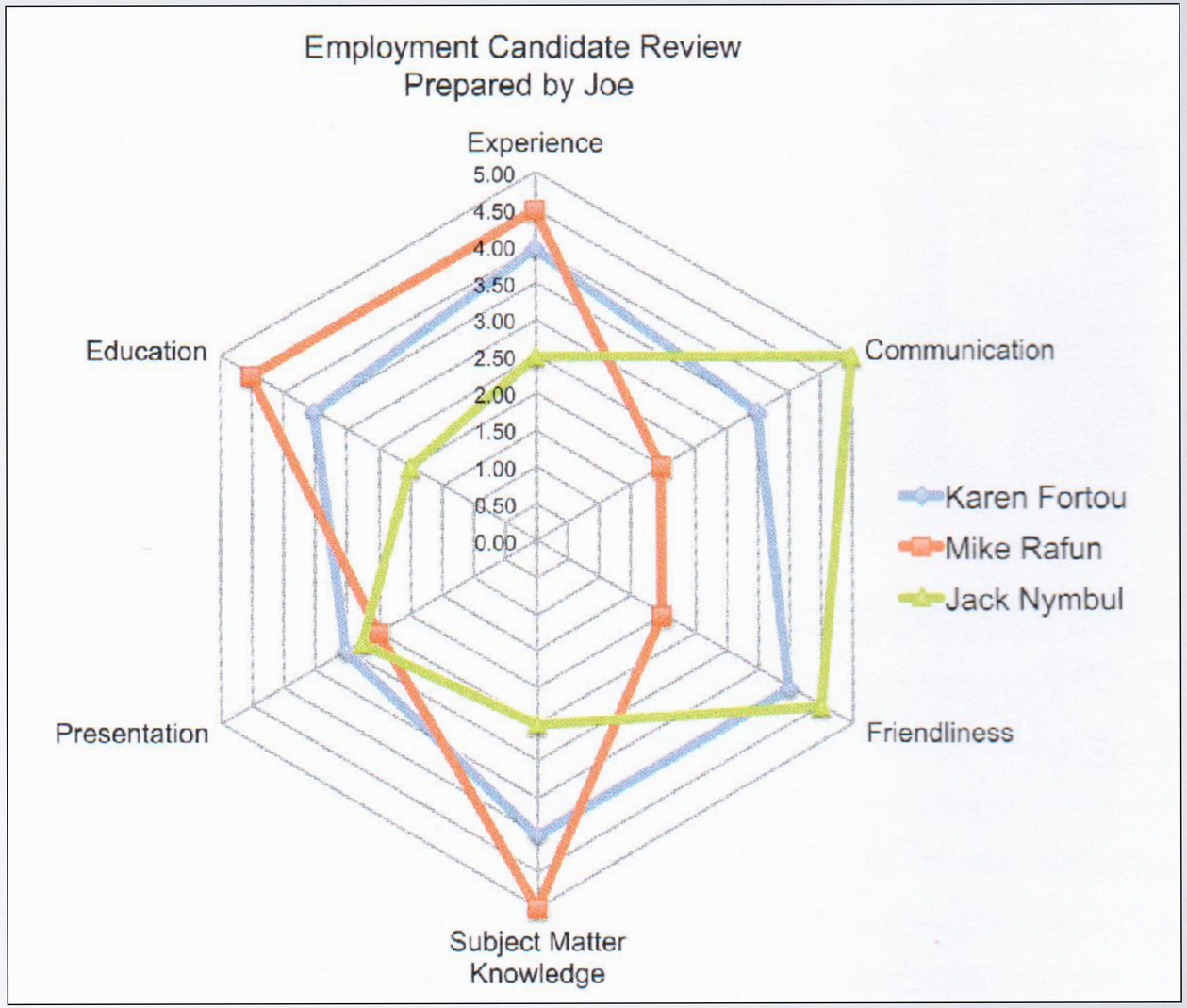


THE BASICS

- Tables and charts are the two primary means to structure and communicate quantitative information
- Quantitative messages are made by two types of data
    - Quantitative
    - Categorical
- Numbers can be categorical

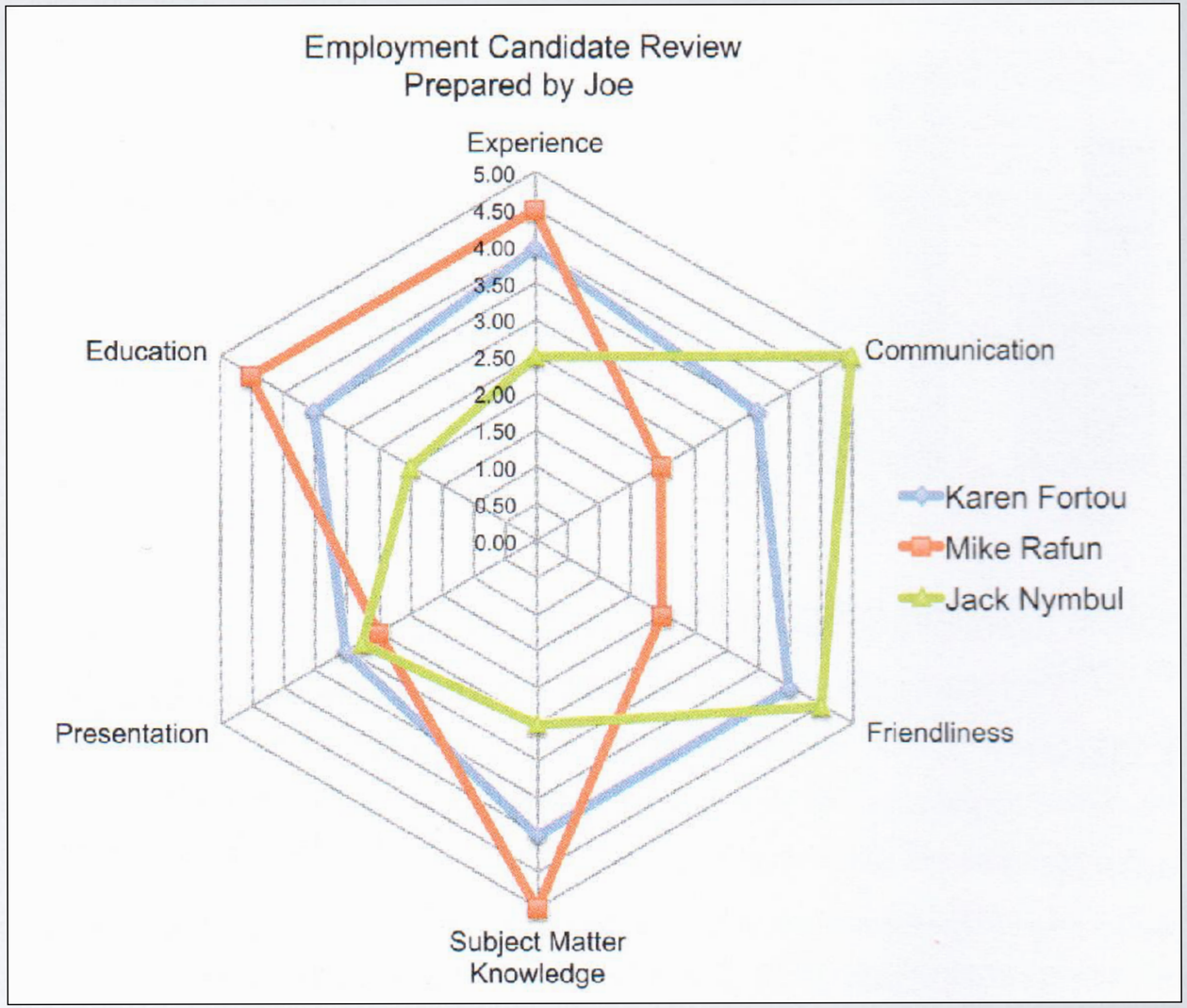
FANCY ≠ EFFECTIVE

- Keep it simple
  - impress ≠ communicate
  - Goal is to communicate
  - Be effective



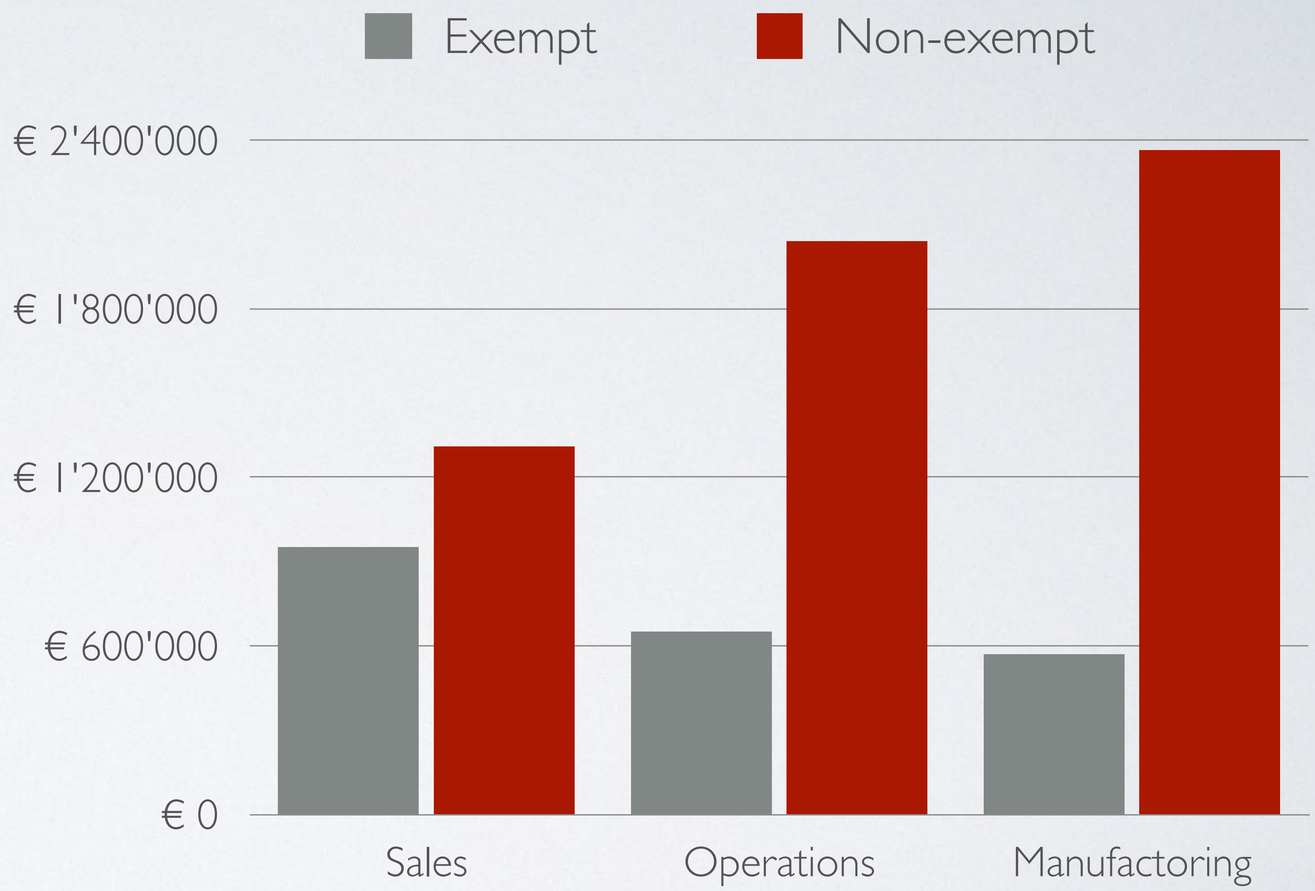
FANCY ≠ EFFECTIVE

Rating areas	K. Fortou	M. Rafun	J. Nymbul
Experience	4.00	4.50	2.50
Communication	3.50	2.00	5.00
Friendliness	4.00	2.00	4.50
Subject matter knowledge	4.00	5.00	2.50
Presentation	3.00	1.50	2.75
Education	3.50	4.50	2.00
Average rating	3.67	3.25	3.21



QUANTITATIVE VS CATEGORICAL DATA

Department	Exempt	Non-exempt
Sales	€ 950'003	€ 1'309'864
Operations	€ 648'763	€ 2'039'927
Manufactoring	€ 568'543	€ 2'367'303
Total	€ 2'167'309	€ 5'717'094



TABLES

- Definition: it renders more than one set of values, where a relationship exists between separated sets of values: related values are aligned by putting them in the same row or column (example: calendar)

- When to use tables?

    - **To look up values**

    - As a means of displaying simple relationships between quantitative values and the categorical subdivisions to which these values are related

    - To compare pairs of related values: values are encoded as text so no translation is required

    - When a high level of precision is necessary (it cannot be provided by graphs)

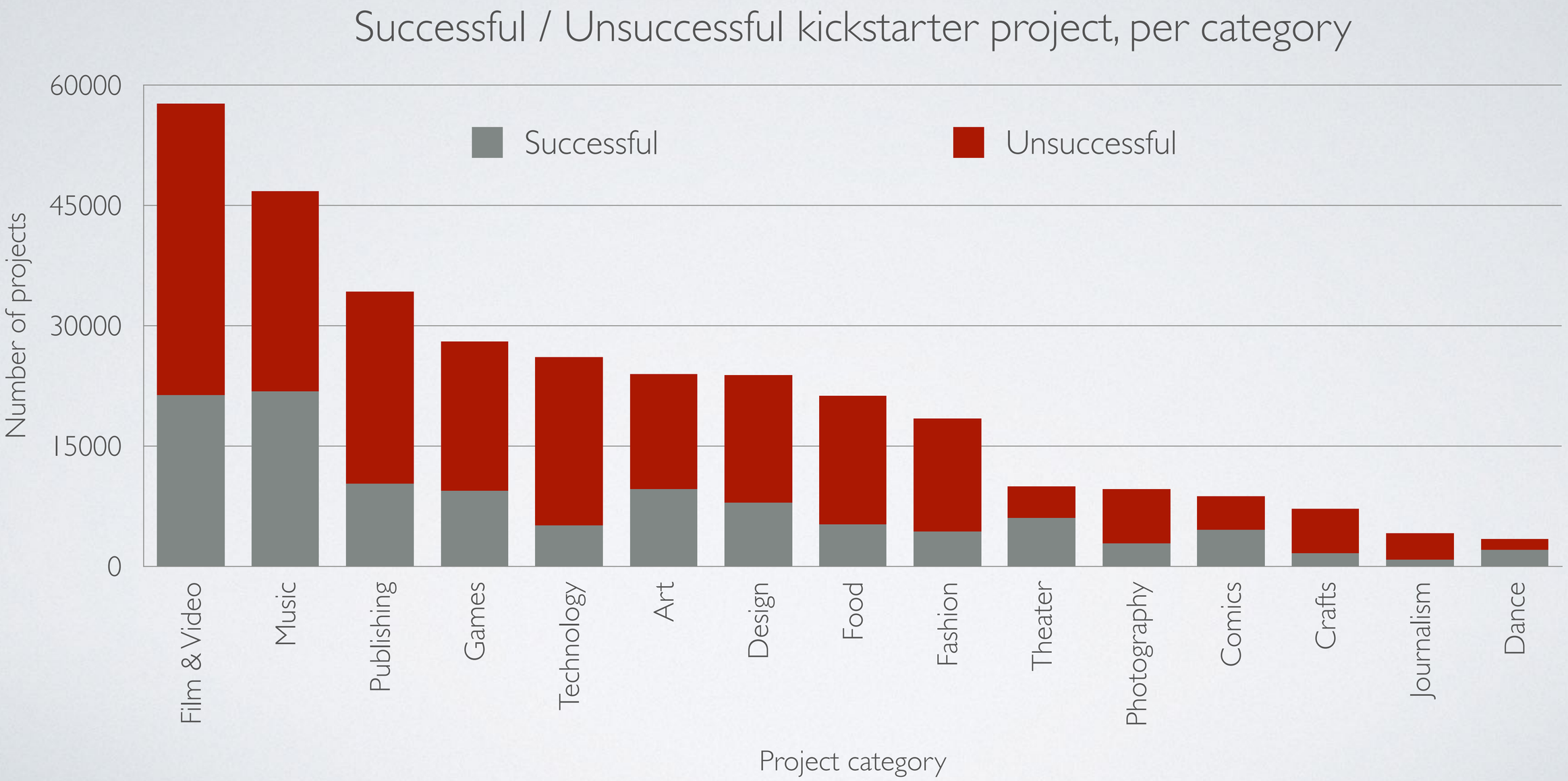
    - When multiple sets are expressed in different units of measure

    - When both summary and detail values are included

GRAPHS

- Definition: a visual display of quantitative information
- Characteristics:
  - It encodes quantitative values visually
  - Values are displayed within an area delimitated by one or more axes
  - Values are encoded as visual objects positioned in relation to the axes
  - Axes provide scale (quantitative and categorical) that are used to assign values and labels to the visual objects

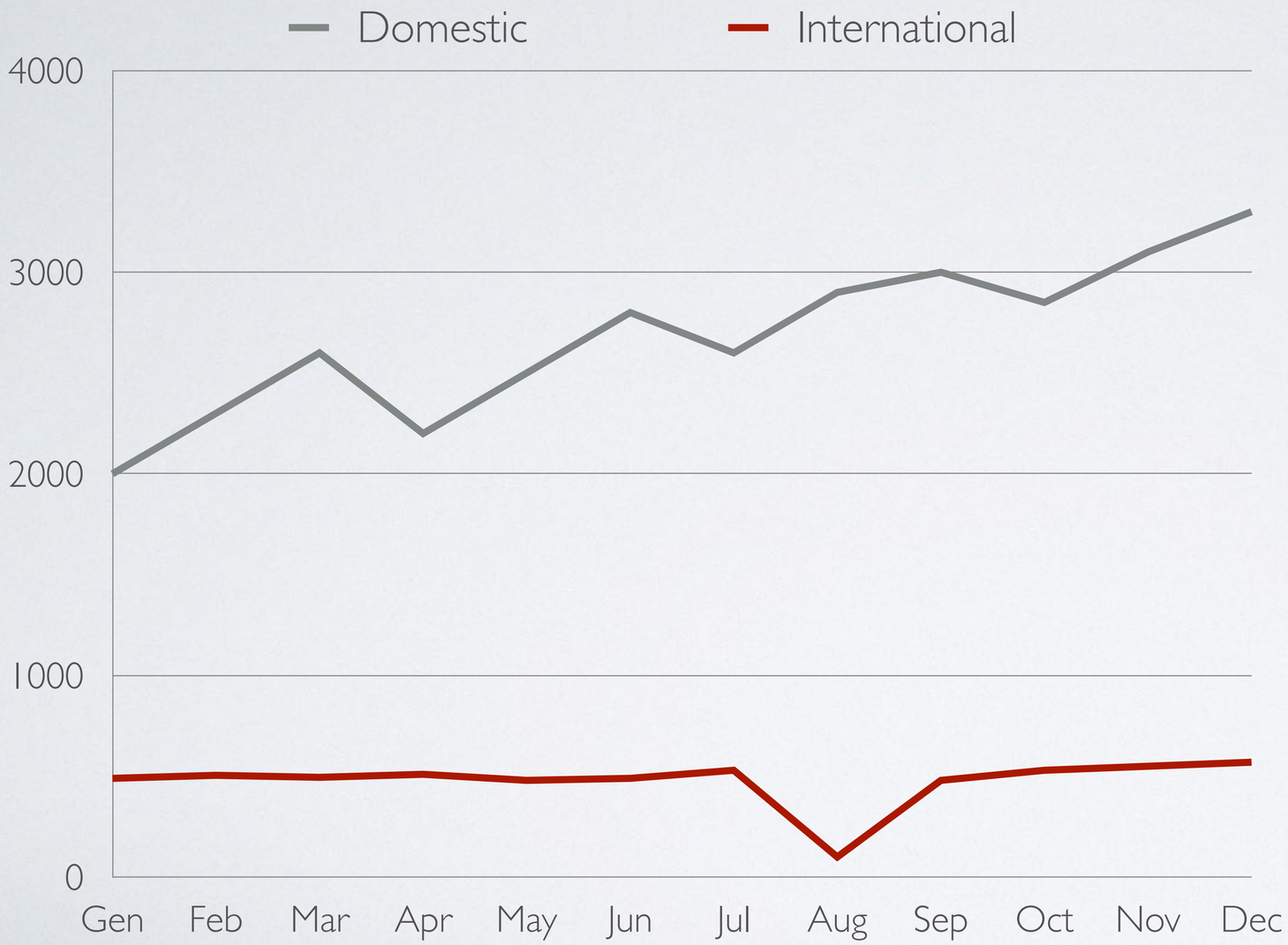
GRAPH EXAMPLE



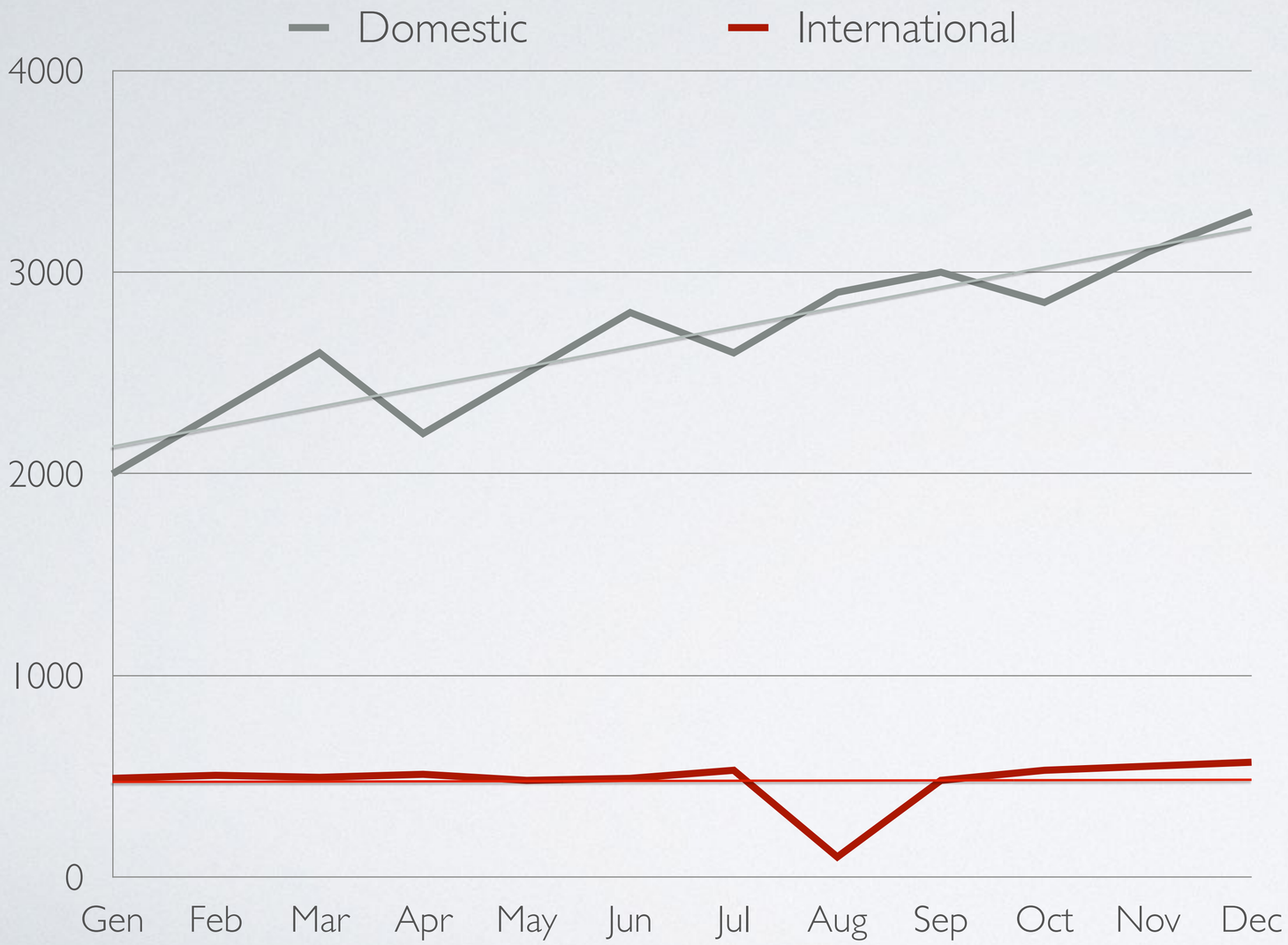
BENEFITS OF GRAPHS

- Graphs reveal much more than a collection of individual values
    - The patterns revealed enable the readers to detect many points of interest in a single collection
- Graphs present the overall shape of the data. They are used to display relationships among multiple quantitative values by giving them shape
    - They reveal patterns like changes, differences, similarities, and exceptions
    - They can communicate quantitative relationships that are much more complex than simple associations between individual quantitative values and categorical subdivisions that are expressed in tables
    - You could compare millions of measures across several years and you would be able to see the nature of the relationship

WHAT CAN BE OBSERVED IN A GRAPH: EXAMPLE

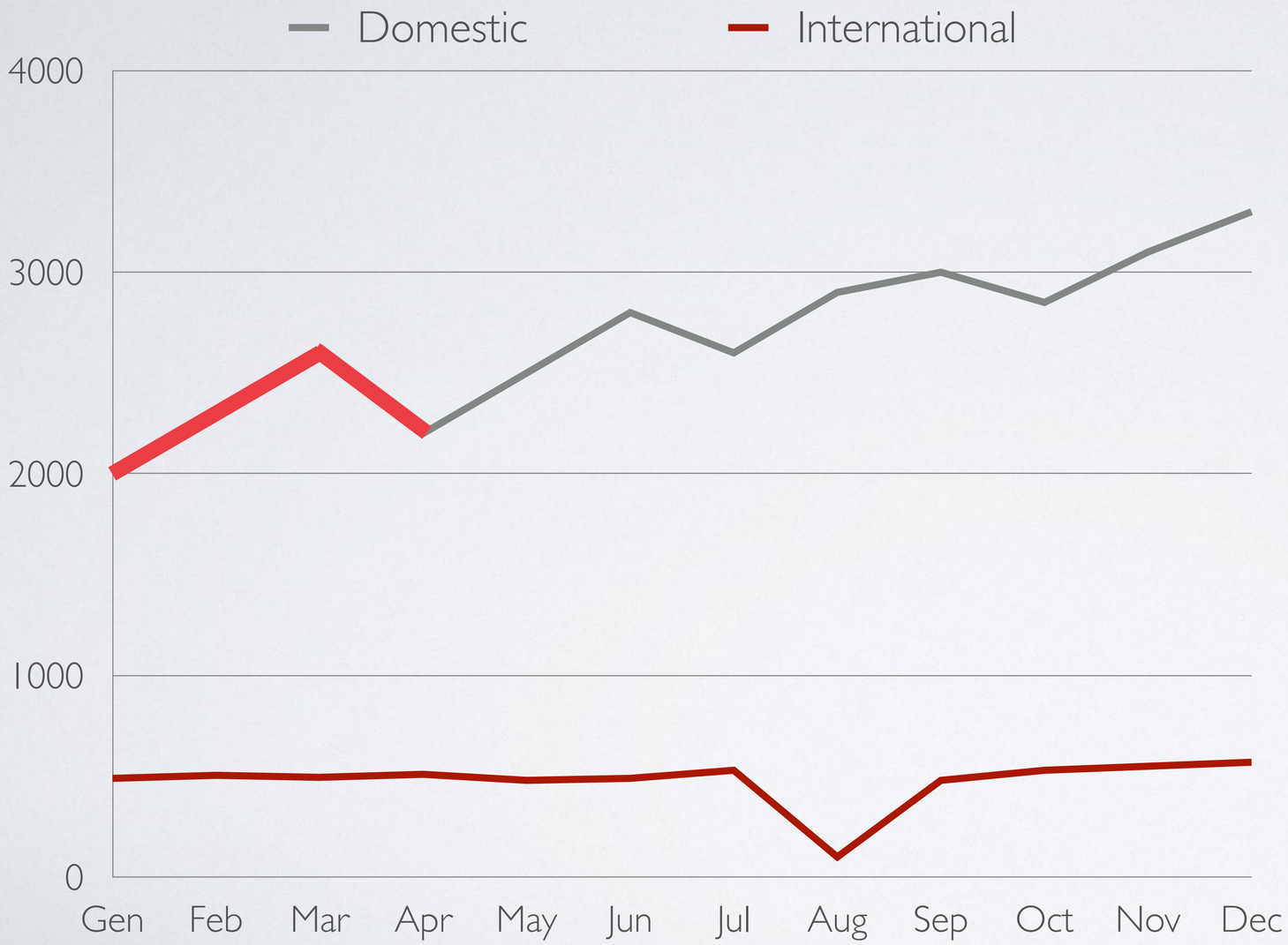


WHAT CAN BE OBSERVED IN A GRAPH: EXAMPLE



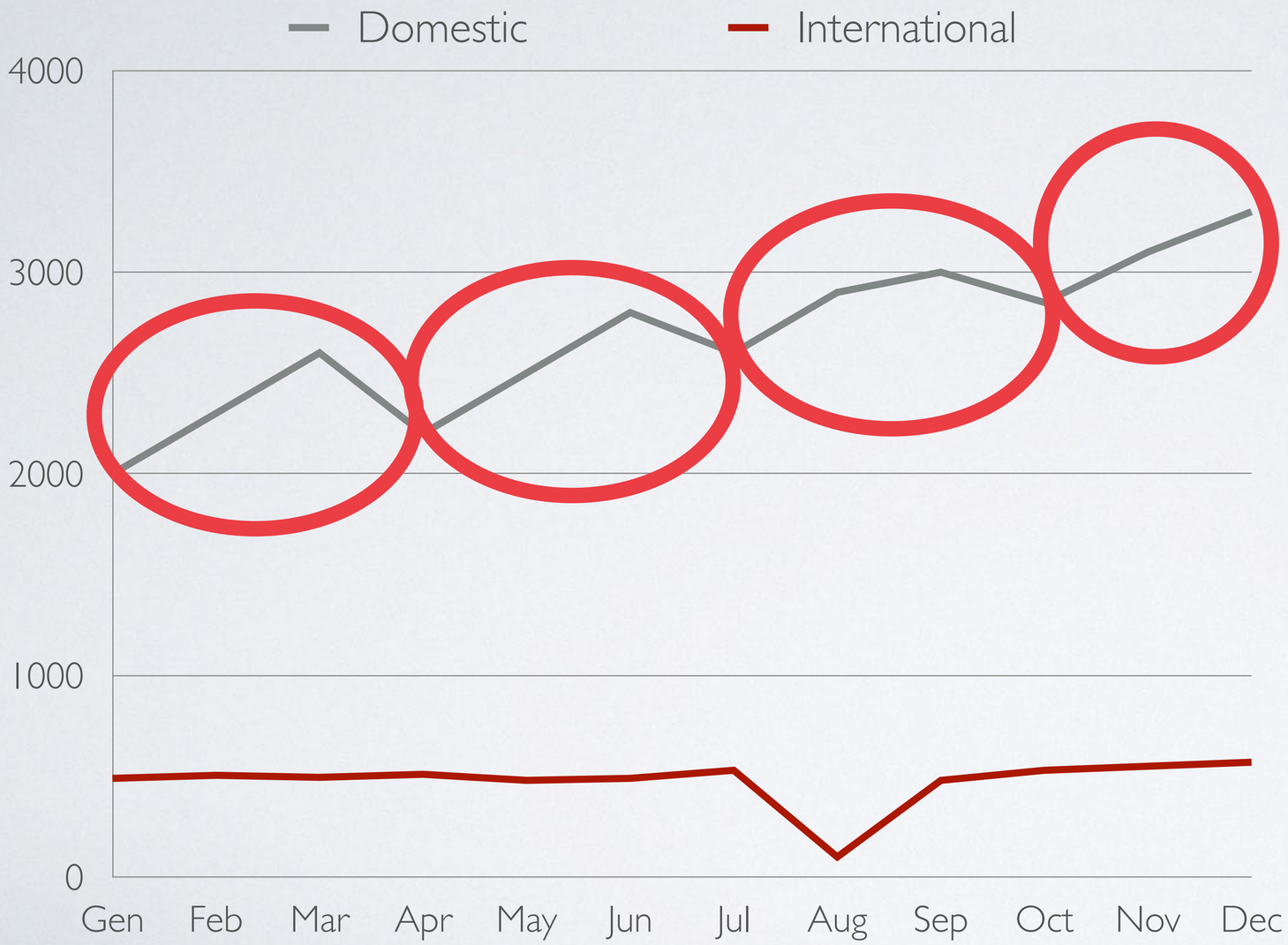
- Pattern goes generally up: it's called trend

WHAT CAN BE OBSERVED IN A GRAPH: EXAMPLE



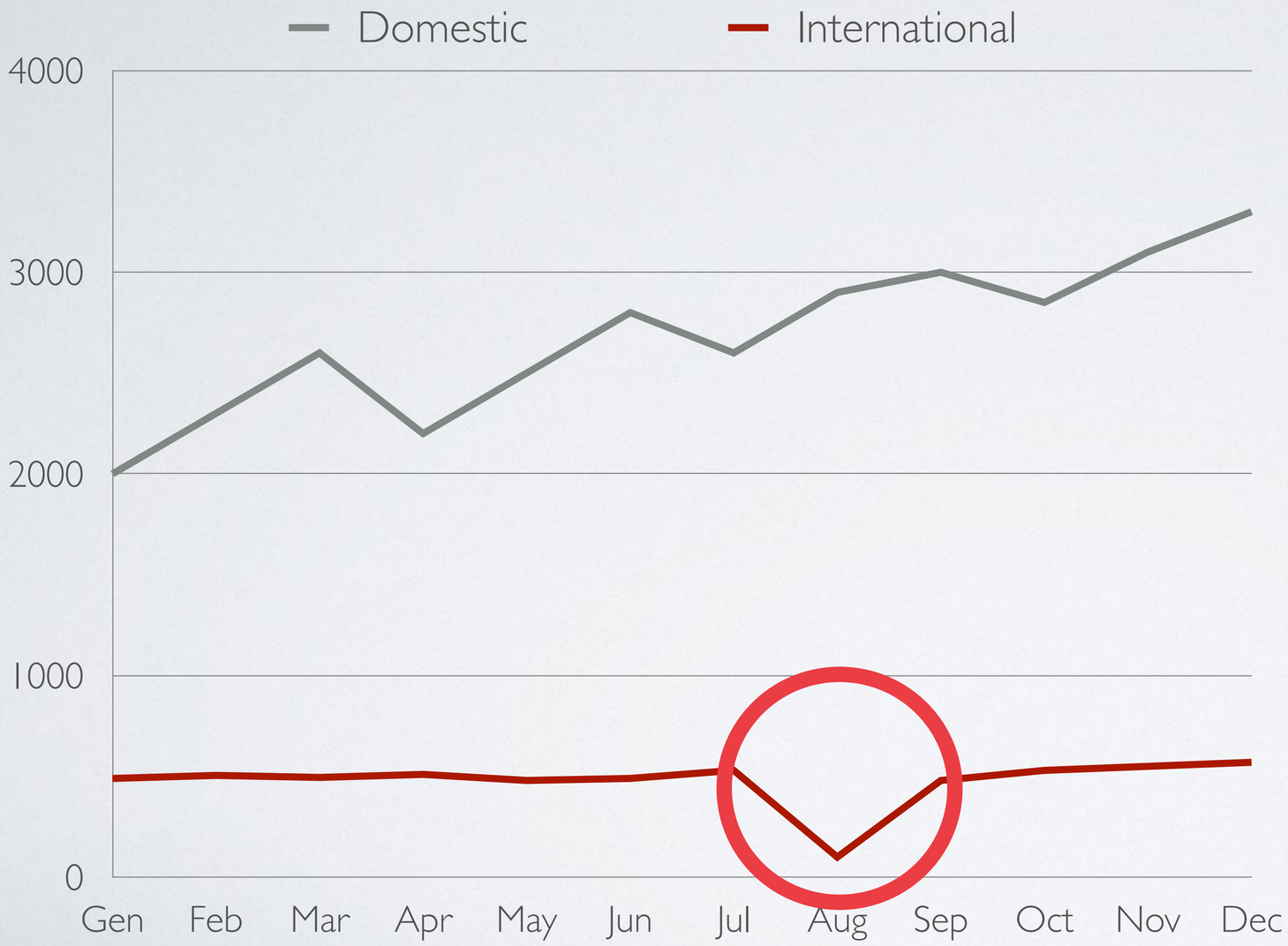
- Pattern goes generally up: it's called trend
- "Hockey stick" behaviour

WHAT CAN BE OBSERVED IN A GRAPH: EXAMPLE



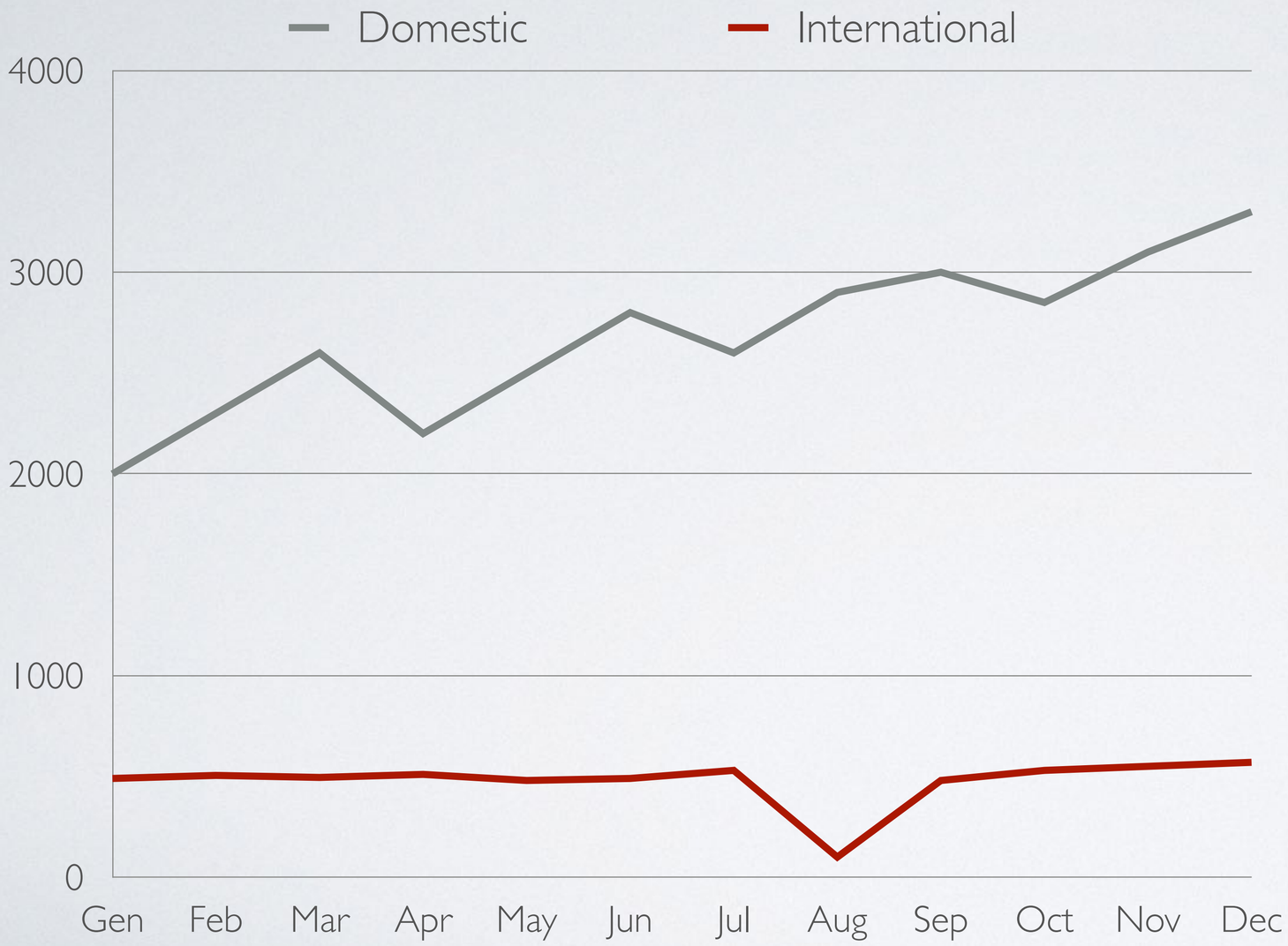
- Pattern goes generally up: it's called trend
- "Hockey stick" behaviour
- Compare between different subset of the same set

WHAT CAN BE OBSERVED IN A GRAPH: EXAMPLE



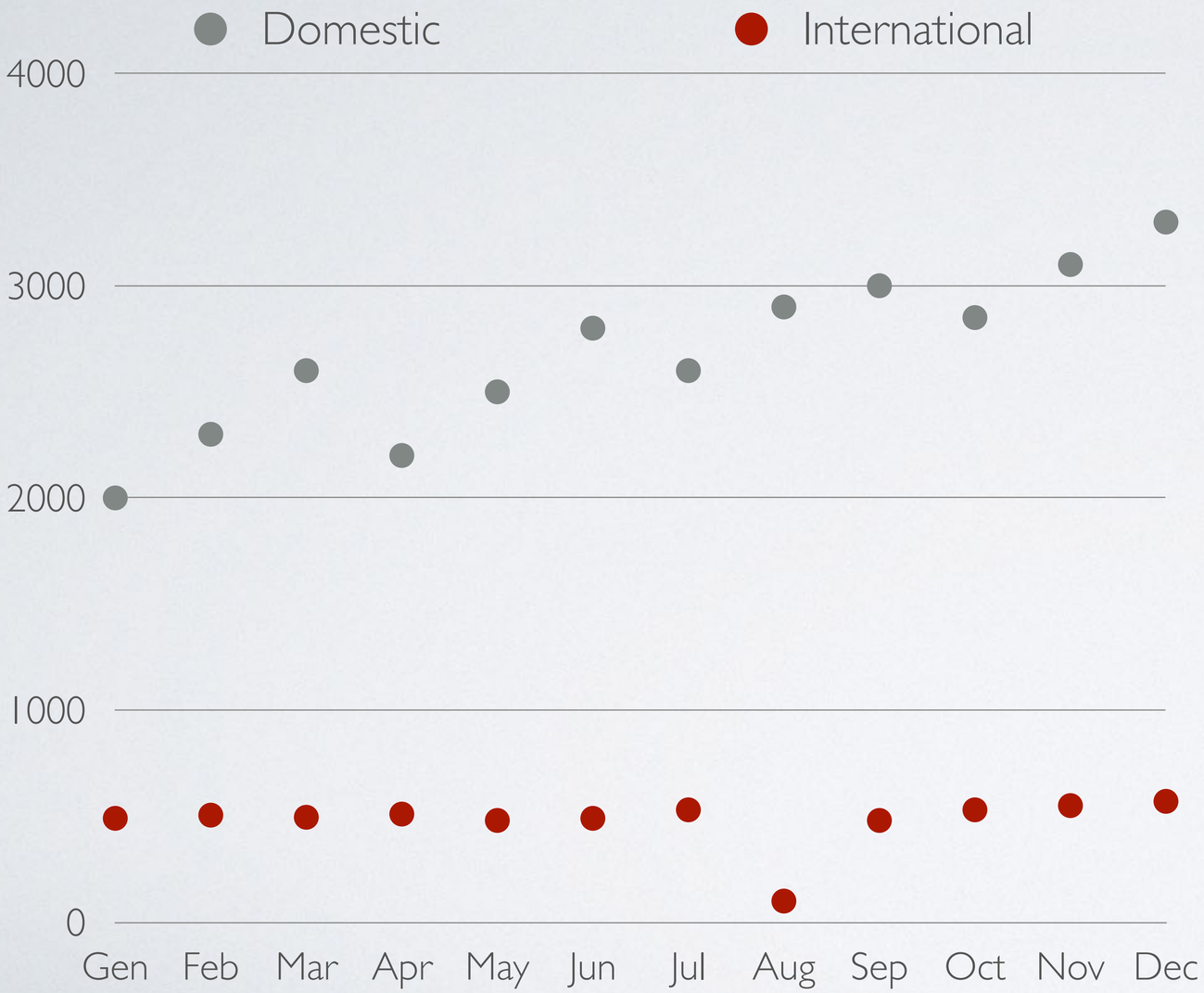
- Pattern goes generally up: it's called trend
- "Hockey stick" behaviour
- Compare between different subset of the same set
- International is steady not affected by season. Only august drop
- Exceptions to general pattern stand out clearly from the rest - would be much harder in a table

WHAT CAN BE OBSERVED IN A GRAPH: EXAMPLE

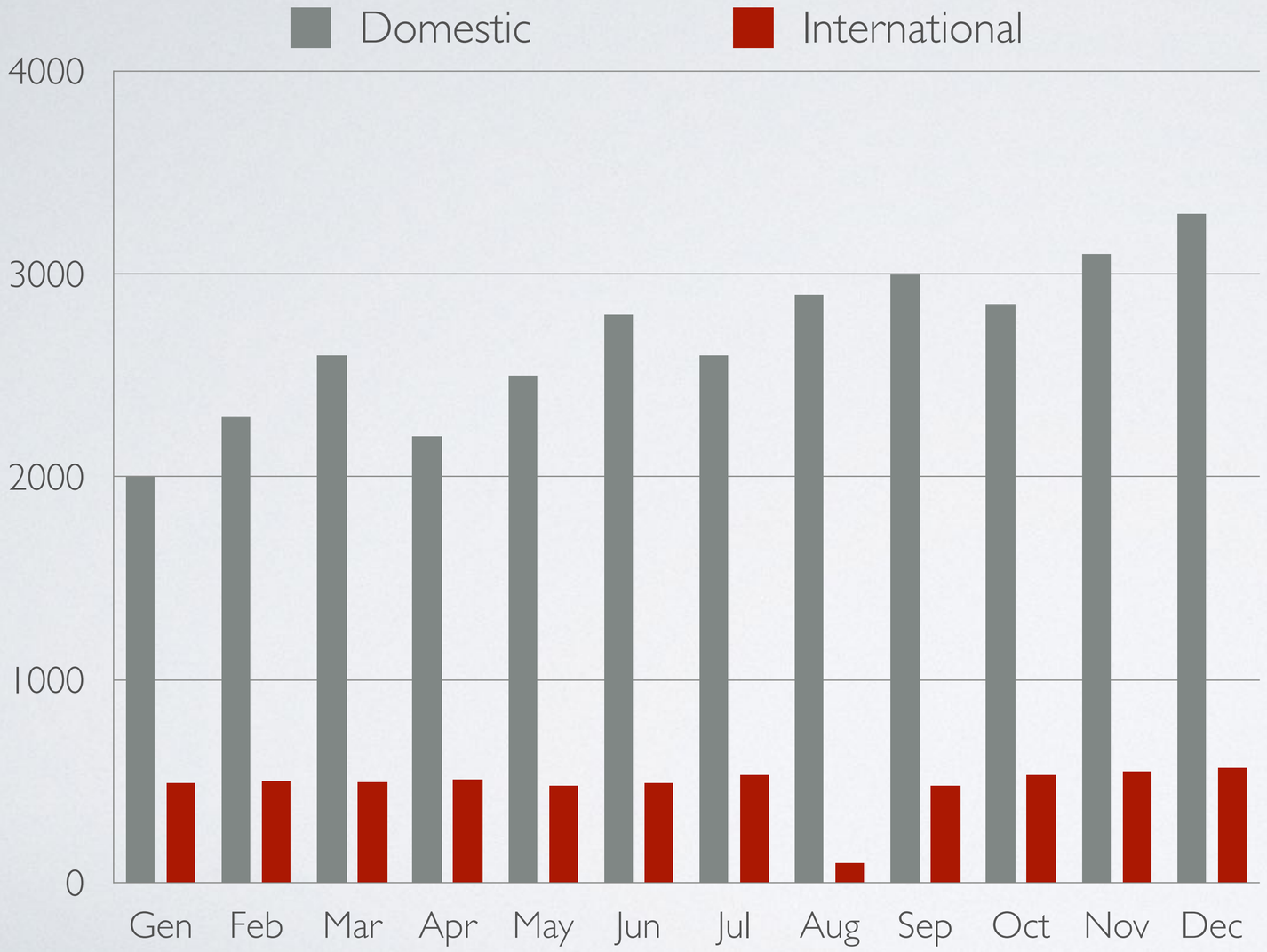


- Pattern goes generally up: it's called trend
- "Hockey stick" behaviour
- Compare between different subset of the same set
- International is steady not affected by season. Only august drop
- Exceptions to general pattern stand out clearly from the rest - would be much harder in a table
- Similarities and differences between two set of values

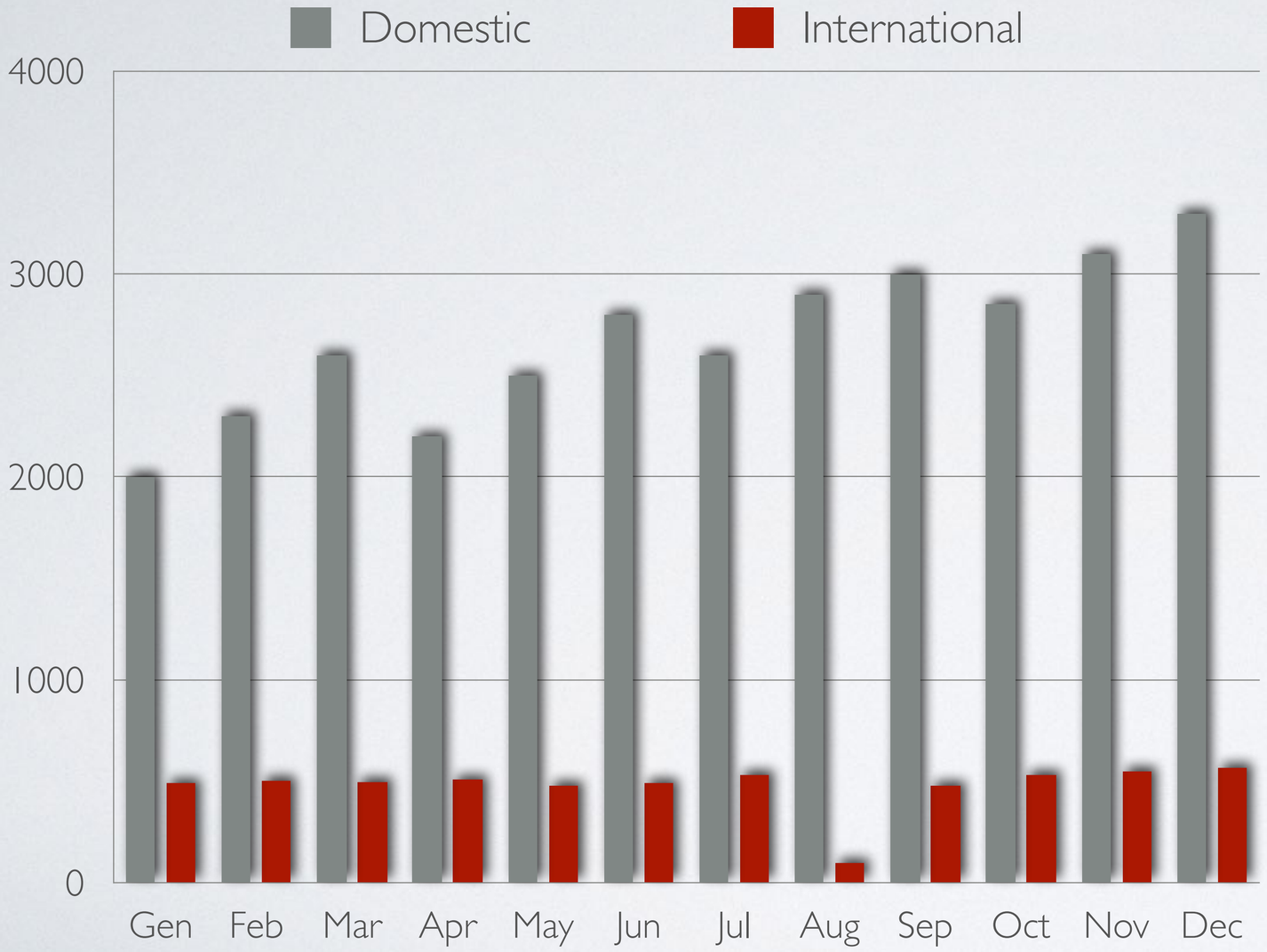
WHAT CAN BE OBSERVED IN A GRAPH: EXAMPLE



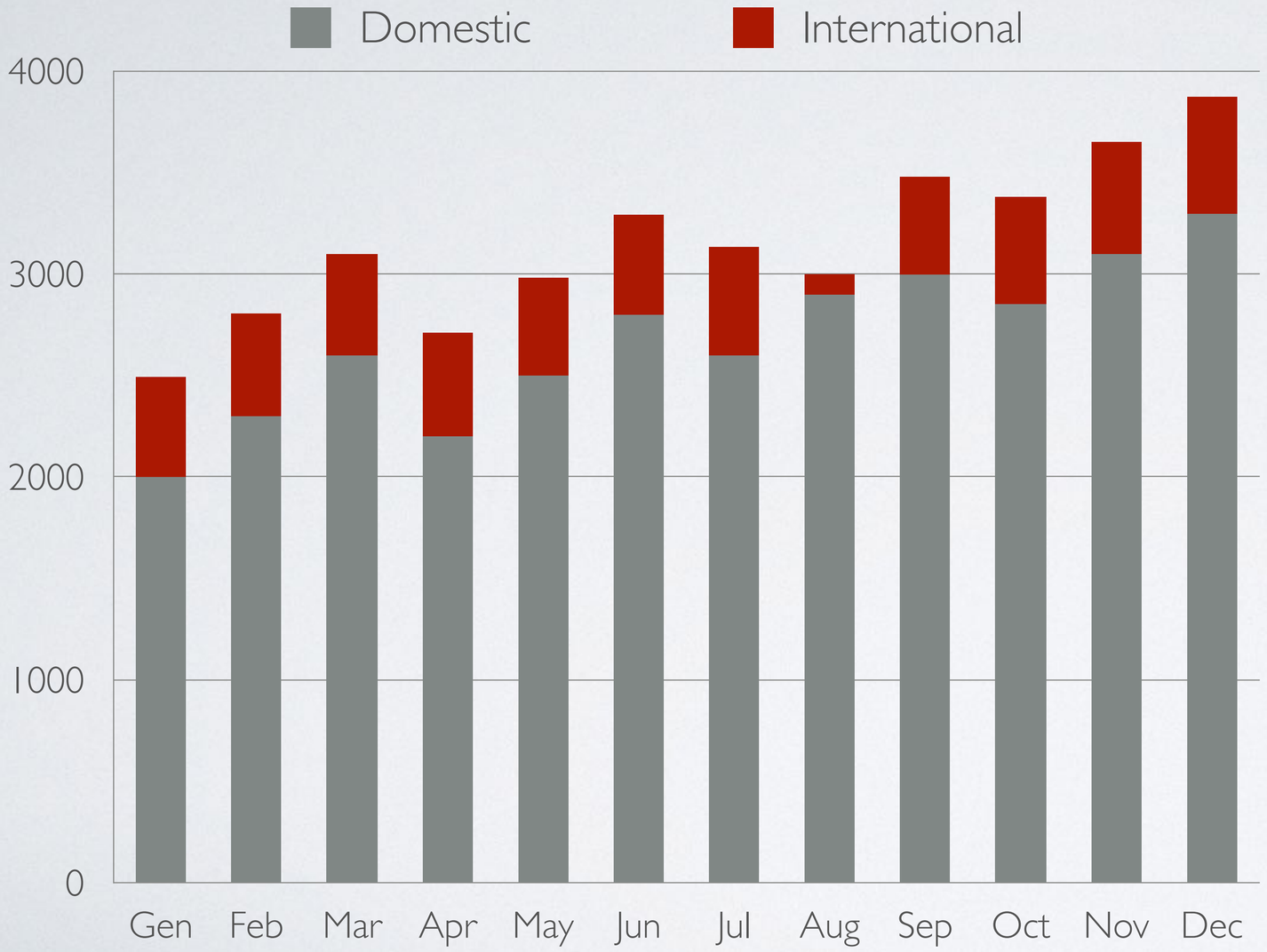
WHAT CAN BE OBSERVED IN A GRAPH: EXAMPLE



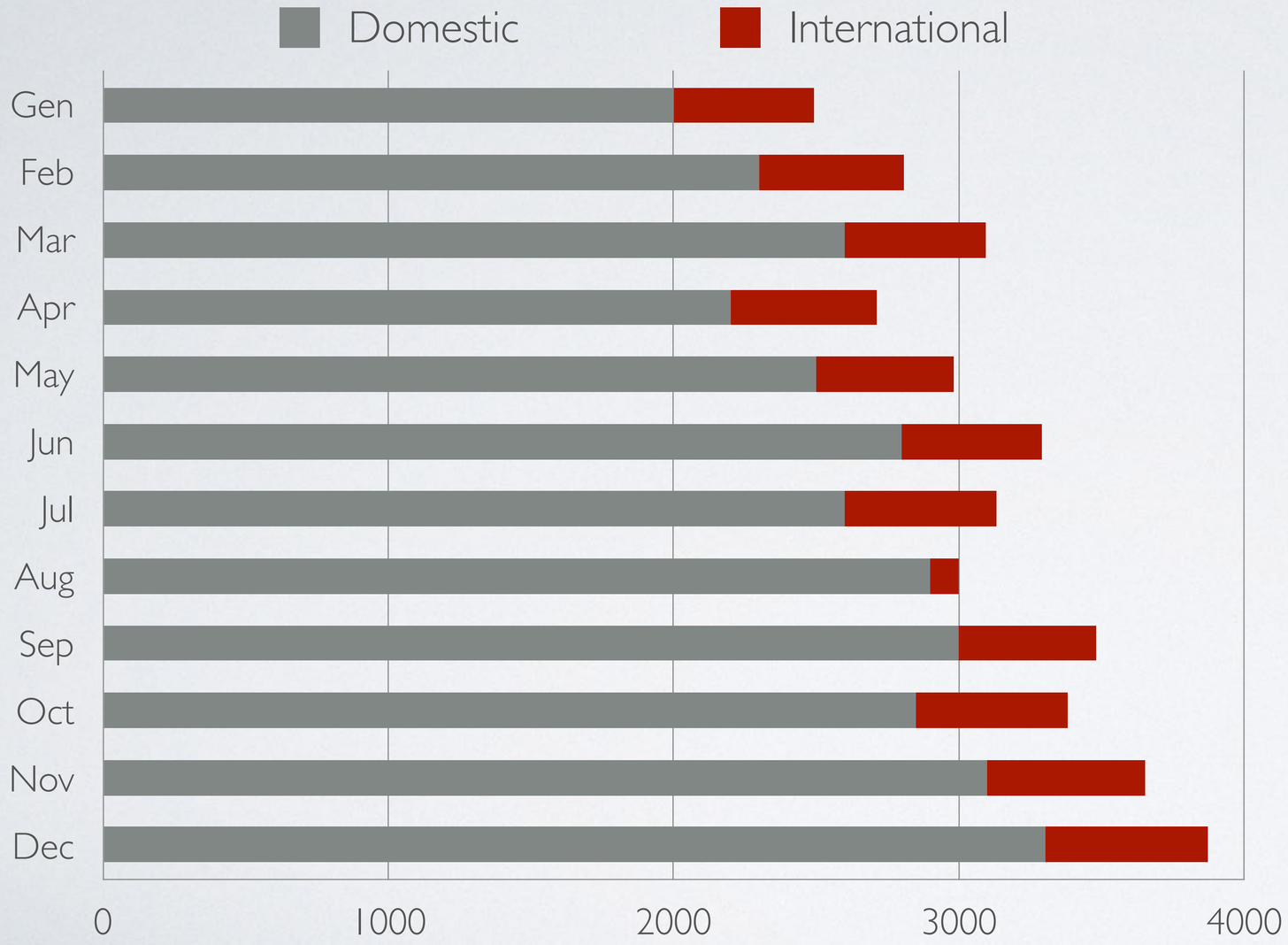
WHAT CAN BE OBSERVED IN A GRAPH: EXAMPLE



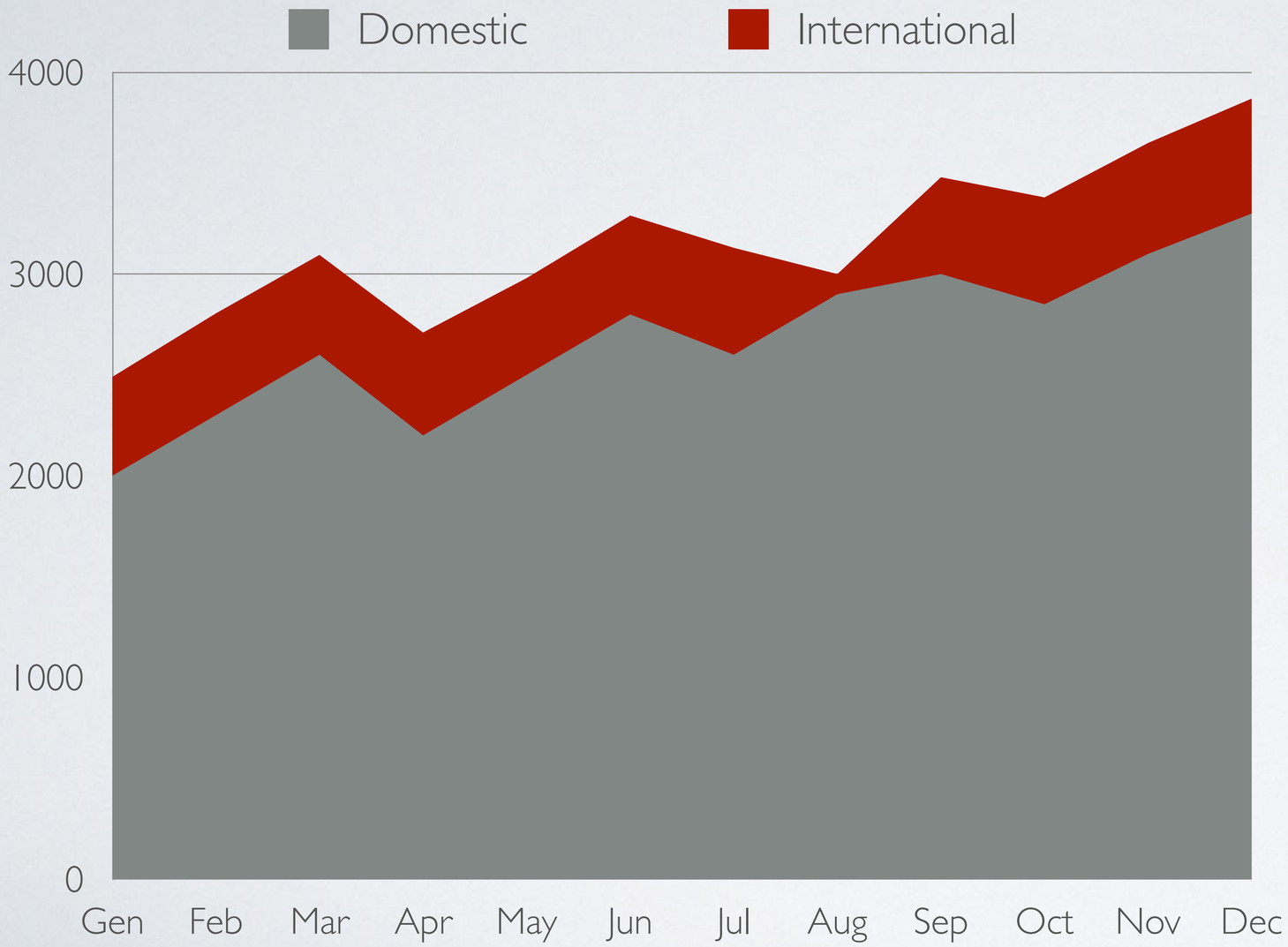
WHAT CAN BE OBSERVED IN A GRAPH: EXAMPLE



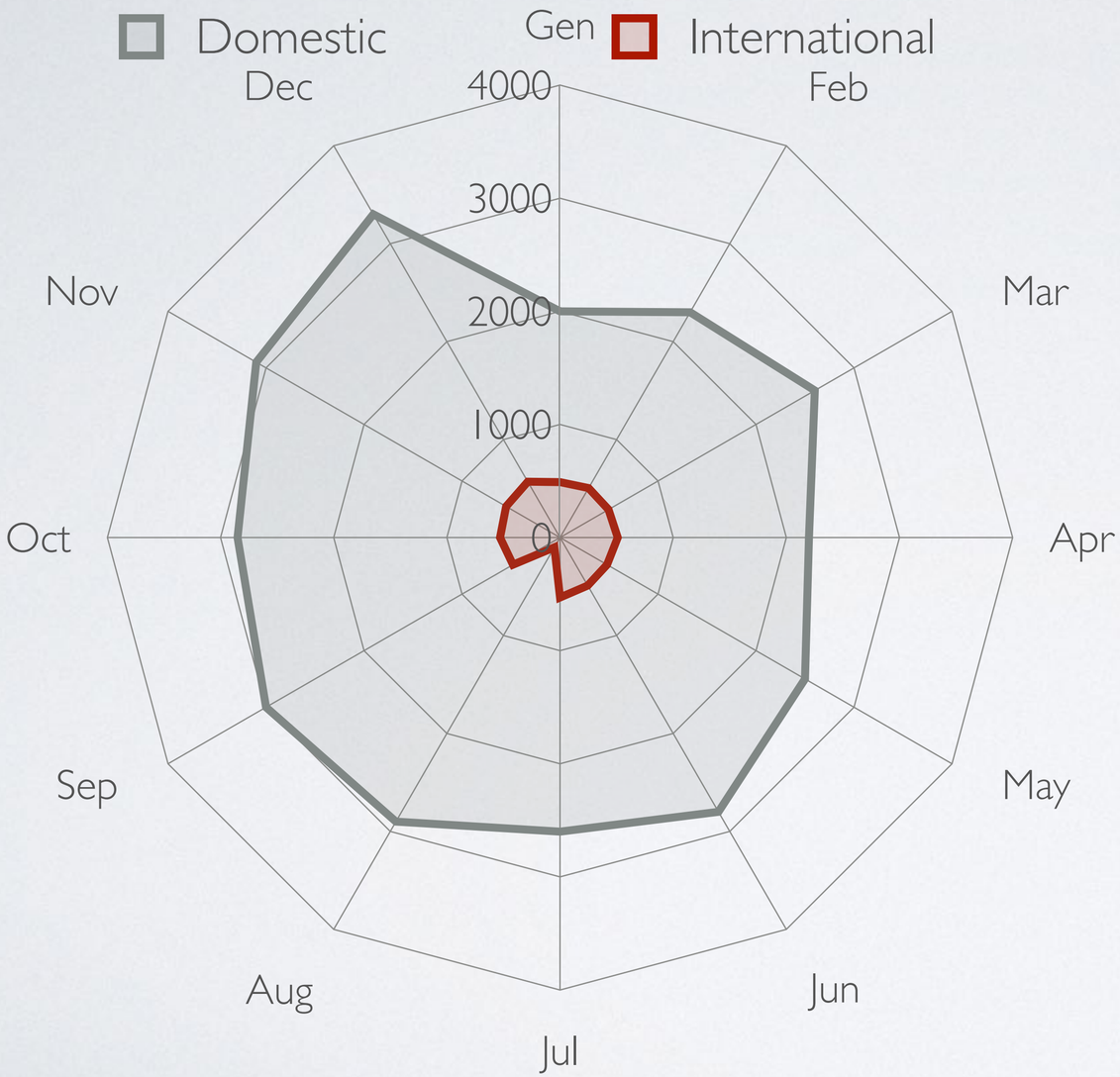
WHAT CAN BE OBSERVED IN A GRAPH: EXAMPLE



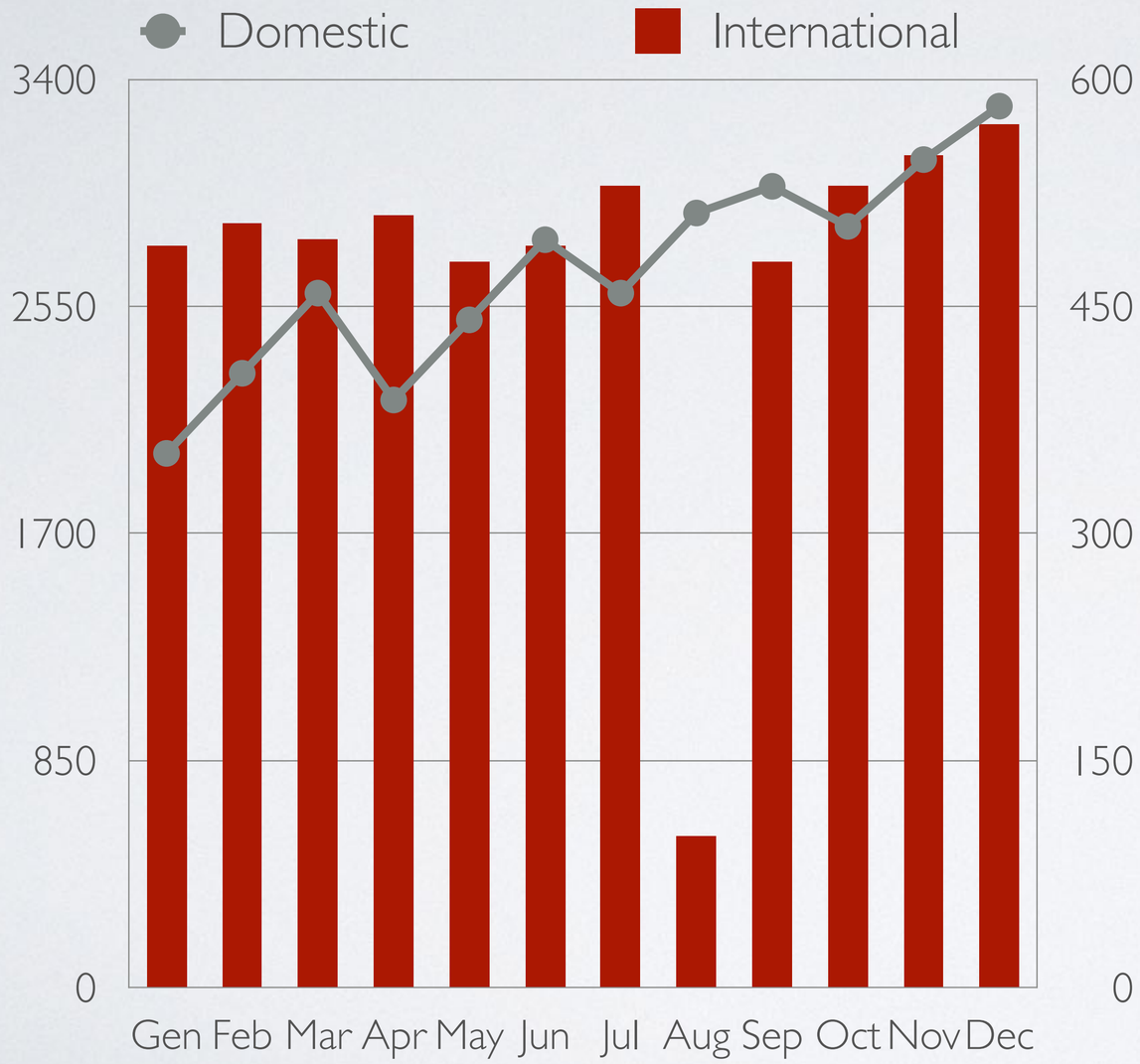
WHAT CAN BE OBSERVED IN A GRAPH: EXAMPLE



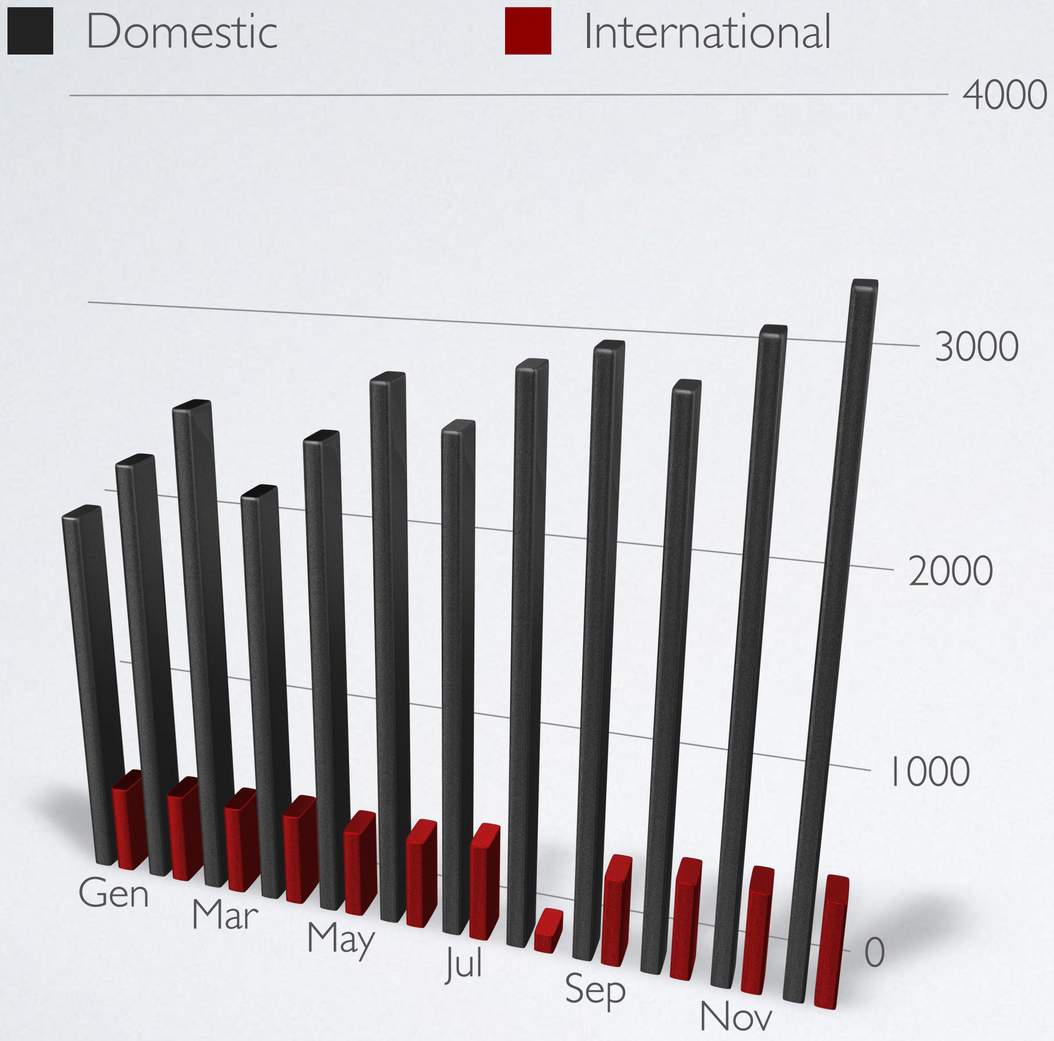
WHAT CAN BE OBSERVED IN A GRAPH: EXAMPLE



WHAT CAN BE OBSERVED IN A GRAPH: EXAMPLE



WHAT CAN BE OBSERVED IN A GRAPH: EXAMPLE



TABLES VS GRAPHS

"Tables usually outperform graphics in reporting of small sets of 20 numbers or less. The special power of graphics comes in the display of large data sets."

–Edward Tufte

- It's true that one's chance of seeing a meaningful pattern in small table of numbers is greater than in one that is large
- However, patters are easier to see in graph

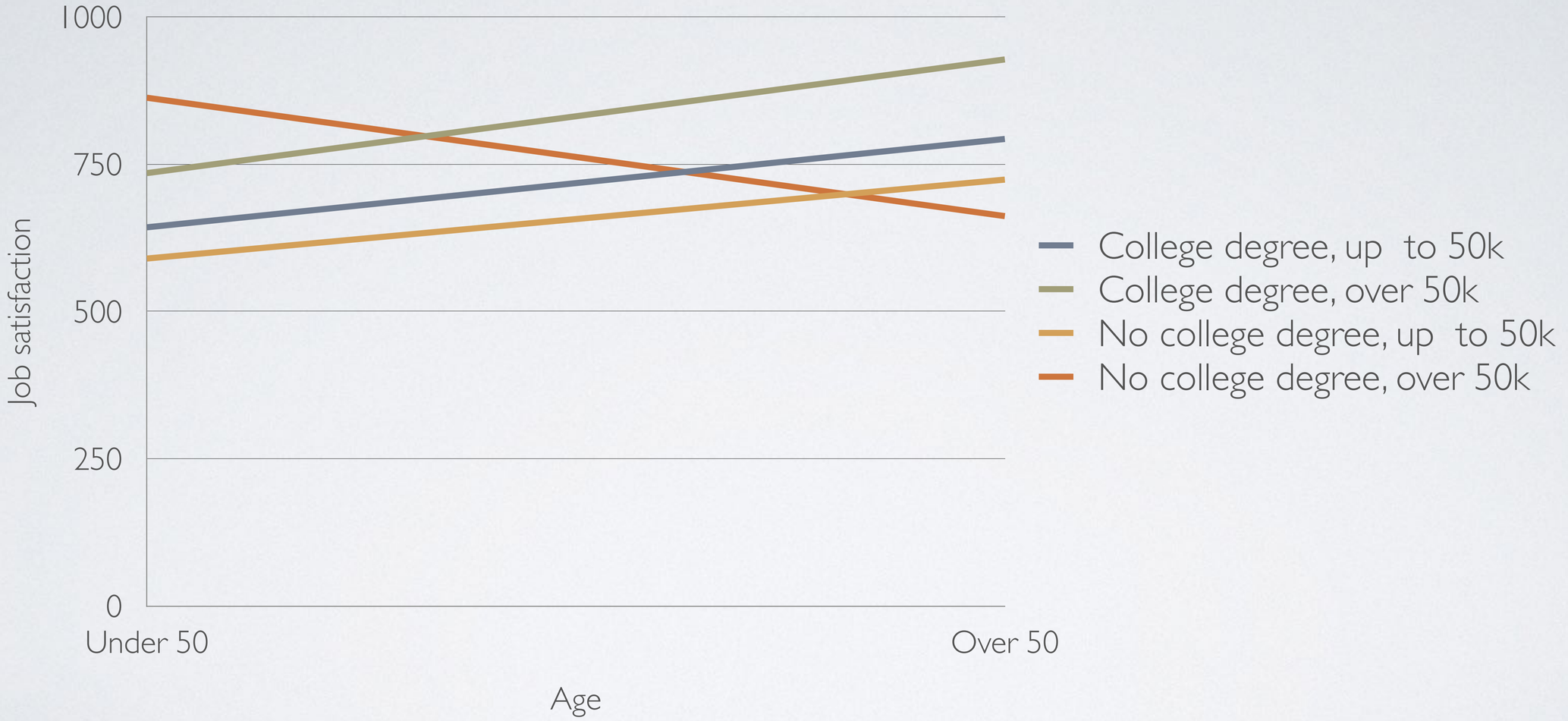
TABLES VS GRAPHS: EXAMPLE

	College degree		No college degree	
Income	Under 50	Over 50	Under 50	Over 50
Up to 50'000	643	793	590	724
Over 50,000	735	928	863	662

Job satisfaction by income, education, and age

Do you see any group exhibiting a different pattern of satisfaction from the others?

TABLES VS GRAPHS: EXAMPLE



Do you see any group exhibiting a different pattern of satisfaction from the others?

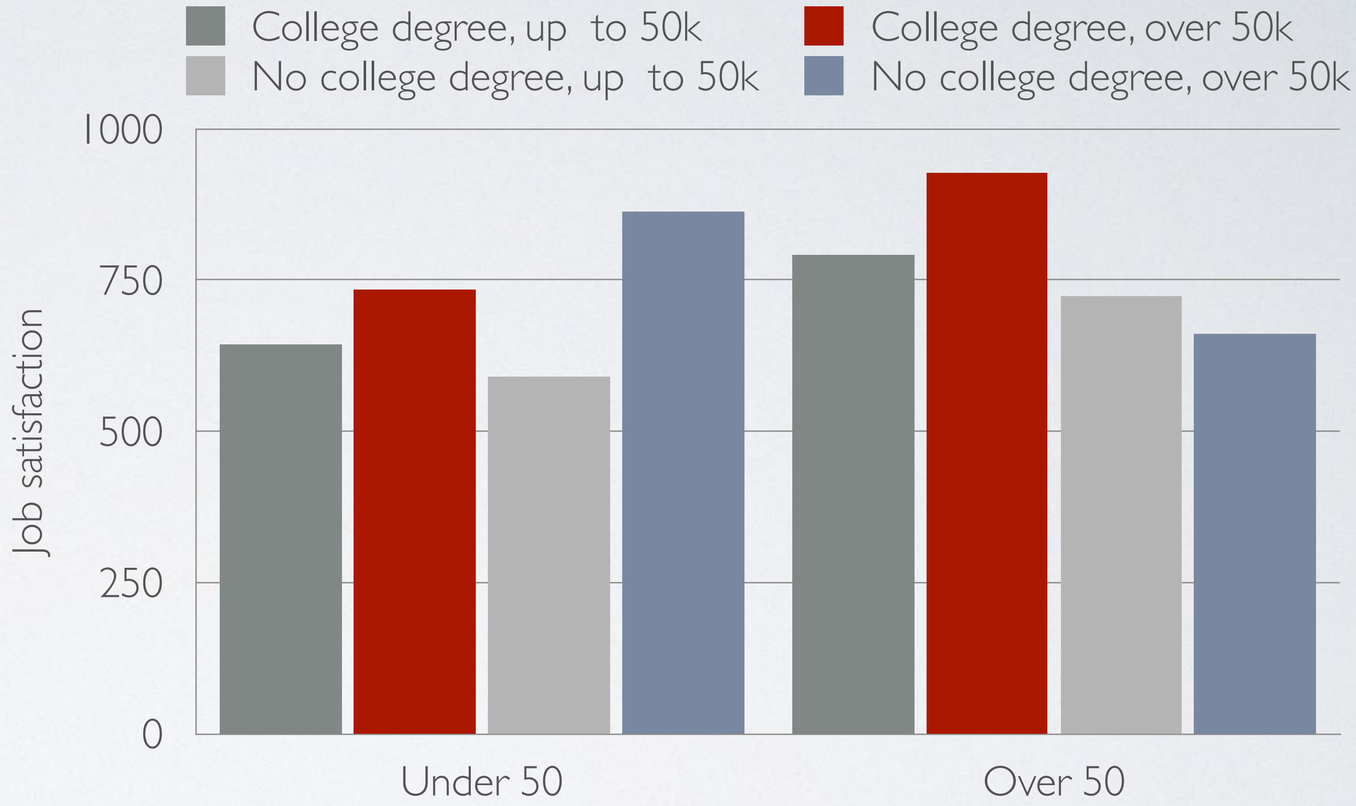
TABLES VS GRAPHS: EXAMPLE

- Even with only 8 numbers, the previous graph did what the table could not: it made the diverging pattern obvious

- This results not only because we used a graph, but because the graph was designed to feature this particular pattern

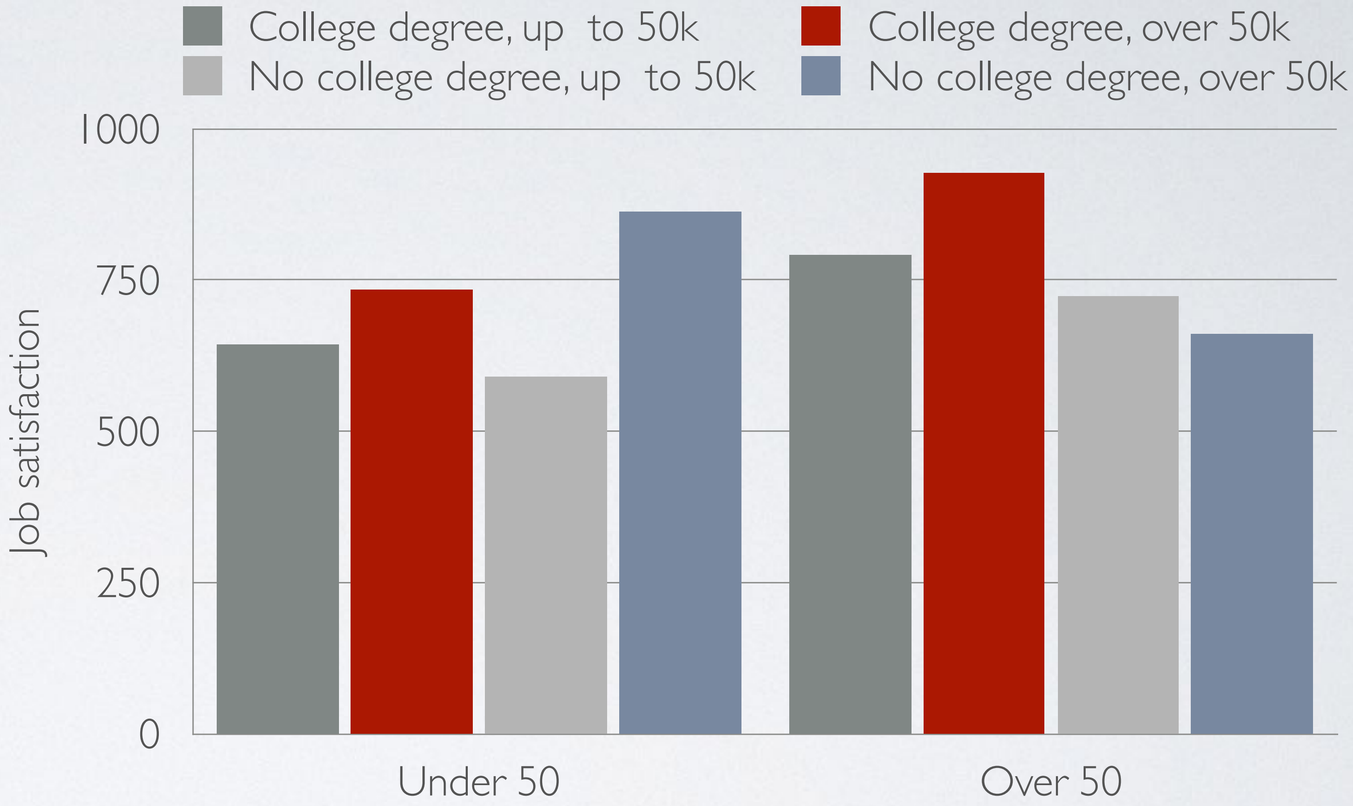
TABLES VS GRAPHS: EXAMPLE

- Even with only 8 numbers, the previous graph did what the table could not: it made the diverging pattern obvious
- This results not only because we used a graph, but because the graph was designed to feature this particular pattern
- The graph on the right for example doesn't show the pattern at all



TABLES VS GRAPHS: EXAMPLE

- Even with only 8 numbers, the previous graph did what the table could not: it made the diverging pattern obvious
- This results not only because we used a graph, but because the graph was designed to feature this particular pattern
- The graph on the right for example doesn't show the pattern at all



- Displaying information in a graph by itself does not make meaningful pattern visible
- We have to design the graph to feature evidence of the particular story we want to tell

WHEN TO USE TABLES VS WHEN TO USE GRAPHS

Use tables when

- Look up individual values
- Compare individual values
- Precise values are required
- The quantitative values require more than one unit of measures
- Both detail and summary values are included

Use graphs when

- The message is contained in the shape of the values (i.e., patterns, trends, exceptions)
- Reveal relationships among whole set of values

TYPE OF TABLES: QUANTITATIVE TO CATEGORICAL

Between one set of quantitative values and one set of categorical subdivision

Salesperson	QTD sales
Robert Jones	13'803
Mandy Rodriguez	20'374
Terri Moore	28'520
Jennifer Taylor	34'786

Between one set of quantitative values and the intersection of multiple categories

Salesperson	Jan	Feb	Mar
Robert Jones	2'834	4'839	6'131
Mandy Rodriguez	5'890	6'482	8'002
Terri Moore	7'398	9'374	11'788
Jennifer Taylor	9'375	12'387	13'024

FUNCTION: LOOK-UP

TYPE OF TABLES: QUANTITATIVE TO CATEGORICAL

Between one set of quantitative values and the intersection of hierarchical categories

Line of product	Product family	Product	Sales
Hardware	Printer	PPS	6'131
Hardware	Printer	PXT	8'001
Hardware	Router	RRZ	11'748
Hardware	Router	RTS	13'024
Software	Business	ACT	12'393
Software	Business	SPR	5'392
Software	Games	ZAP	10'363
Software	Games	ZOW	13'881

FUNCTION: LOOK-UP

TYPE OF TABLES: QUANTITATIVE TO QUANTITATIVE

Among one set of quantitative values
associated with multiple categorical subdivisions

Salesperson	Jan	Feb	Mar
Robert Jones	2'834	4'839	6'131
Mandy Rodriguez	5'890	6'482	8'002
Terri Moore	7'398	9'374	11'788
Jennifer Taylor	9'375	12'387	13'024

FUNCTION: COMPARISON

TYPE OF TABLES: QUANTITATIVE TO QUANTITATIVE

Among distinct sets of quantitative values associated with the same categorical subdivision

Salesperson	Sales	Returns	Net sales
Robert Jones	2'834	4'839	6'131
Mandy Rodriguez	5'890	6'482	8'002
Terri Moore	7'398	9'374	11'788
Jennifer Taylor	9'375	12'387	13'024

FUNCTION: COMPARISON

TABLE DESIGN: UNIDIRECTIONAL TABLES

Department	Headcount	Expenses
Finance	26	202'202
Sales	93	983'393
Operations	107	933'200
Total	226	2'118'795

Categorical subdivision organised down the rows

Categorical subdivision arranged across the columns

Department	Finance	Sales	Operations	Total
Headcount	26	93	107	226
Expenses	202'202	983'393	933'200	2'118'795

TABLE DESIGN: UNIDIRECTIONAL TABLES

Table with 2 sets of categorical subdivisions

Department	Expense type	Expenses
Finance	Compensation	160'383
Finance	Supplies	5'038
Finance	Travel	10'385
Sales	Compensation	683'879
Sales	Supplies	193'378
Sales	Travel	125'705
Total		1'178'768

TABLE DESIGN: BIDIRECTIONAL TABLES

Also called crosstab or pivot tables, they have categorical subdivisions both on the rows and on the columns. Bidirectional tables make a more efficient use of space.

Department	Finance	Sales	Total
Compensation	160'383	683'879	844'262
Supplies	5'038	193'378	198'416
Travel	10'385	125'705	136'090
Total	175'806	1'002'962	1'178'768

SUMMARY OF TABLE VARIATIONS

Relationship	Unidirectional	Bidirectional
Quantitative to categorical relationships		
Between one set of quantitative values and one set of categorical subdivision	Yes	Not applicable
Between one set of quantitative values and the intersection of multiple categories	Yes, sometimes better because of convention	Yes, it saves space
Between one set of quantitative values and the intersection of hierarchical categories	Yes, it clearly shows the hierarchy	Yes, although the hierarchy doesn't stand out

SUMMARY OF TABLE VARIATIONS

Relationship	Unidirectional	Bidirectional
Quantitative to categorical relationships		
Between one set of quantitative values and one set of categorical subdivision	Yes	Not applicable
Between one set of quantitative values and the intersection of multiple categories	Yes, sometimes better because of convention	Yes, it saves space
Between one set of quantitative values and the intersection of hierarchical categories	Yes, it clearly shows the hierarchy	Yes, although the hierarchy doesn't stand out
Quantitative to quantitative relationships		
Among one set of quantitative values associated with multiple categorical subdivisions	Yes	Yes. Works well because quantitative values are close together for easy comparison
Among distinct sets of quantitative values associated with the same categorical subdivision	Yes	Yes, although it tends to get messy as you add multiple sets

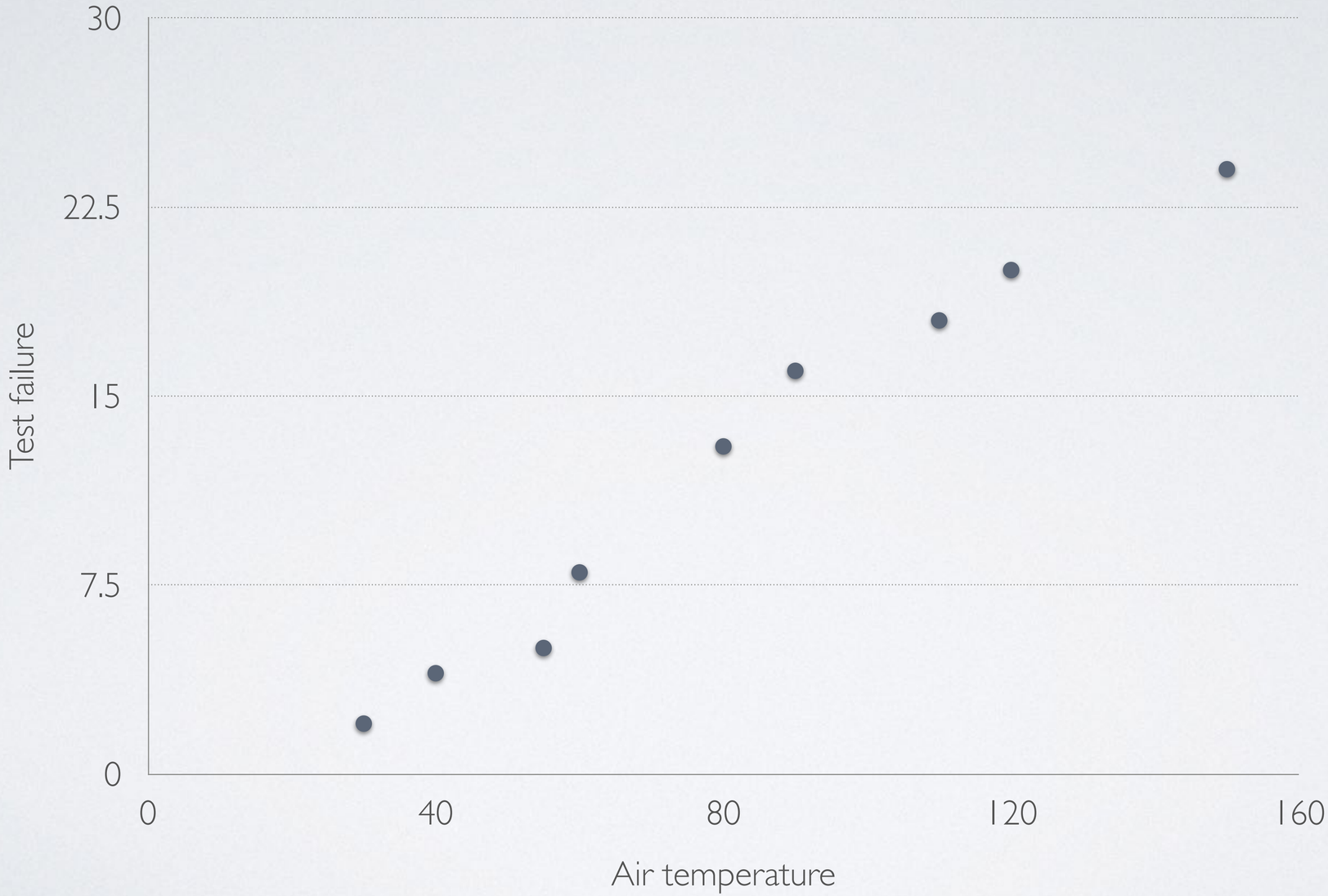
GRAPHS VARIATIONS

- Different type of quantitative information requires different types of graphs
- Graphs displays information about relationships
- Their strength is the ability to present complex relationships so that we can see them quickly and easily
- Variations in graphs are defined primarily by differences in the components that encode quantitative values

ENCODING QUANTITATIVE VALUES: POINTS

- A simple small geometric shape used to mark a specific location on the graph
- Often they are simple dots
- Position maps the quantitative values
- They are seldom used alone without lines to connect them expect in scatter plot, where their small size and lack of direction give them the unique ability to simultaneously encode values along both the X and Y axes
- Points + lines: not redundant, points highlights individual values, lines help us focus on the changes between values and the overall trend of those changes

EXAMPLE: SCATTER PLOT



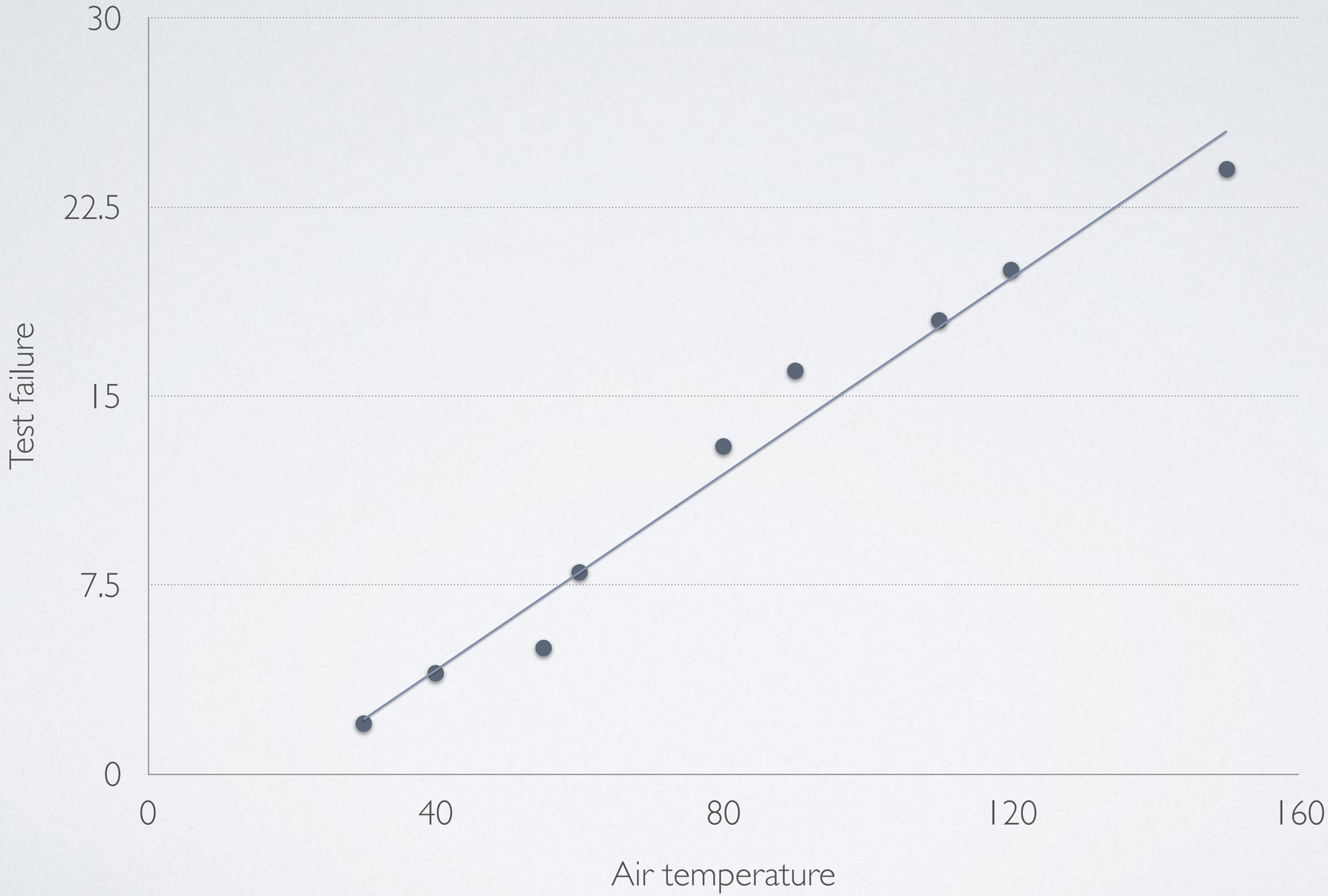
ENCODING QUANTITATIVE VALUES: LINES

- A collection of contiguous individual points extending in a single direction through space
- Encode quantitative values in two ways:
    - Connect individual data points
    - To display the trend of a series of data points (linear regression in scatter plot)
- When a line is used to connect individual values, is not necessary to display those values as points in addition to the line

EXAMPLE: TIME SERIES

EXAMPLE: TIME SERIES

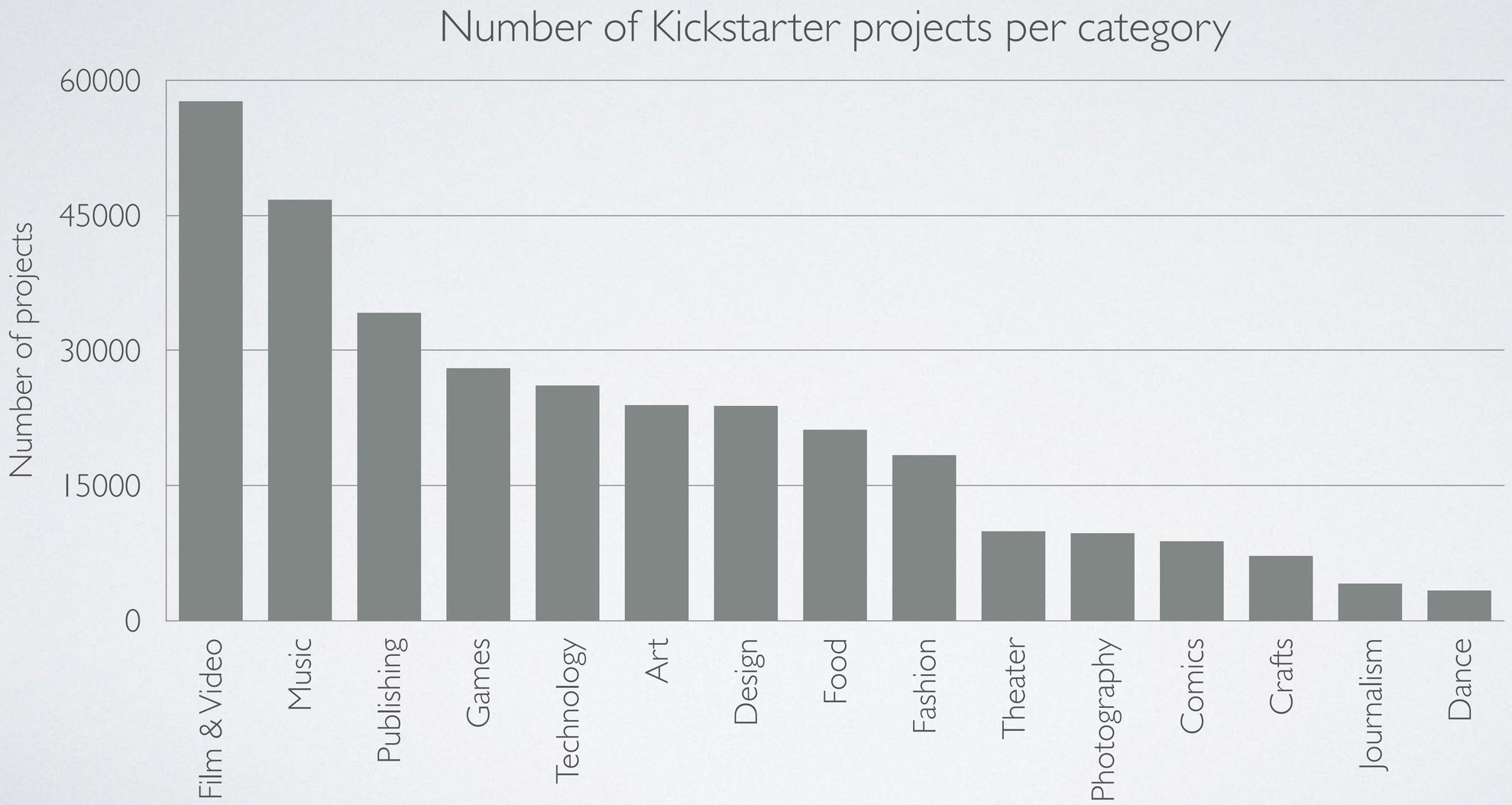
EXAMPLE: LINE + POINTS - TREND LINE OR LINE BEST FIT



ENCODING QUANTITATIVE VALUES: BARS

- A thick line (theoretically a rectangle)
- The width is not intended to convey any meaning in a graph
- Quantitative values are mapped into the length and not the position
- Important that they have the same thickness since our brain assigns bigger values to things that are bigger - so it might incorrectly influence our interpretation
- Can be vertical or horizontal
- We could use points mapping the position to value
    - Due to their visual weight, they stand out so clearly and distinctly from one another that they do a great job of representing individual values discretely

EXAMPLE:


Number of Kickstarter projects per category
Number of projects
60000
45000
30000
15000
0
Film & Video
Music
Publishing
Games
Technology
Art
Design
Food
Fashion
Theater
Photography
Comics
Crafts
Journalism
Dance

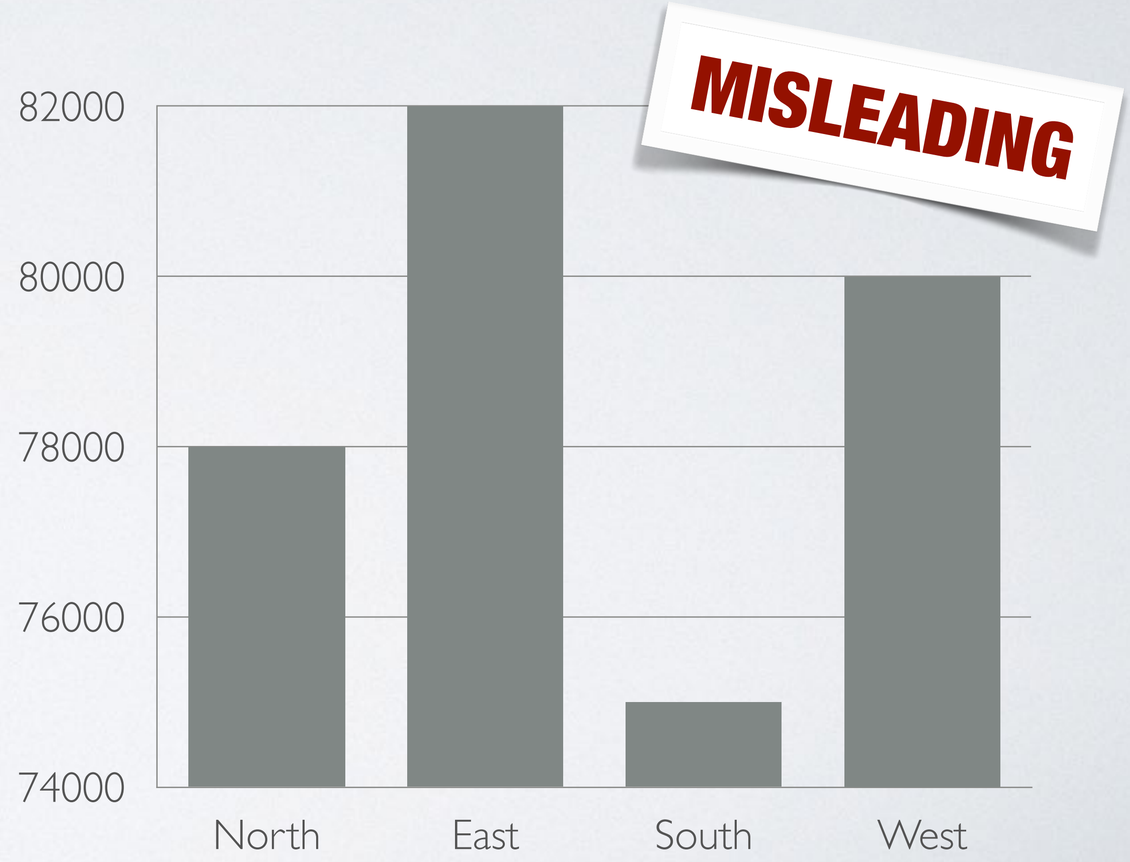

82000
61500
41000
20500
0
North
East
South
West


82000
61500
41000
20500
0
North
East
South
West


82000
80000
78000
76000
74000
North
East
South
West

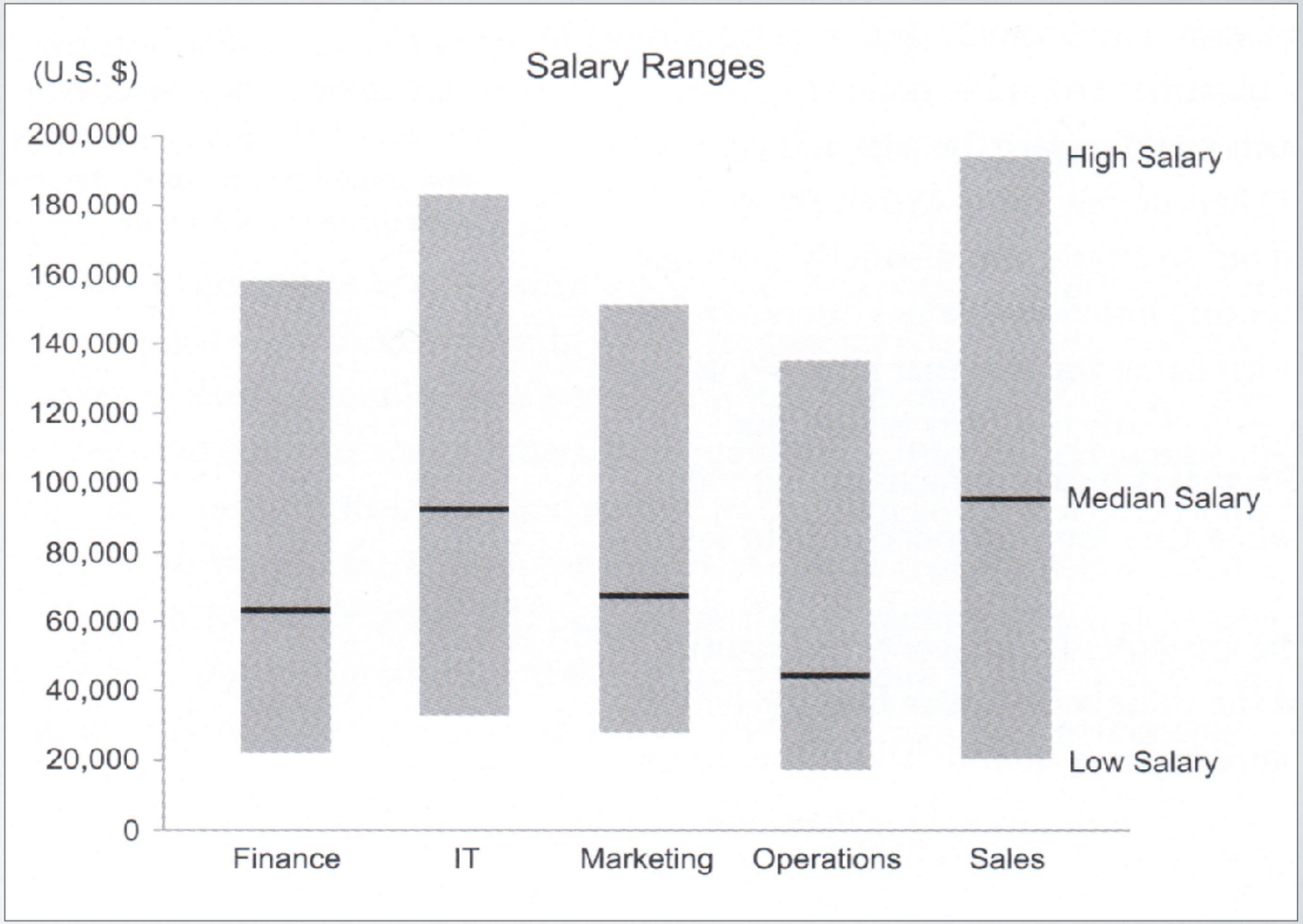
BARS SHOULD ALWAYS START AT ZERO

Because the length of a bar encode its quantitative value, its base should always begin at the value zero, otherwise it would not support accurate comparison of quantitative values



ENCODING QUANTITATIVE VALUES: BOXES

- Less known than all other types
- Used to compare distribution
- Encode an entire range of values rather than a single value
- We can compare the median, the lowest/highest values, and the range



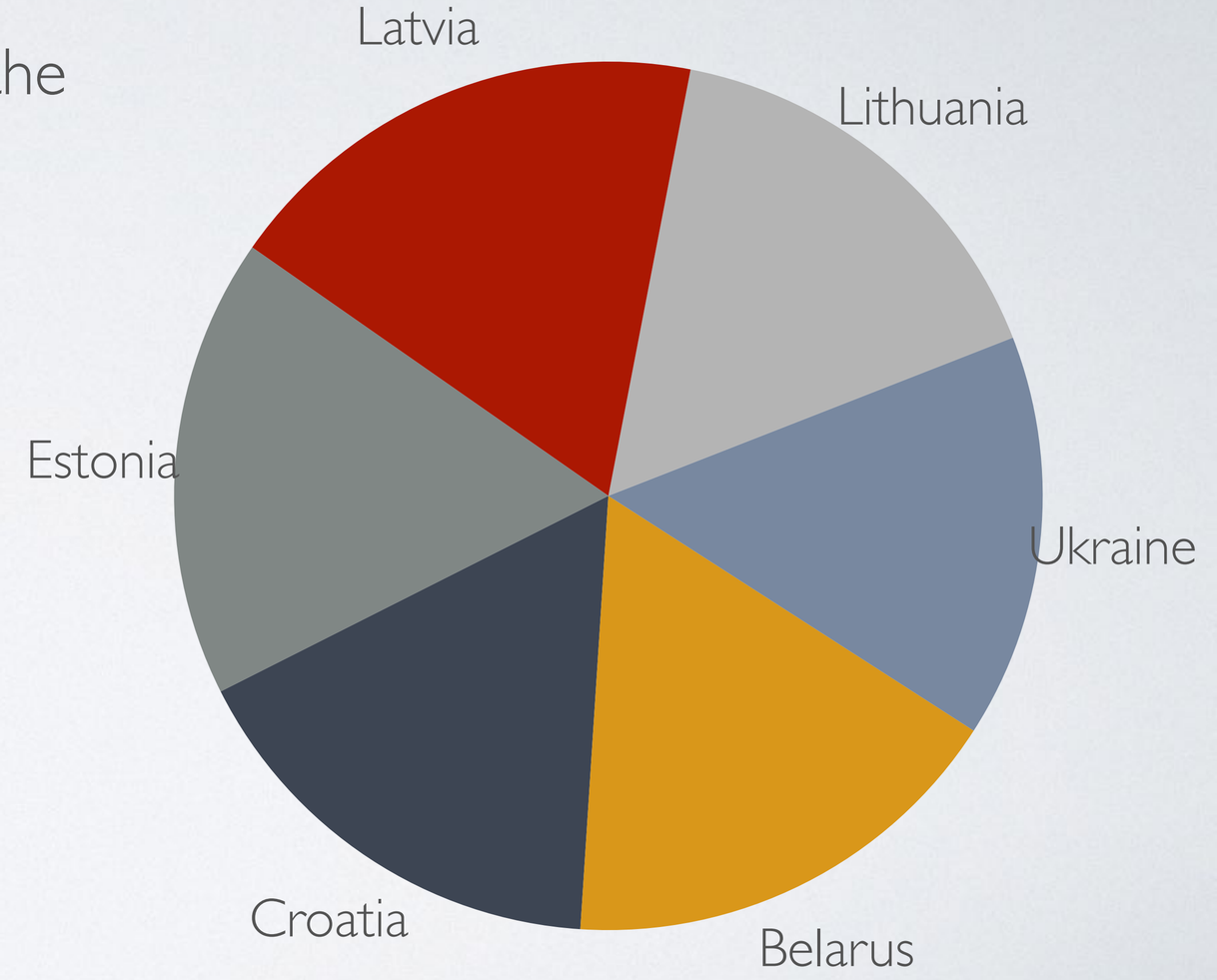
ENCODING QUANTITATIVE VALUES: SHAPES WITH 2D AREA

- Bi-dimensional shapes that represent value proportional to area rather than location on the graph

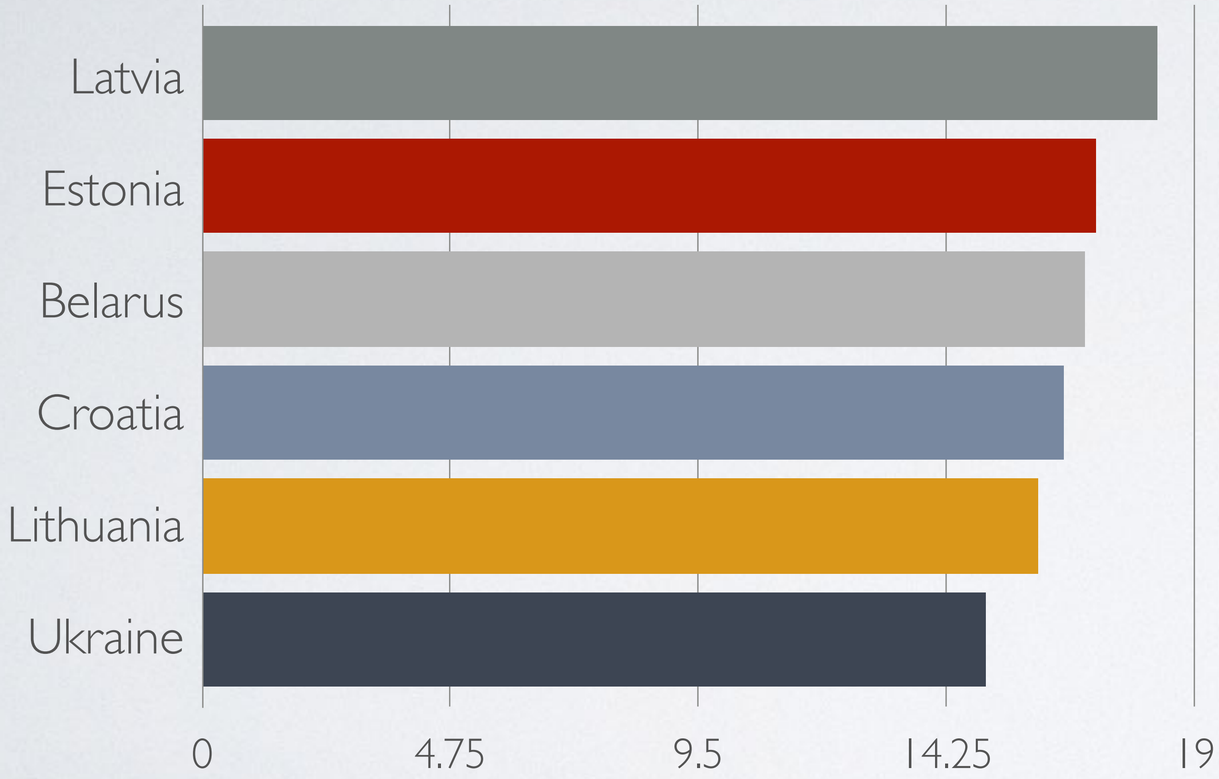
- Example: pie chart

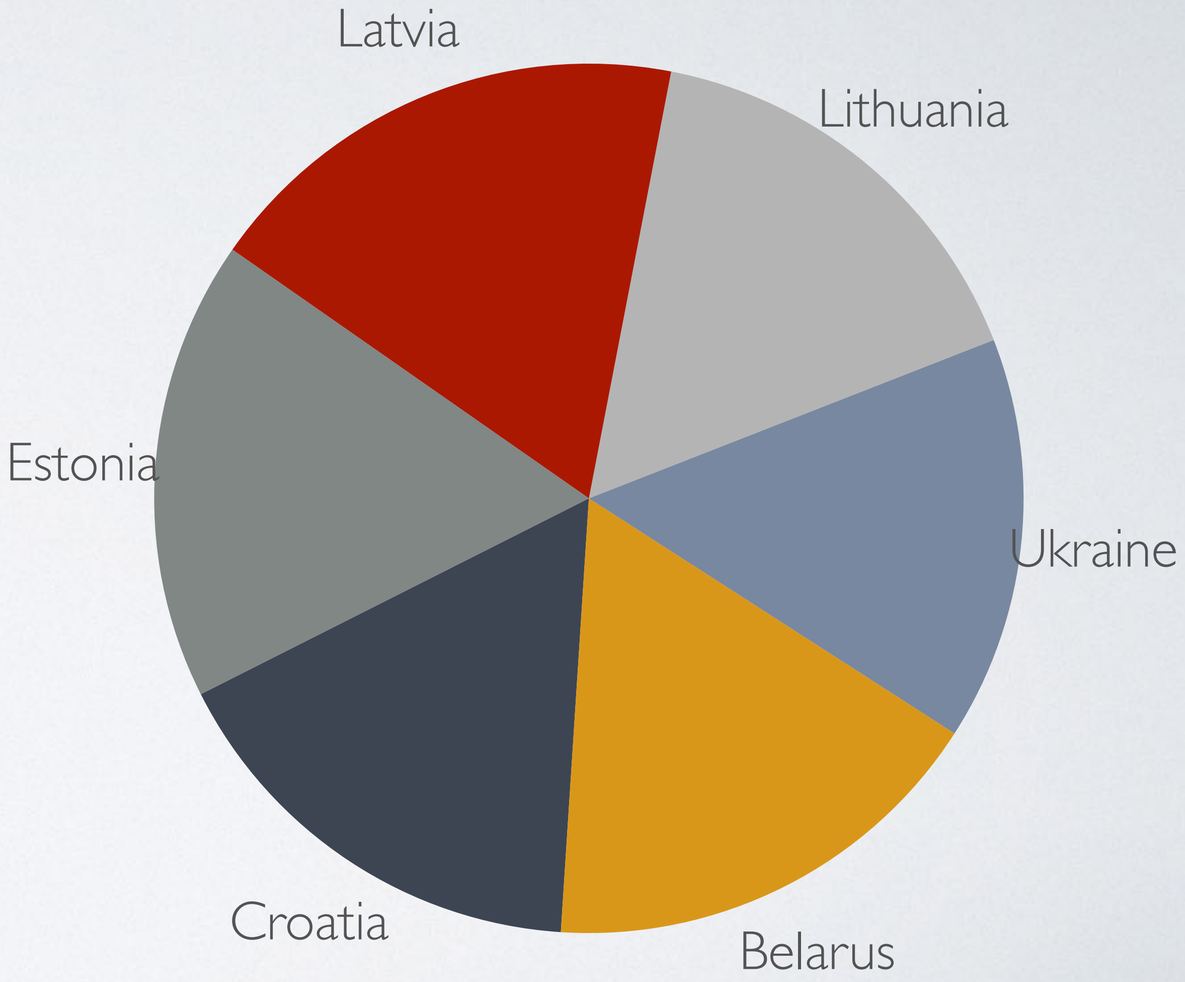
  - It is part of a larger family of area graphs which use 2D area to encode quantitative values

  - The axis is not a straight line but the perimeter of the circle



ENCODING QUANTITATIVE VALUES: SHAPES WITH 2D AREA





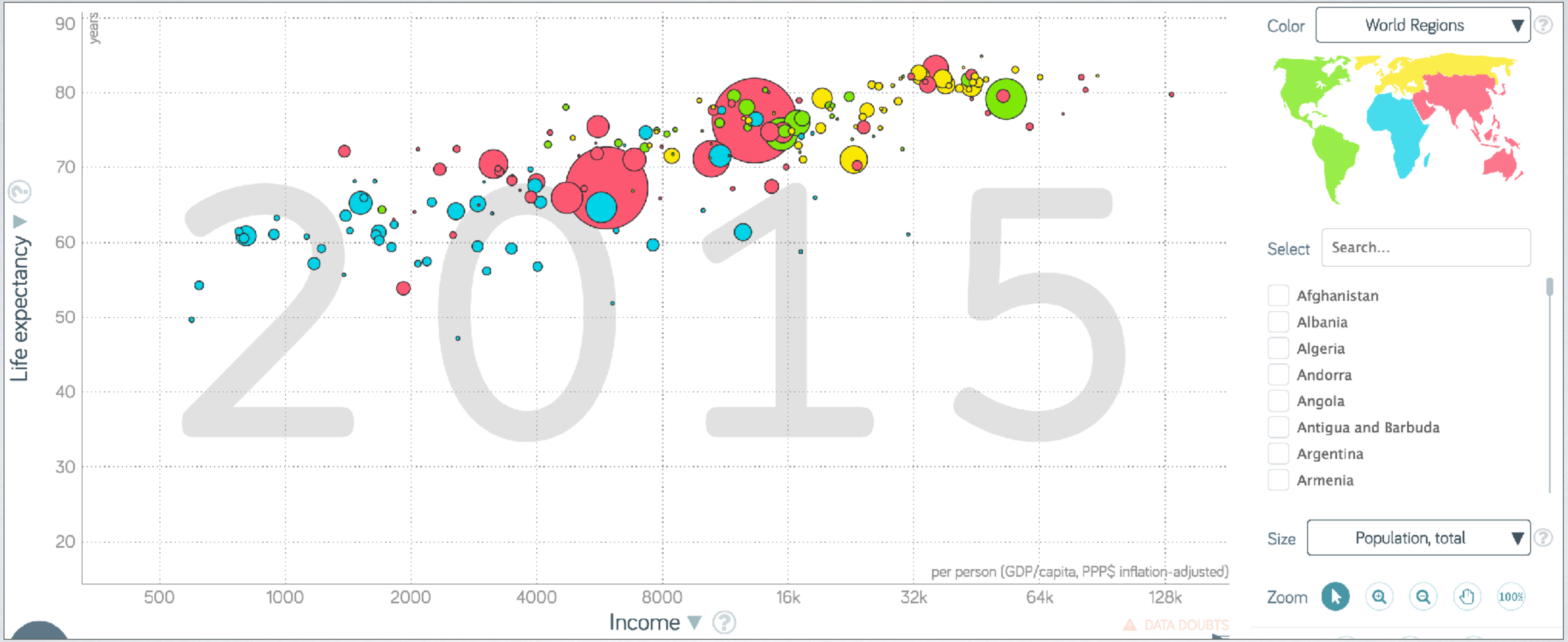
2D AREA IS NOT EFFECTIVE TO ENCODE QUANTITATIVE VALUES

- Better to abandon pie chart: they are difficult to read and communicate poorly (all 2D areas but especially pie)
- Our visual perception is not designed to assign quantitative values to 2D areas (and it is even harder when a 3rd dimension is added)
- Slices are difficult to measure (and rectangles too)
  - If they are close in size, impossible to tell which one is larger
  - If not close in size, you can't really tell by how much
- There are exceptions, e.g., size in bubble charts, but extraordinary care when using them since they are harder to interpret (normally for correlation of more than 2 variables)

BUBBLES

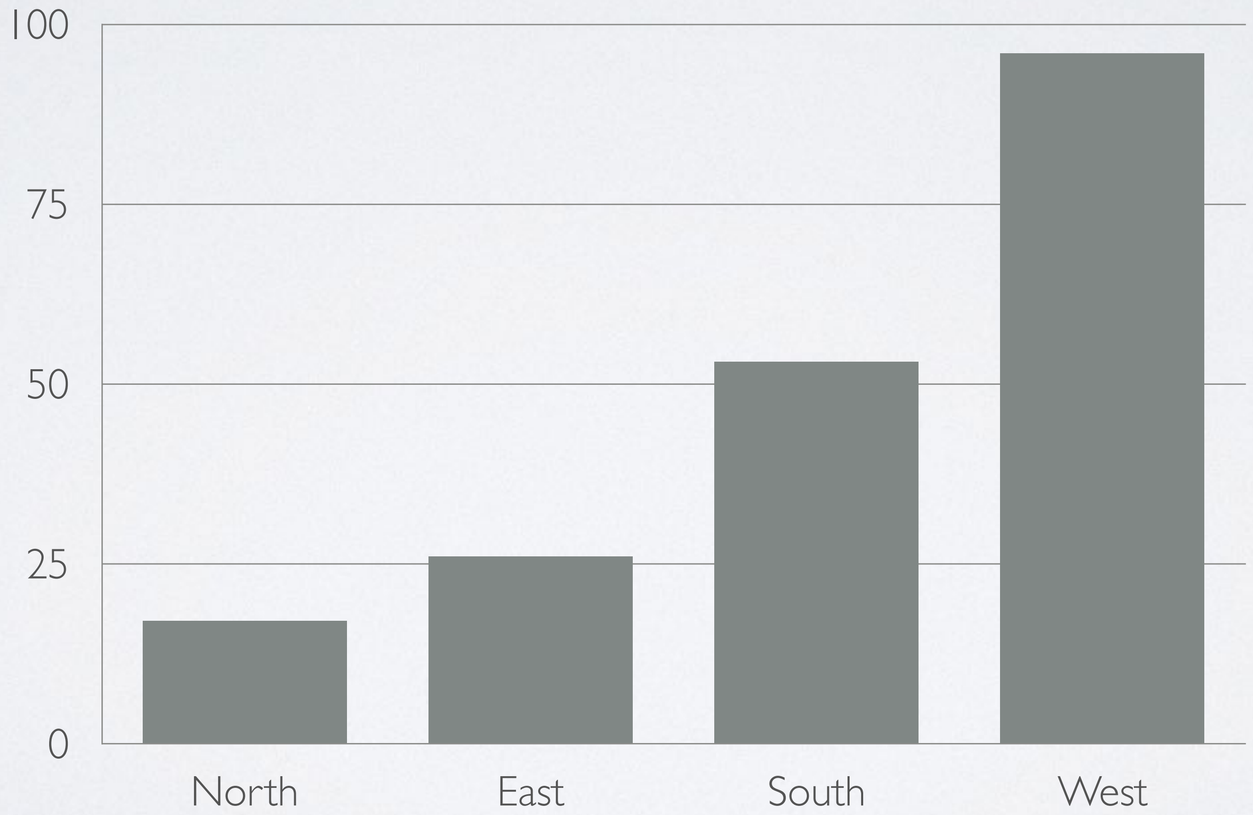
- Another means to encode values in 2D areas involves varying the size of points
- It has the benefit that another set of values can be added
- Use it only when approximate values are all that's needed, and the third variable adds enough to the story to justify the increased complexity

BUBBLES: EXAMPLE



ENCODE CATEGORICAL DATA: POSITION

2D position is the clearest means to associate quantitative values to categories

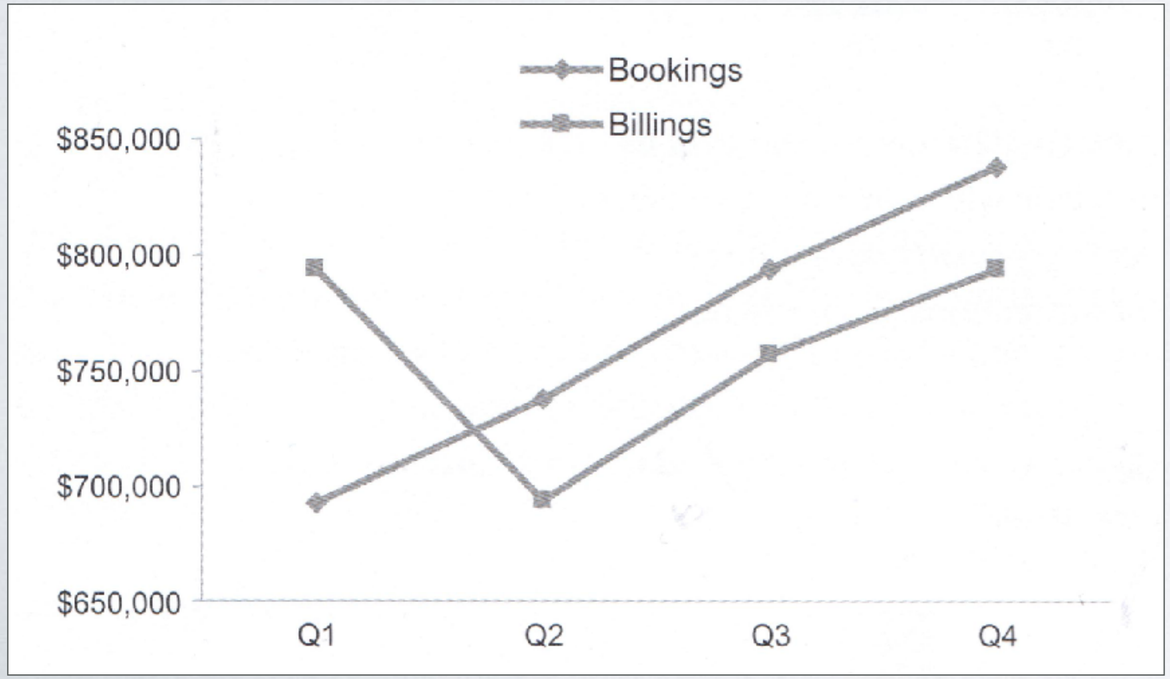


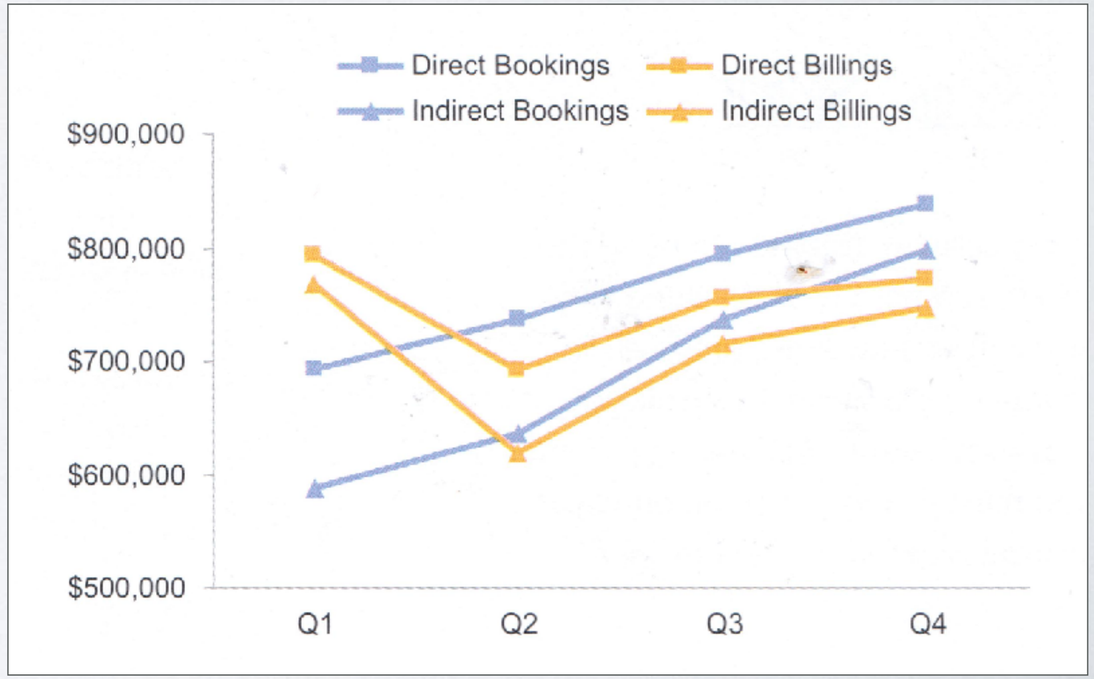
ENCODE CATEGORICAL DATA: COLOR

If the position is already used, to add a distinction between categorical subdivision you can use the color. Note that the reader needs to read the legend and store the mapping in memory

ENCODE CATEGORICAL DATA: POINT SHAPE

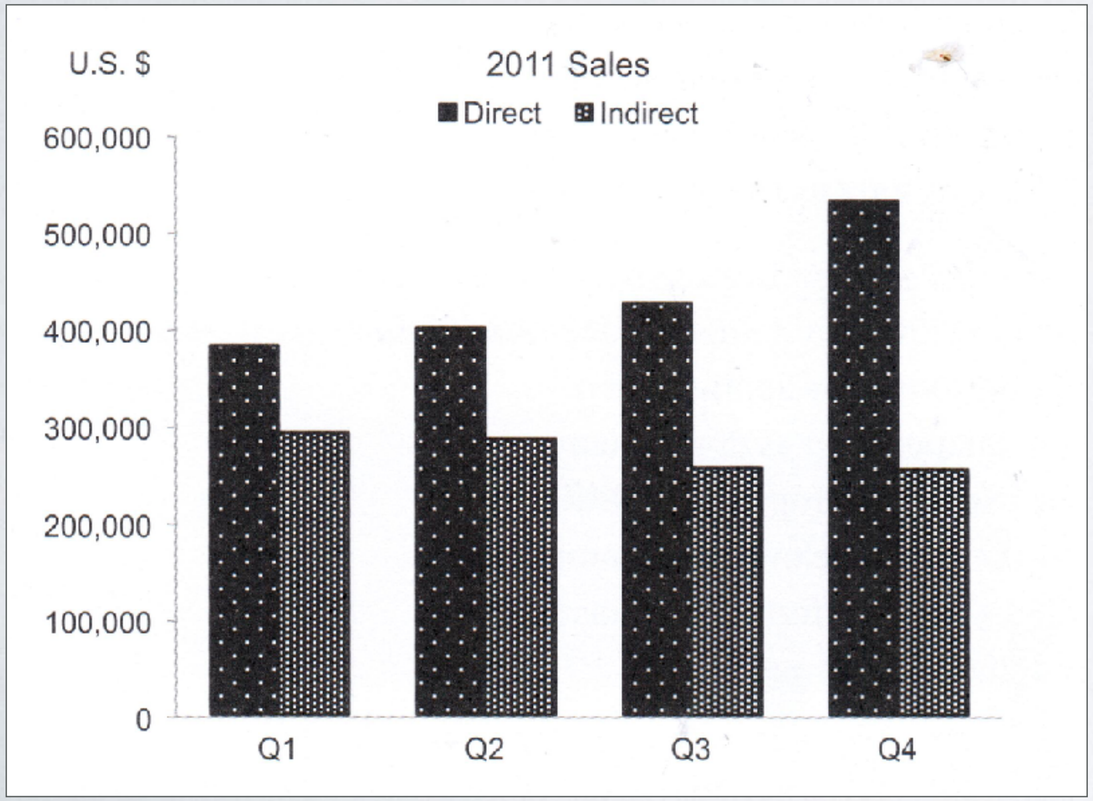
- Only when points are used to encode quantitative values, different shapes can be used to distinguish quantitative values that belong to different categories
- Different shape can be easily distinguished, but not as easily as different colours
- Can be useful when you already used position and color

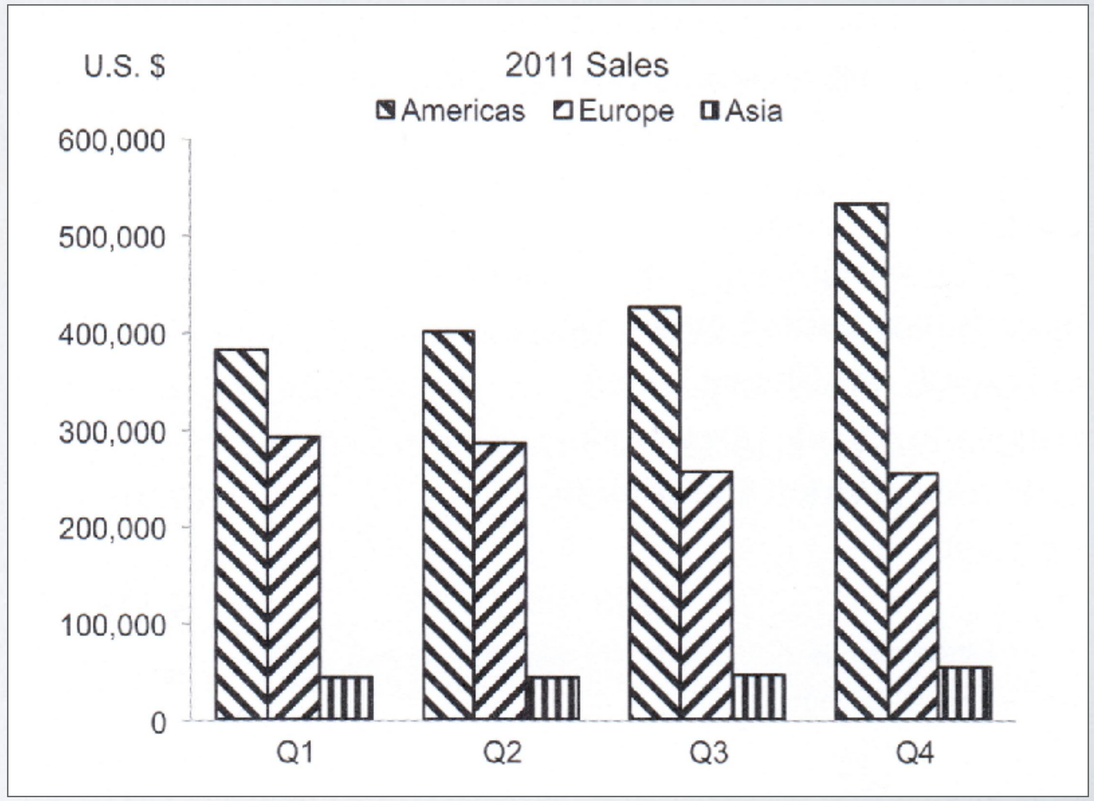




ENCODE CATEGORICAL DATA: FILL PATTERN

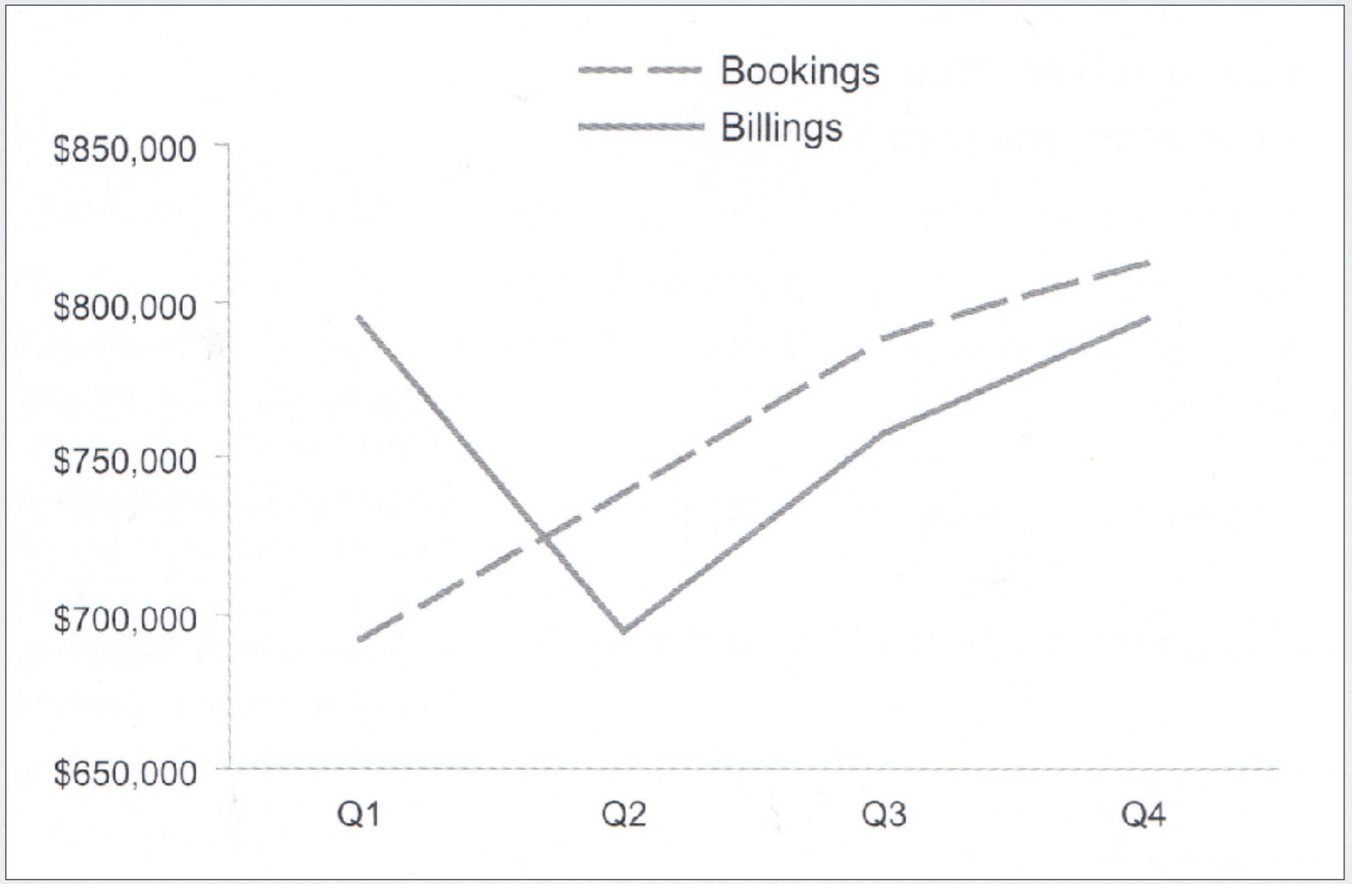
- Only when quantitative values are encoded as bars
- To be used as a last resort because it is harder to distinguish and can cause Moiré vibration (dizzying effect)





ENCODE CATEGORICAL DATA: LINE STYLE

It doesn't work as well as color, because the breaks in the lines disrupt smooth perception of their flow

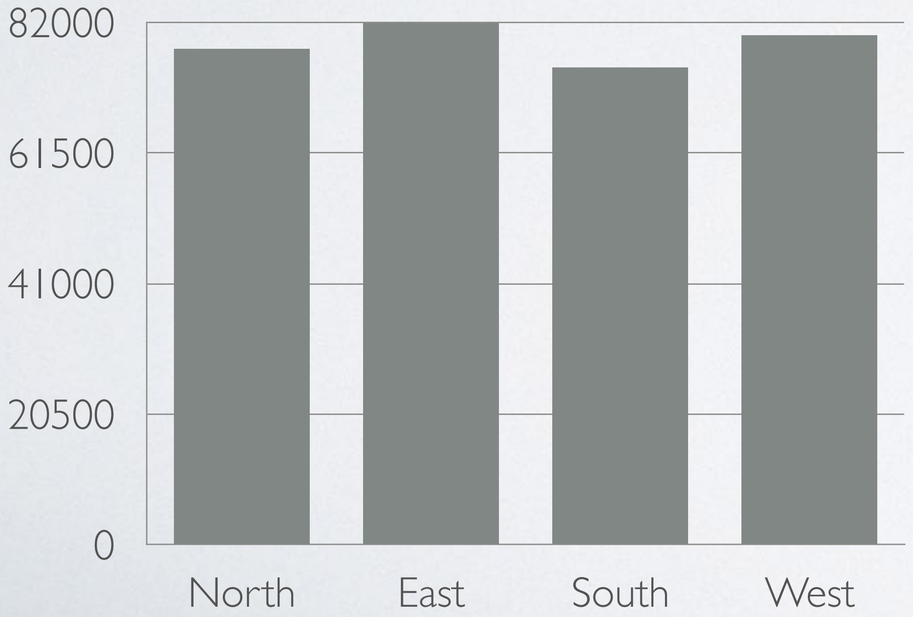


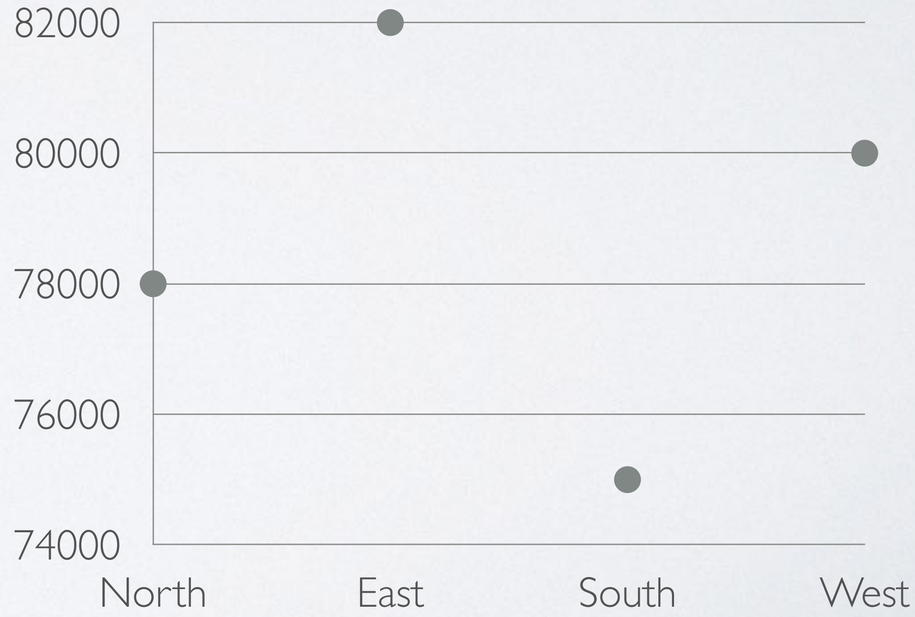
8 TYPES OF RELATIONSHIPS WE USE GRAPHS TO DISPLAY

1. Nominal comparison
2. Time series
3. Ranking
4. Part-to-whole
5. Deviation
6. Distribution
7. Correlation
8. Geospatial

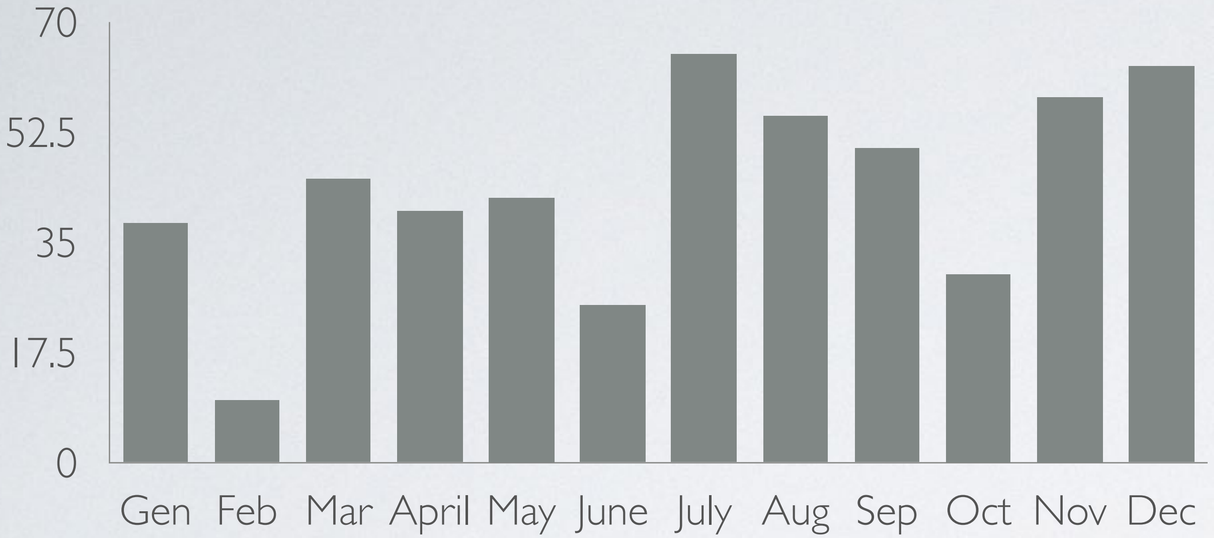
NOMINAL COMPARISON

- Item on the categorical scale have no order because they represent a nominal scale
- Display series of discrete quantitative values to be compared
- We can use horizontal or vertical bars, or points (dot plots) which don't require a zero-based scale



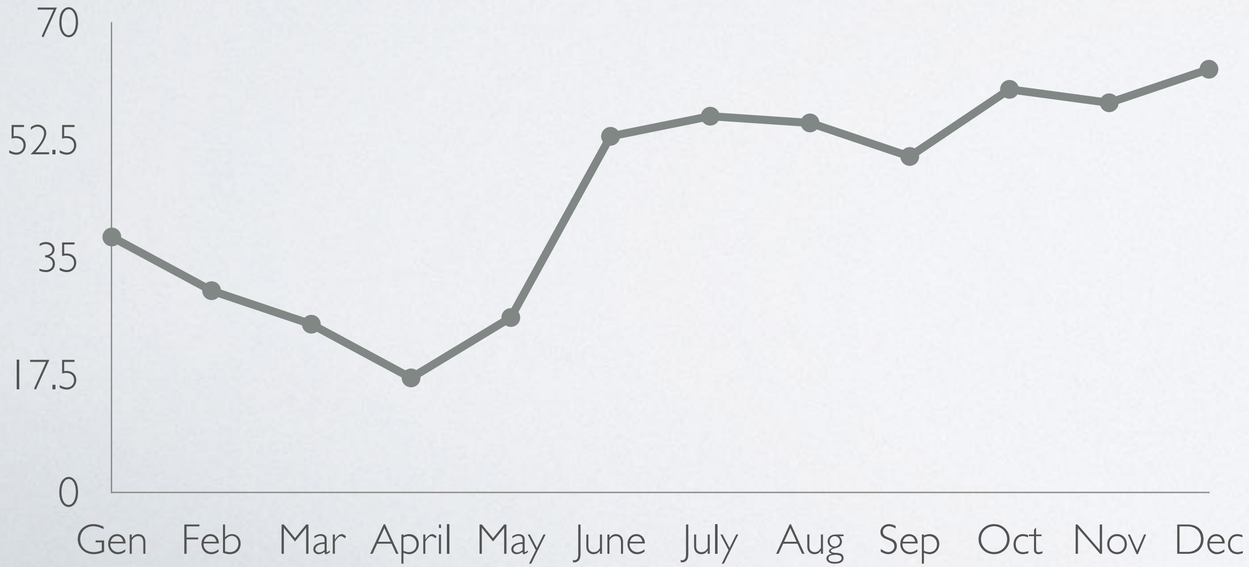


TIME SERIES



- Series of relationships between quantitative values that are associated with categorical subdivision of time (e.g., year, millisecond)
- It is a powerful means to show trends over time
- It is obvious and well established convention how to display time
- Using bars is not effective to show the pattern of change over time. Using dots (only) does not give the sense of continuity that is required to display time

TIME SERIES



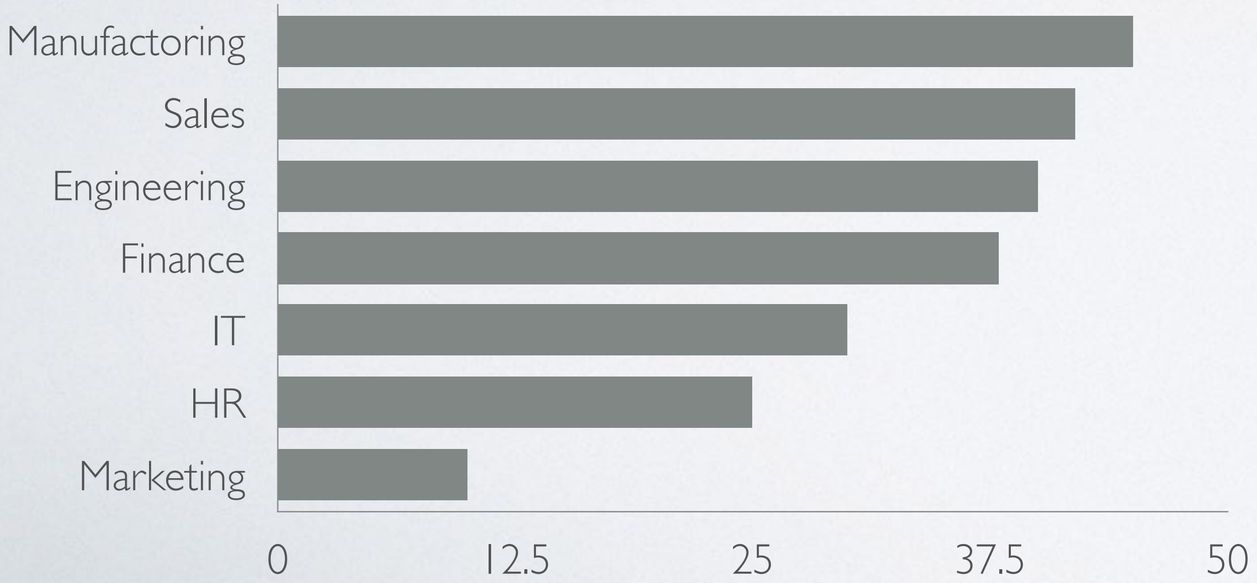
- Series of relationships between quantitative values that are associated with categorical subdivision of time (e.g., year, millisecond)
- It is a powerful means to show trends over time
- It is obvious and well established convention how to display time
- Using bars is not effective to show the pattern of change over time. Using dots (only) does not give the sense of continuity that is required to display time

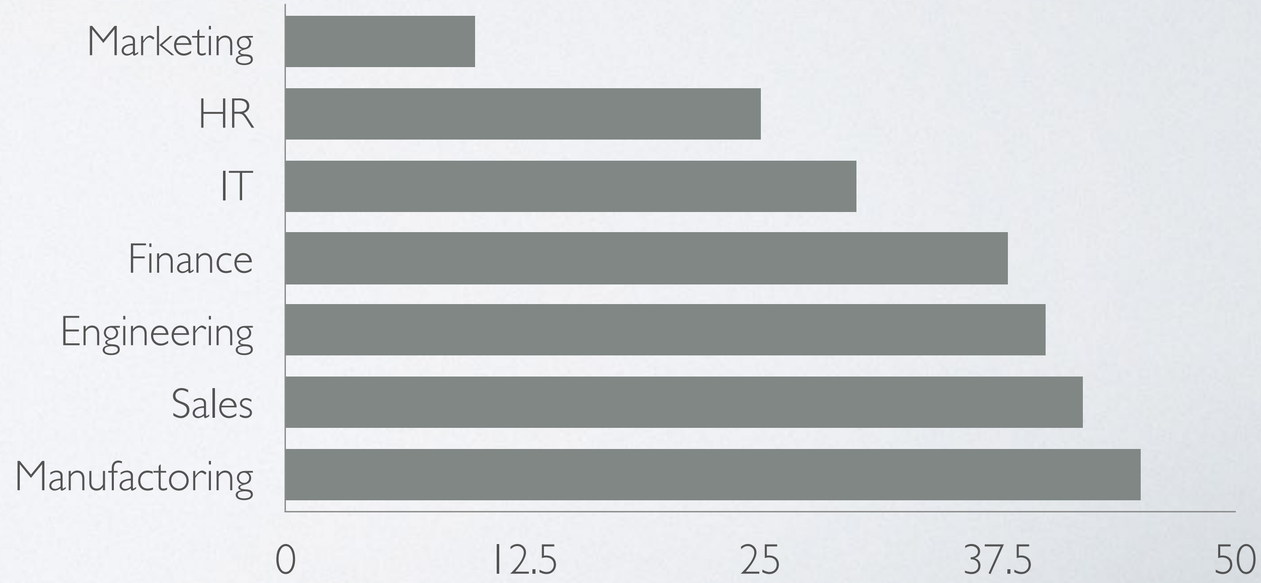
RANKING

- How individual quantitative values associated with a set of categorical subdivision relate to each other sequentially sorted by size

- Default choice is bar, although points can be used in case non-zero start is needed

    - Highlight the highest/lowest: highest/lowest on top (vertical) or left (western culture, read left to right)

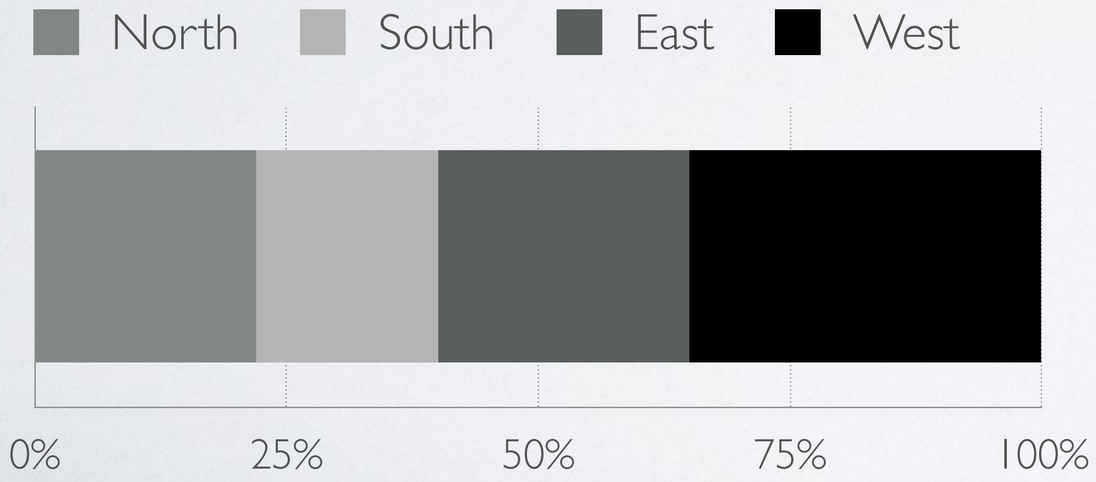




PART-TO-WHOLE

- How individual quantitative values associated with a set of categorical subdivision relate to the complete set of values and thus to each other

- No pie chart, right? So what should we use?

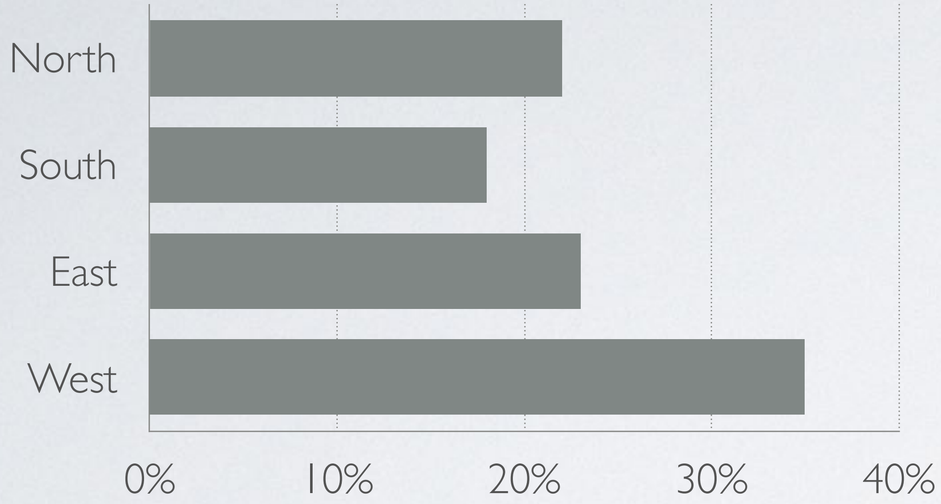


Stacked bar graphs

- Only slightly better than pie chart

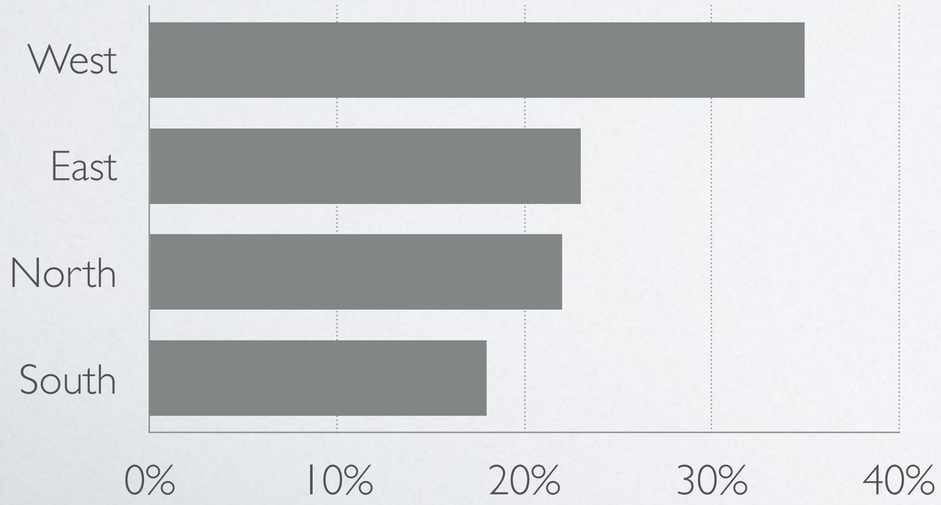
- Easy to determine the percentage associated with the north region, but harder for any other region

PART-TO-WHOLE



Unstacking

- Display each component as a separate bar
- It expresses clearly the contribution of each component



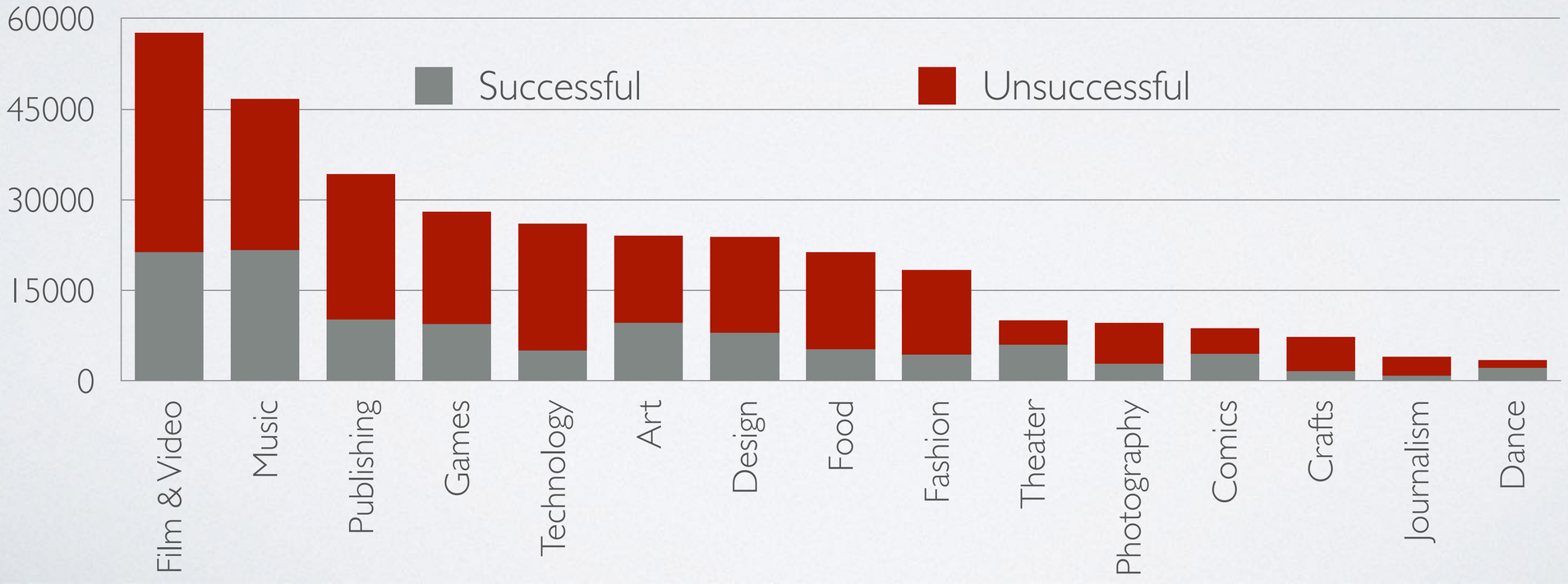
Unstacking + sorting

- It makes it easier to compare values that are close in size

PART-TO-WHOLE

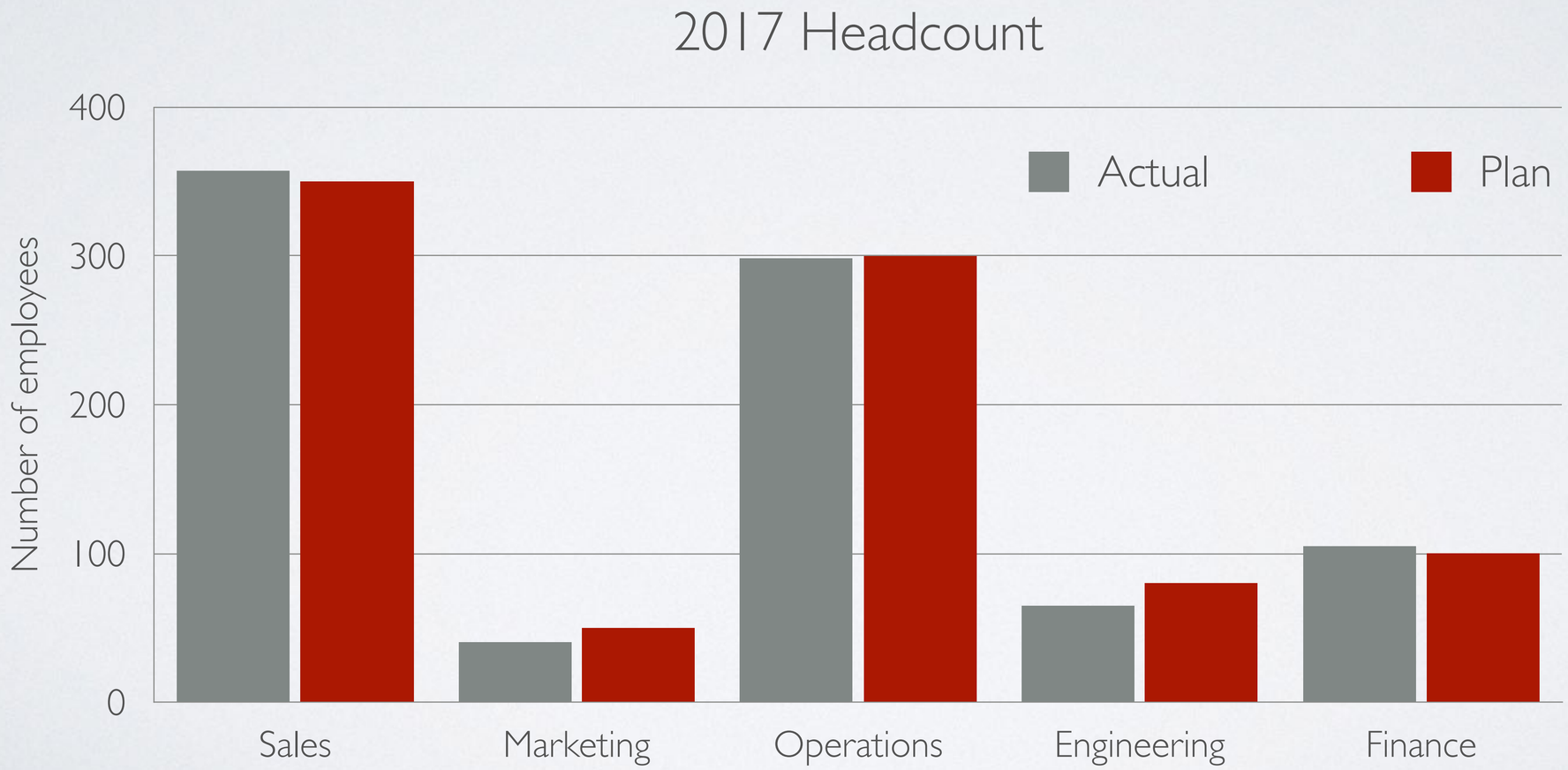
- Points and lines don't work, as they don't carry enough visual weight

- Stacked bar graphs can be useful, despite the disadvantages associated with stacked bars, to display the whole using a unit of measure other the percentage, but also wish to provide an approximate sense of the relative proportion of the parts



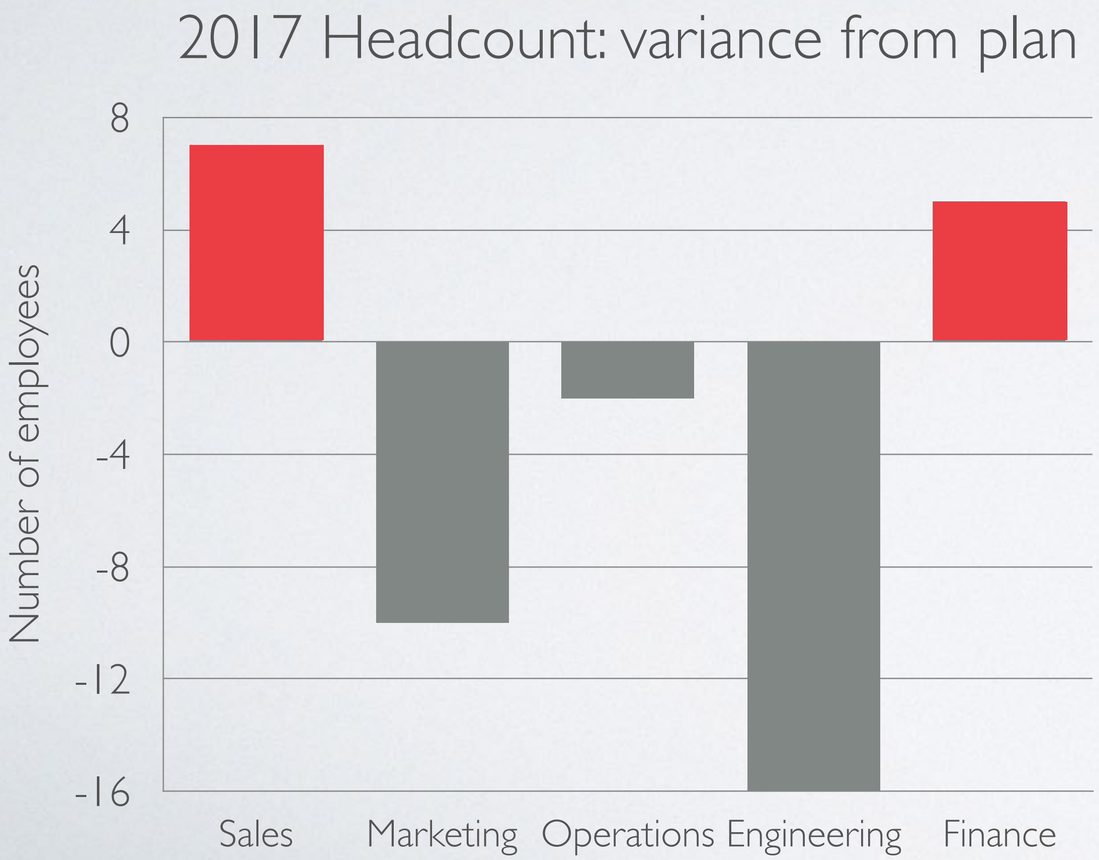
DEVIATION

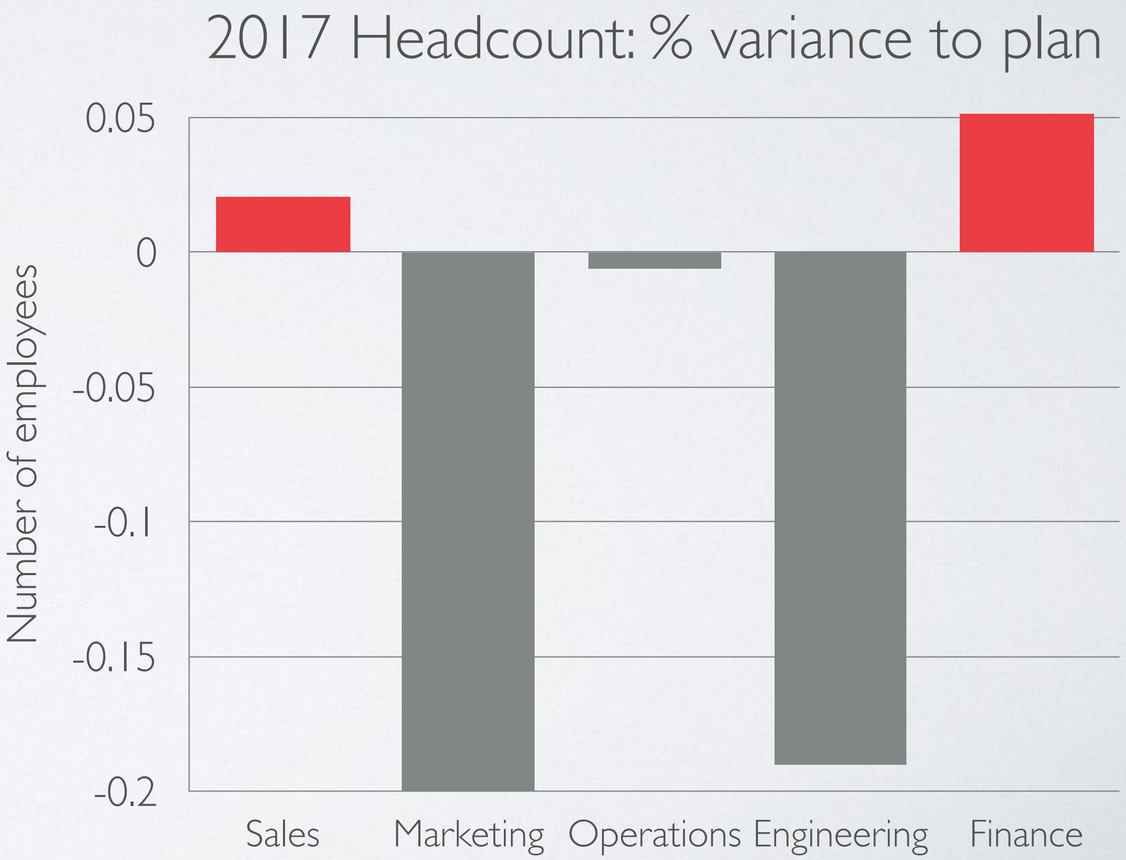
How one or more sets of quantitative values differ from a primary set of values



DEVIATION

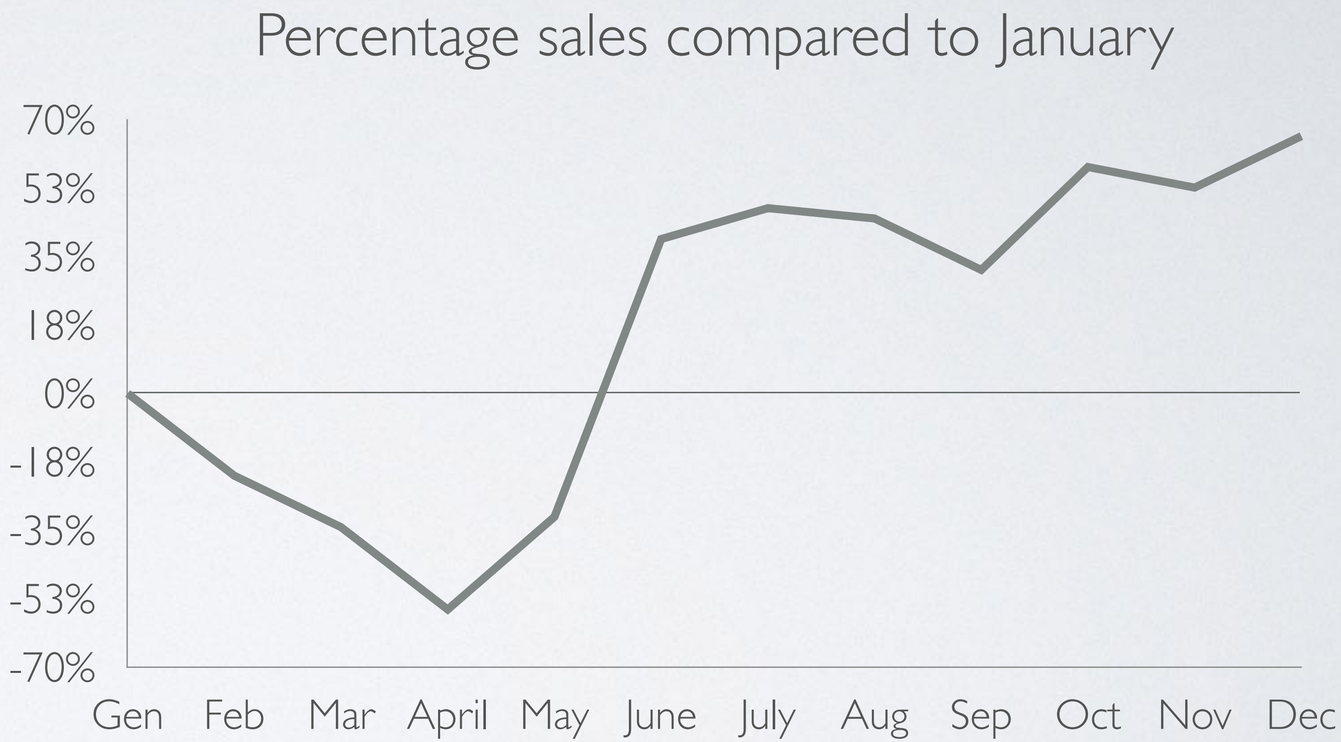
A graph should always display as directly as possible the information that the reader needs. Deviation can be better expressed as a difference, in relation to a primary set





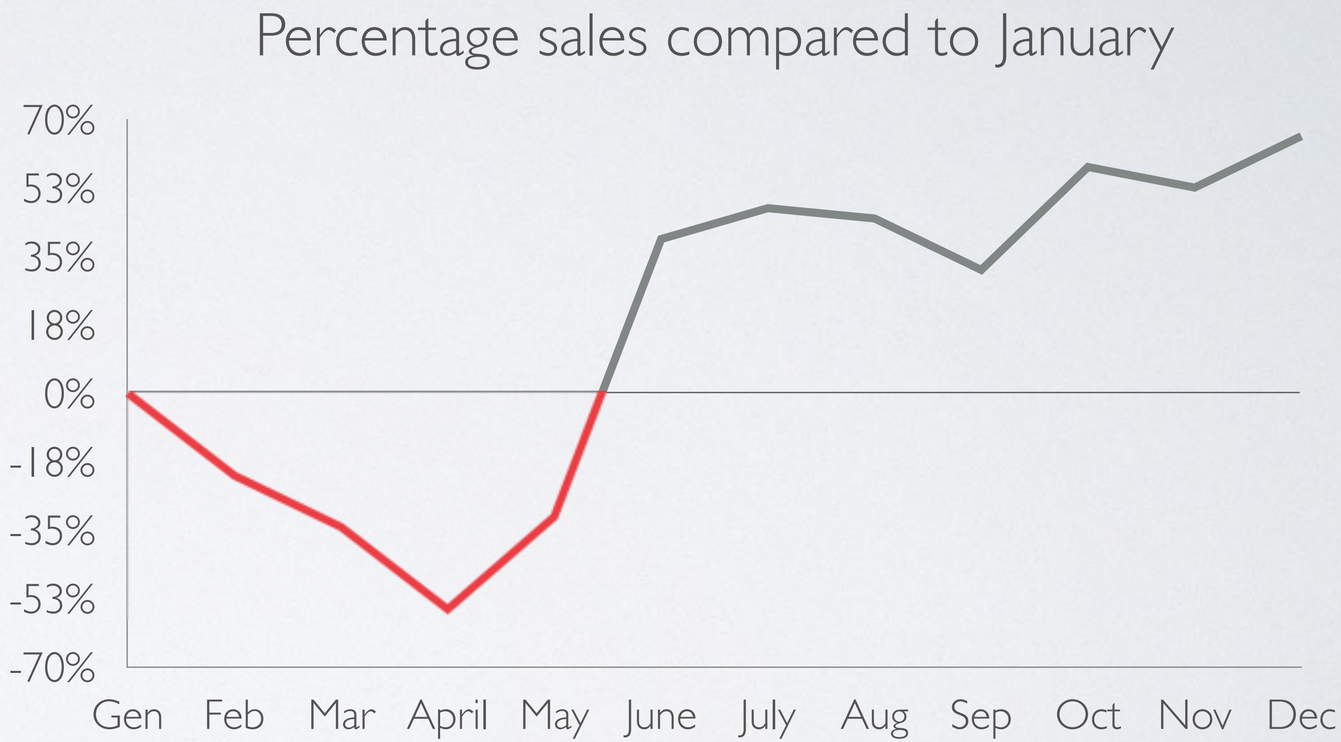
DEVIATION

- Deviation relationships are often teamed with other relationships
    - time series (most common), using lines
    - Ranking or part-to-whole, using bars (as example before)



DEVIATION

- Deviation relationships are often teamed with other relationships
    - time series (most common), using lines
    - Ranking or part-to-whole, using bars (as example before)
- Colors can be used to encode negative or particularly positive situations



DISTRIBUTION

- How one or more sets of quantitative values are distributed across their full quantitative range, from lowest to highest

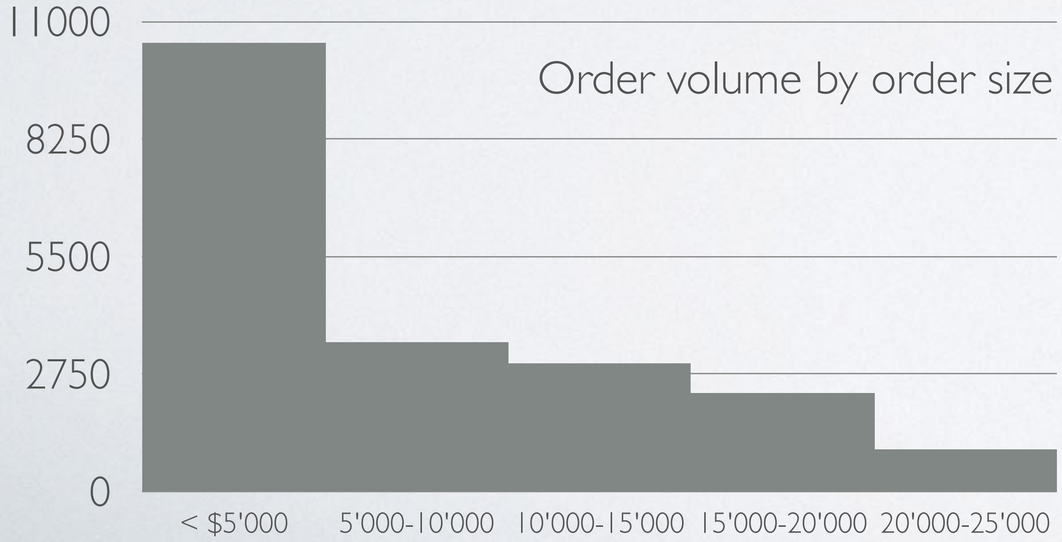
- Single distribution: how often something occurs, distributed across a series of consecutive quantitative ranges (a.k.a. intervals). It is called frequency distribution

DISTRIBUTION: HISTOGRAM

- How one or more sets of quantitative values are distributed across their full quantitative range, from lowest to highest

- Single distribution: how often something occurs, distributed across a series of consecutive quantitative ranges (a.k.a. intervals). It is called frequency distribution

Histogram: if you want to emphasise the number of occurrences in each range



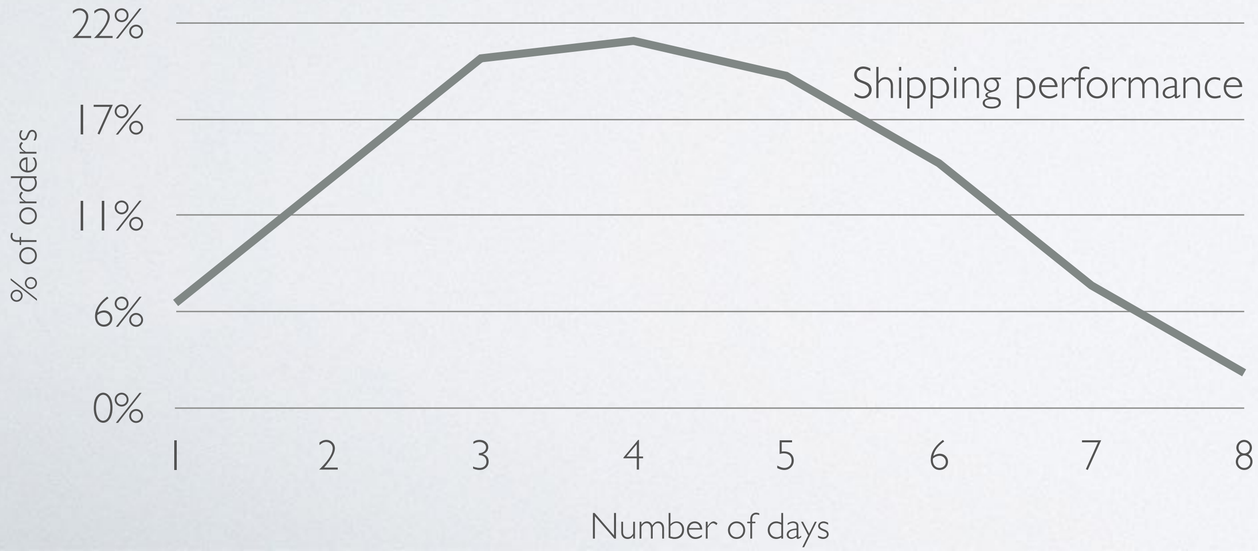
Horizontal bars would work equally well, but the convention is to use vertical bars

DISTRIBUTION: FREQUENCY POLYGON

- How one or more sets of quantitative values are distributed across their full quantitative range, from lowest to highest

- Single distribution: how often something occurs, distributed across a series of consecutive quantitative ranges (a.k.a. intervals). It is called frequency distribution

Frequency polygon: if you want to emphasise the shape of the distribution



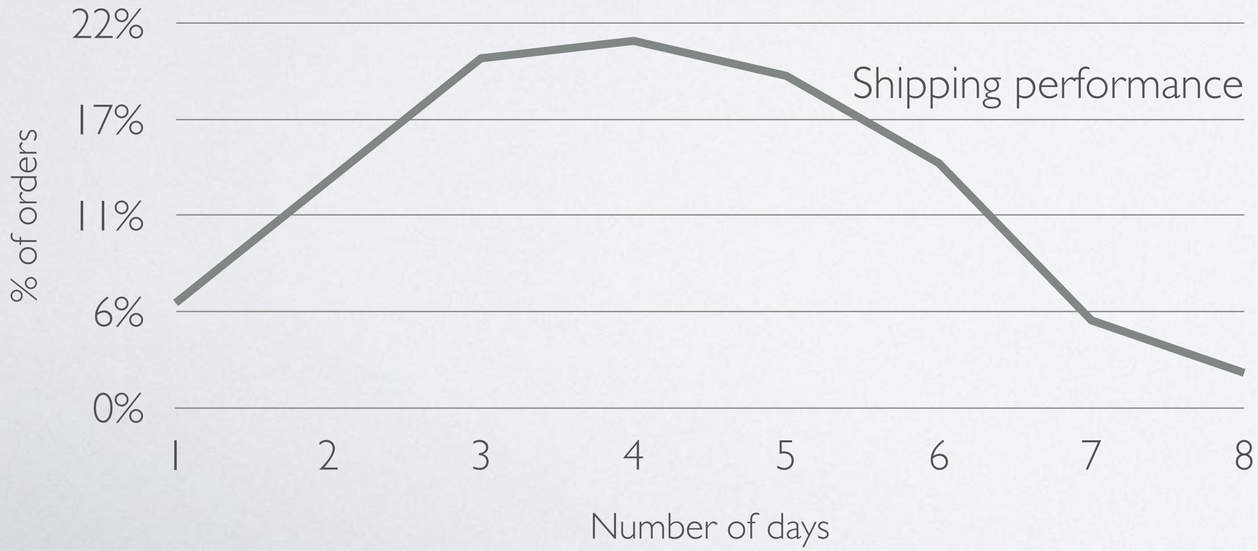
It focuses our attention to the shape by eliminating any visual component that would draw our eyes to the values of the individual ranges

DISTRIBUTION: FREQUENCY POLYGON

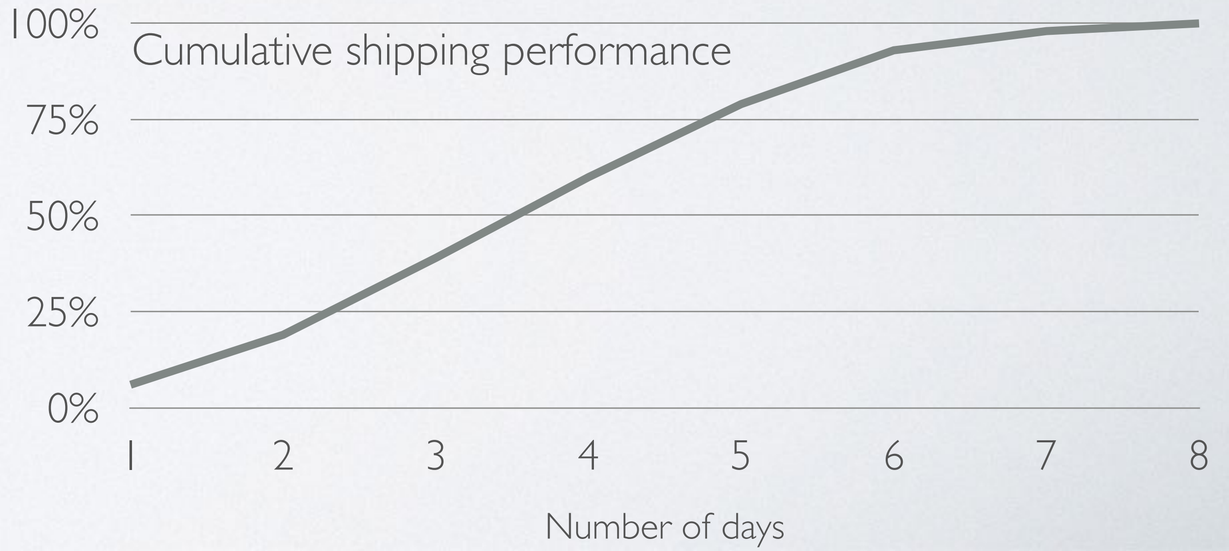
- How one or more sets of quantitative values are distributed across their full quantitative range, from lowest to highest

- Single distribution: how often something occurs, distributed across a series of consecutive quantitative ranges (a.k.a. intervals). It is called frequency distribution

Frequency polygon: if you want to emphasise the shape of the distribution



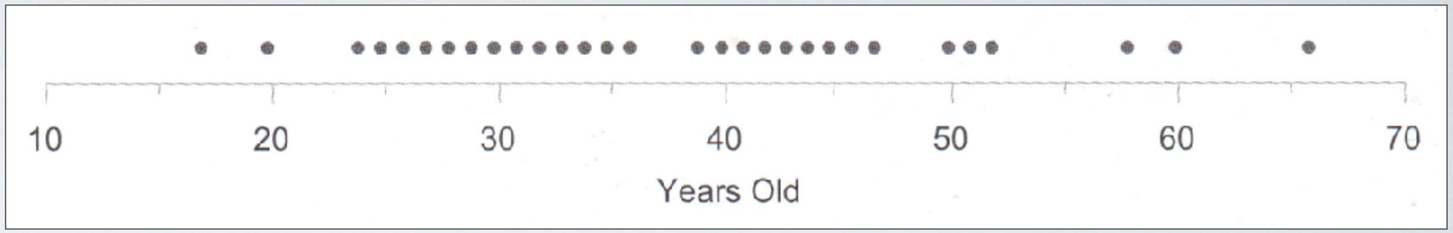
It works really well for cumulative distribution



DISTRIBUTION: STRIP PLOT

- Strip plot display each value in the dataset rather than aggregating the number or percentages of values into intervals
- They don't show the shape of the distribution very well, but they are useful when you have a small set of values and you want to show precisely where they are quantitatively

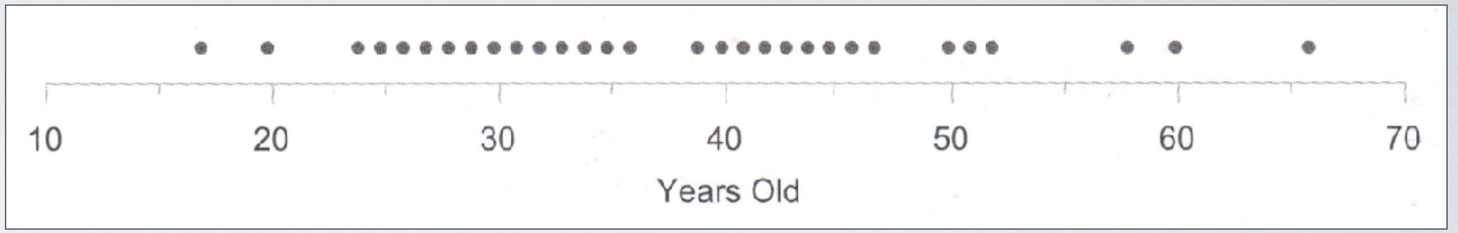
Employees in the Philadelphia office by age



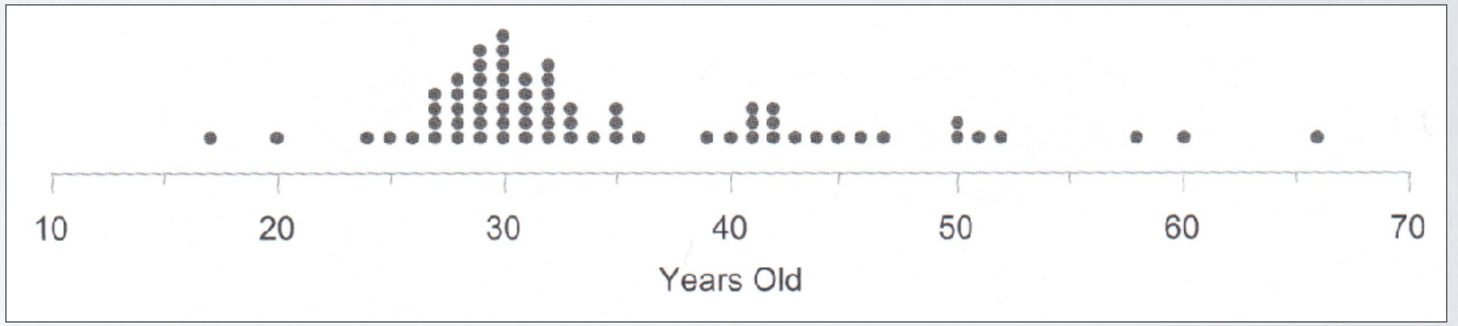
DISTRIBUTION: STRIP PLOT

- Strip plot display each value in the dataset rather than aggregating the number or percentages of values into intervals
- They don't show the shape of the distribution very well, but they are useful when you have a small set of values and you want to show precisely where they are quantitatively

Employees in the Philadelphia office by age



Employees in the New York office by age

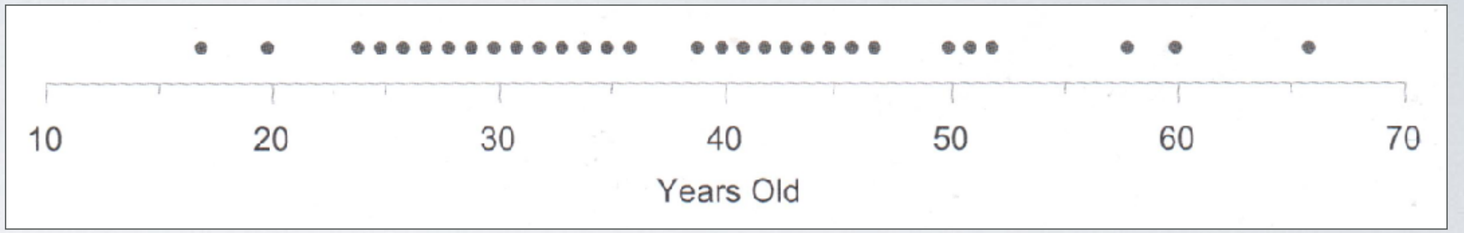


- When multiple points have the same value, you won't see them. We can solve this problem by jittering or by making the points transparent

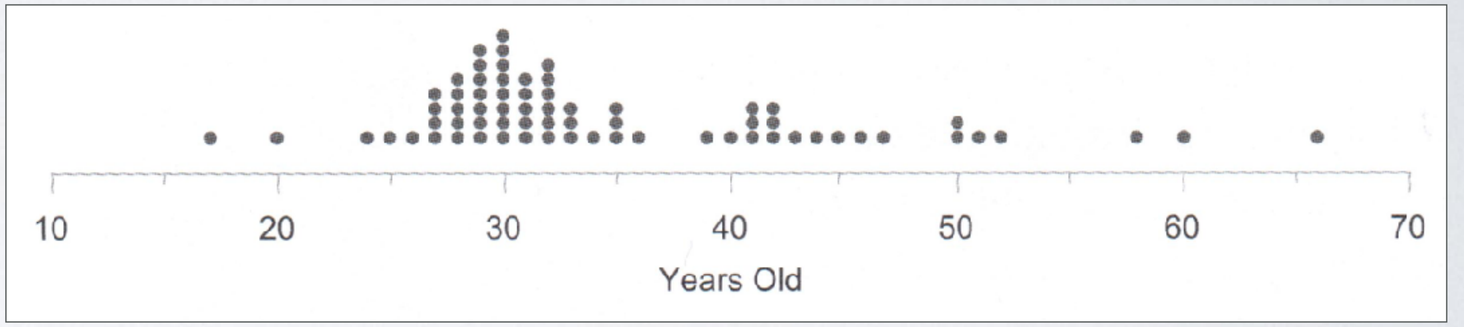
DISTRIBUTION: STRIP PLOT

- Strip plot display each value in the dataset rather than aggregating the number or percentages of values into intervals
- They don't show the shape of the distribution very well, but they are useful when you have a small set of values and you want to show precisely where they are quantitatively

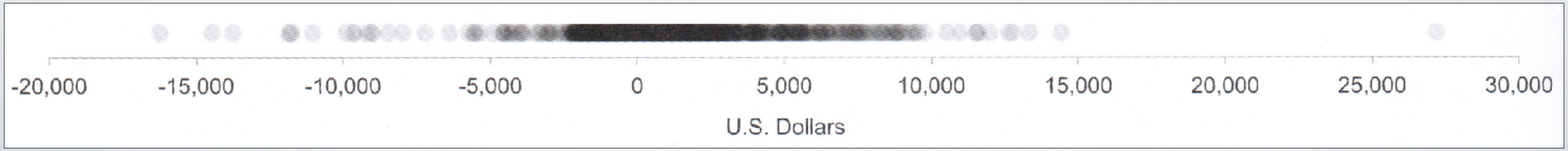
Employees in the Philadelphia office by age



Employees in the New York office by age

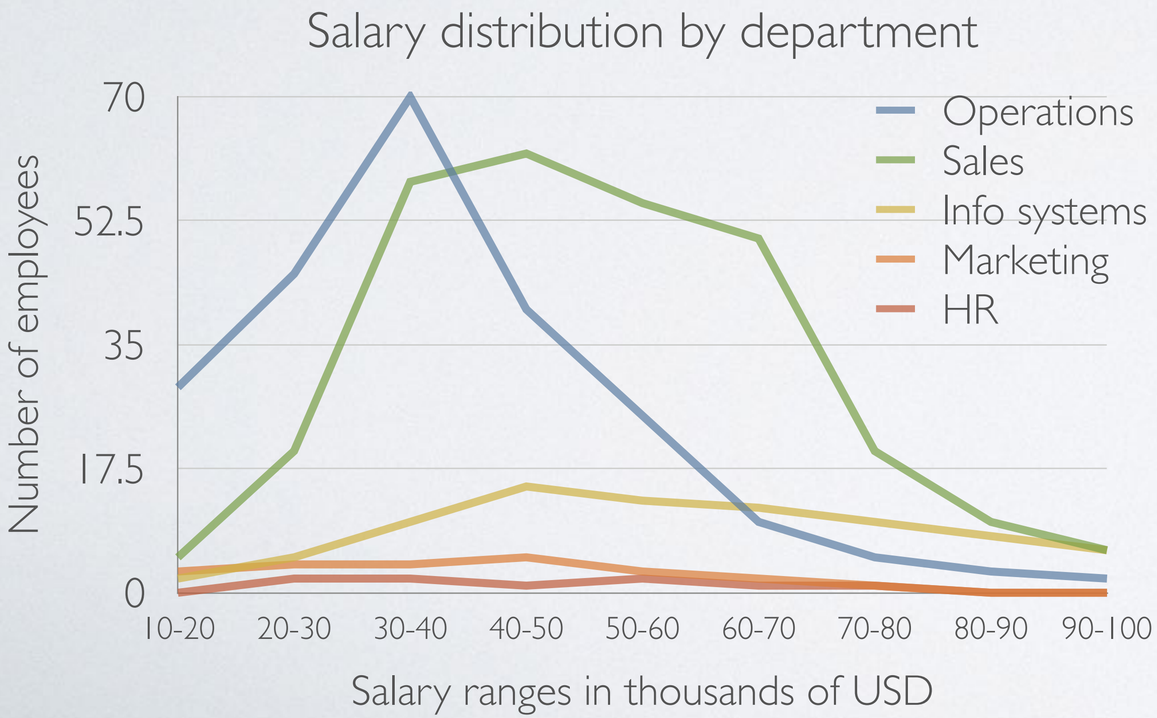


- When multiple points have the same value, you won't see them. We can solve this problem by jittering or by making the points transparent



MULTIPLE DISTRIBUTIONS

- Multiple frequency polygons can be displayed in a single graph, so that they can be compared
- Color is used to map the categorical information
- Better not to display more than 5 polygons (max 5 colors), otherwise the graph would be cluttered and hard to read

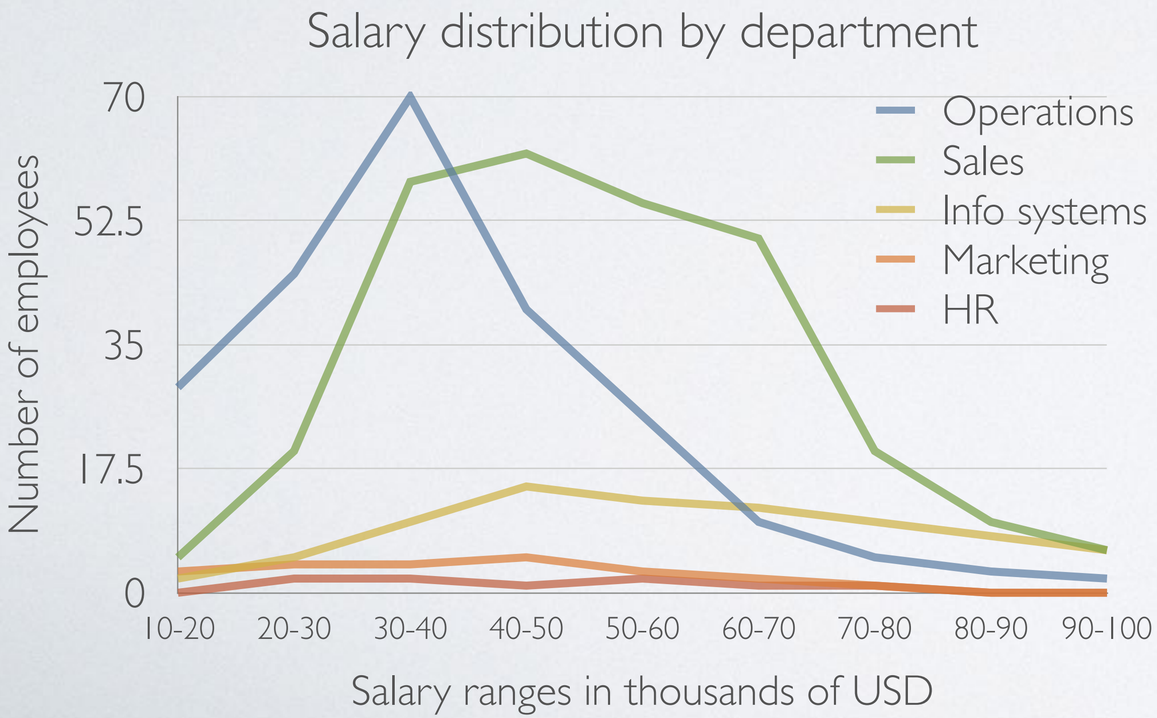


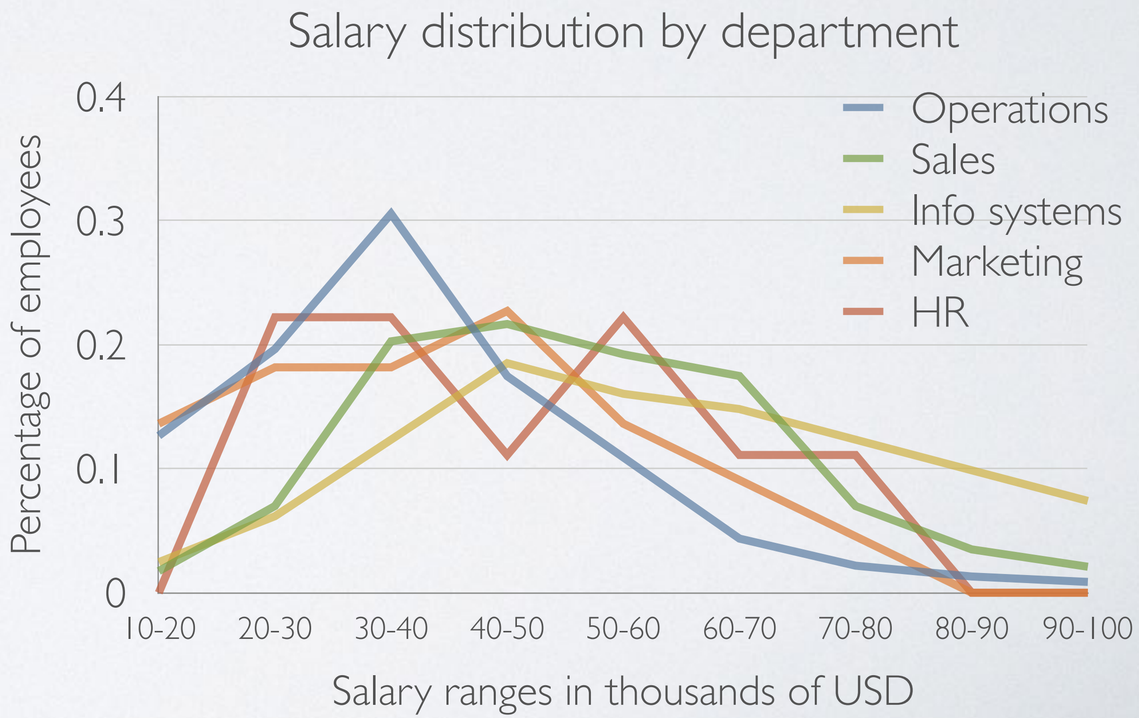
Because the number of employee in each department varies dramatically, some lines are relatively flat, which makes it hard to see their shape.

If we instead express the distribution in terms of percentages (with the values along each line summing to 100%), the shapes would be easier to see and compare

MULTIPLE DISTRIBUTIONS

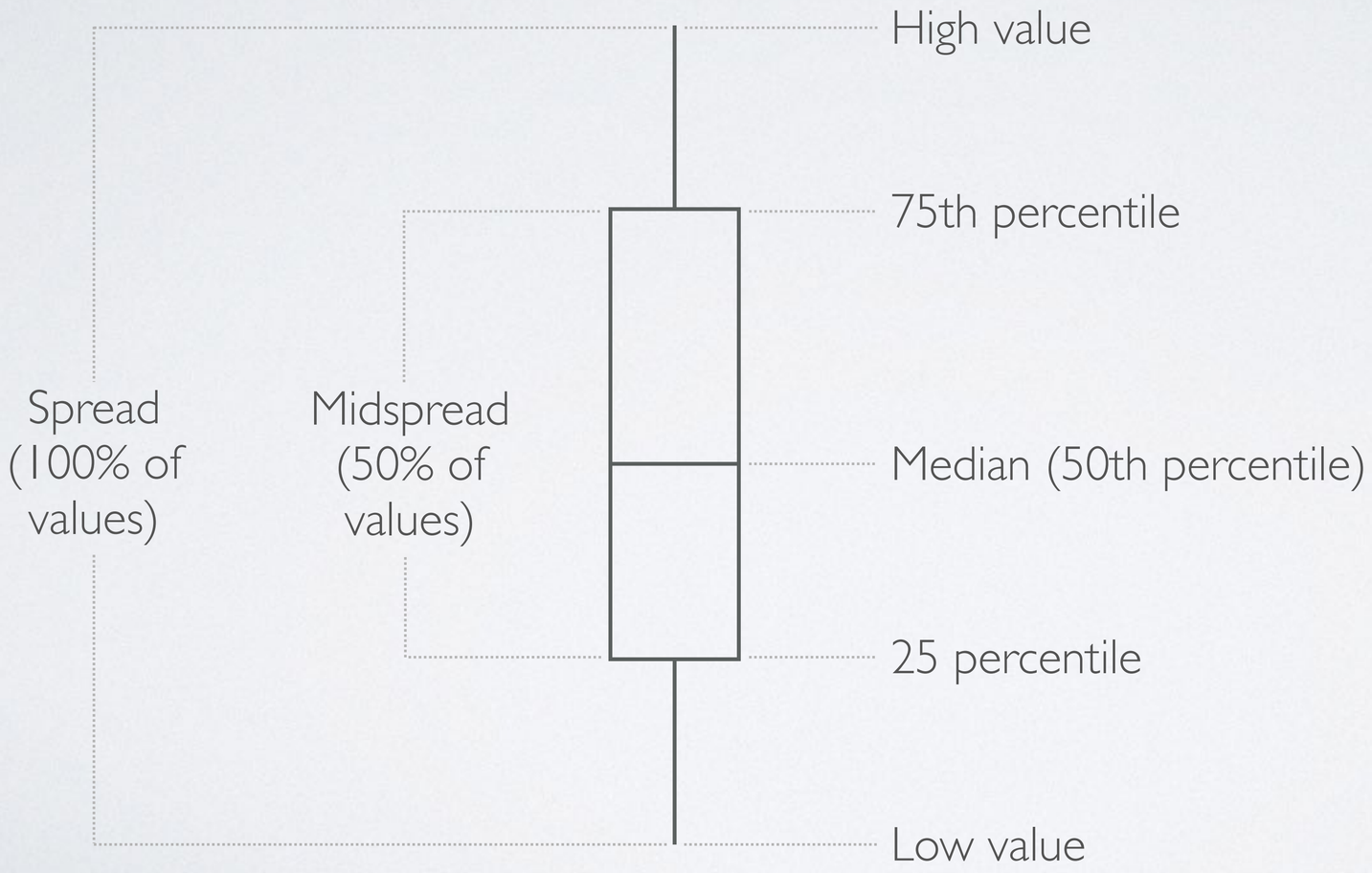
- Multiple frequency polygons can be displayed in a single graph, so that they can be compared
- Color is used to map the categorical information
- Better not to display more than 5 polygons (max 5 colors), otherwise the graph would be cluttered and hard to read



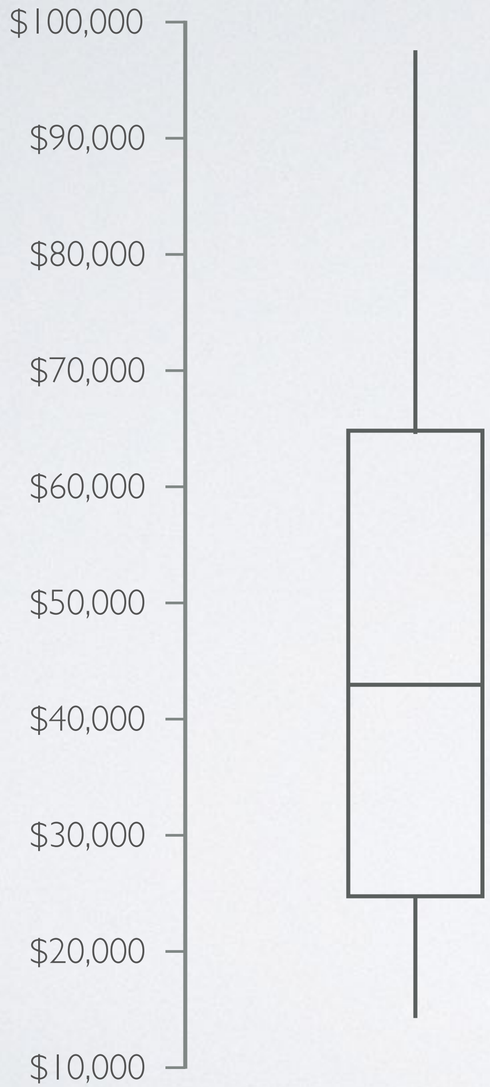


MORE THAN 5 DISTRIBUTIONS: BOX-AND-WHISKER PLOT

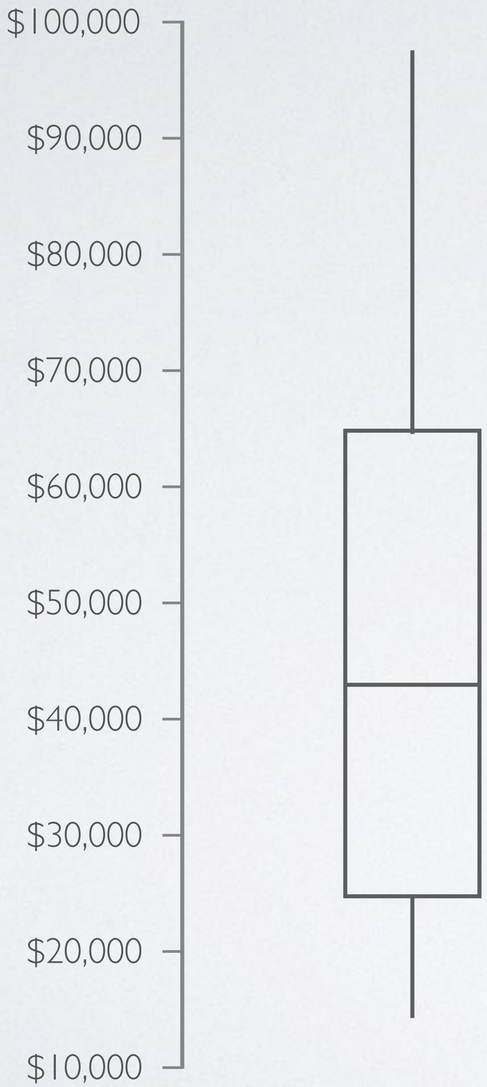
A simple combination of bar and data points communicate a lot about the distribution



BOX-AND-WHISKER PLOT EXAMPLE

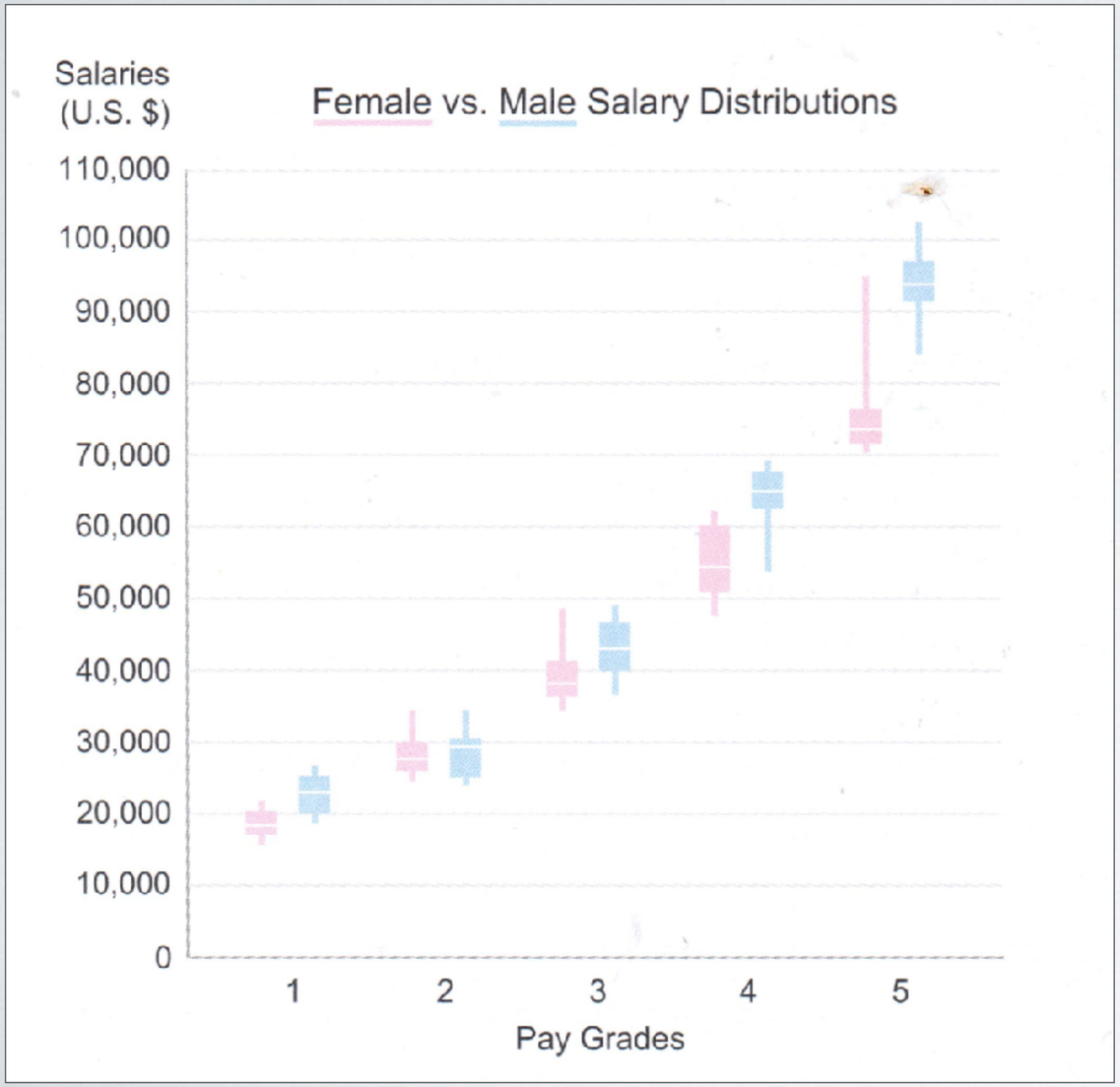


BOX-AND-WHISKER PLOT EXAMPLE

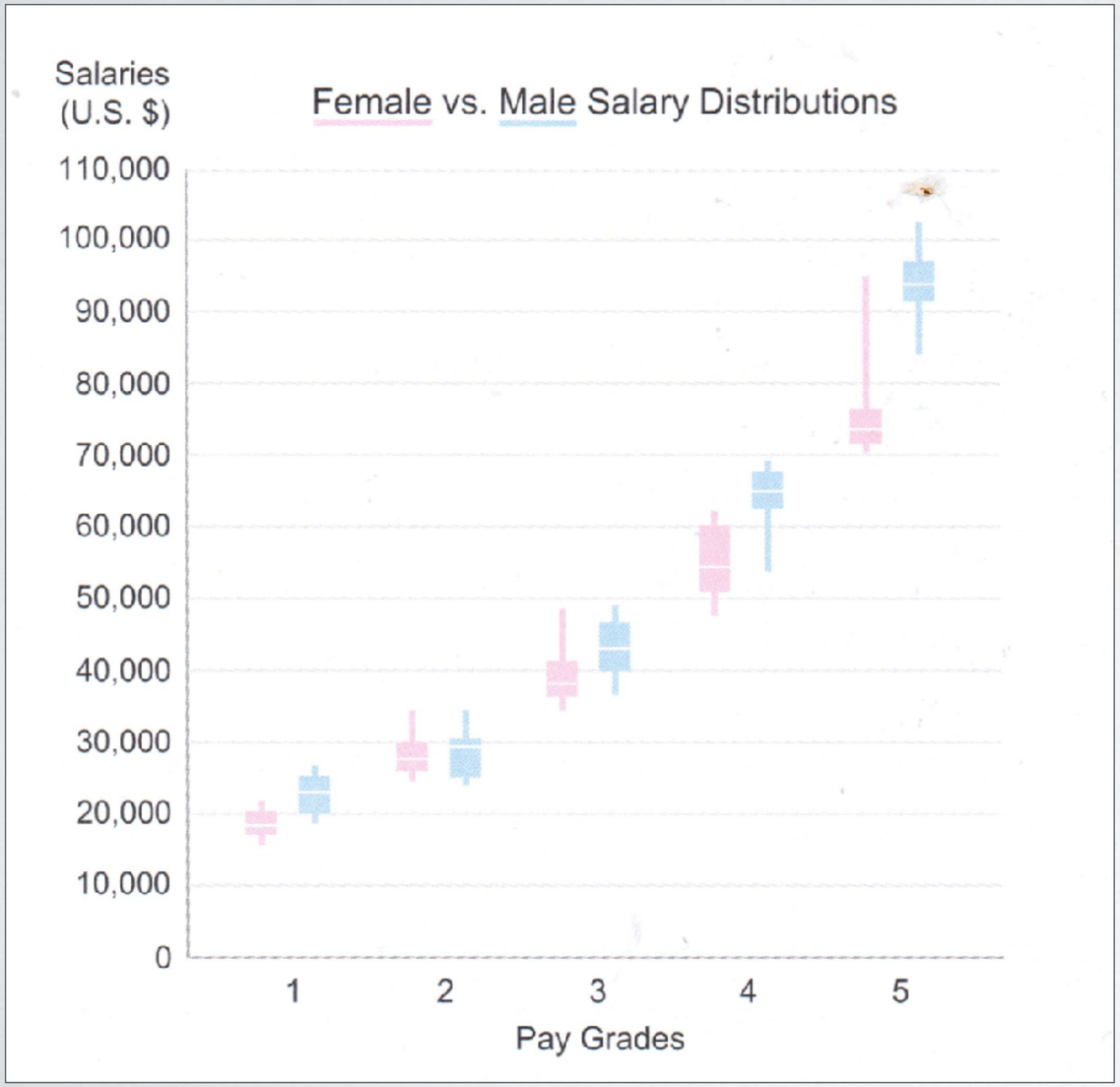


- Full range of salaries is large, from ~15,000 to ~97,000
- Salaries are not evenly spread across the entire range
- More people earn salaries toward the lower end
- Half of the people earn between ~25,000 and ~65,000
- The 25% of people who earn the lowest salaries are grouped closely, in a ~10,000 range (between ~15,000 and ~25,000)
- On the contrary, the 25% top salaries have a large spread of ~32,000 (between ~65,000 and ~97,000)

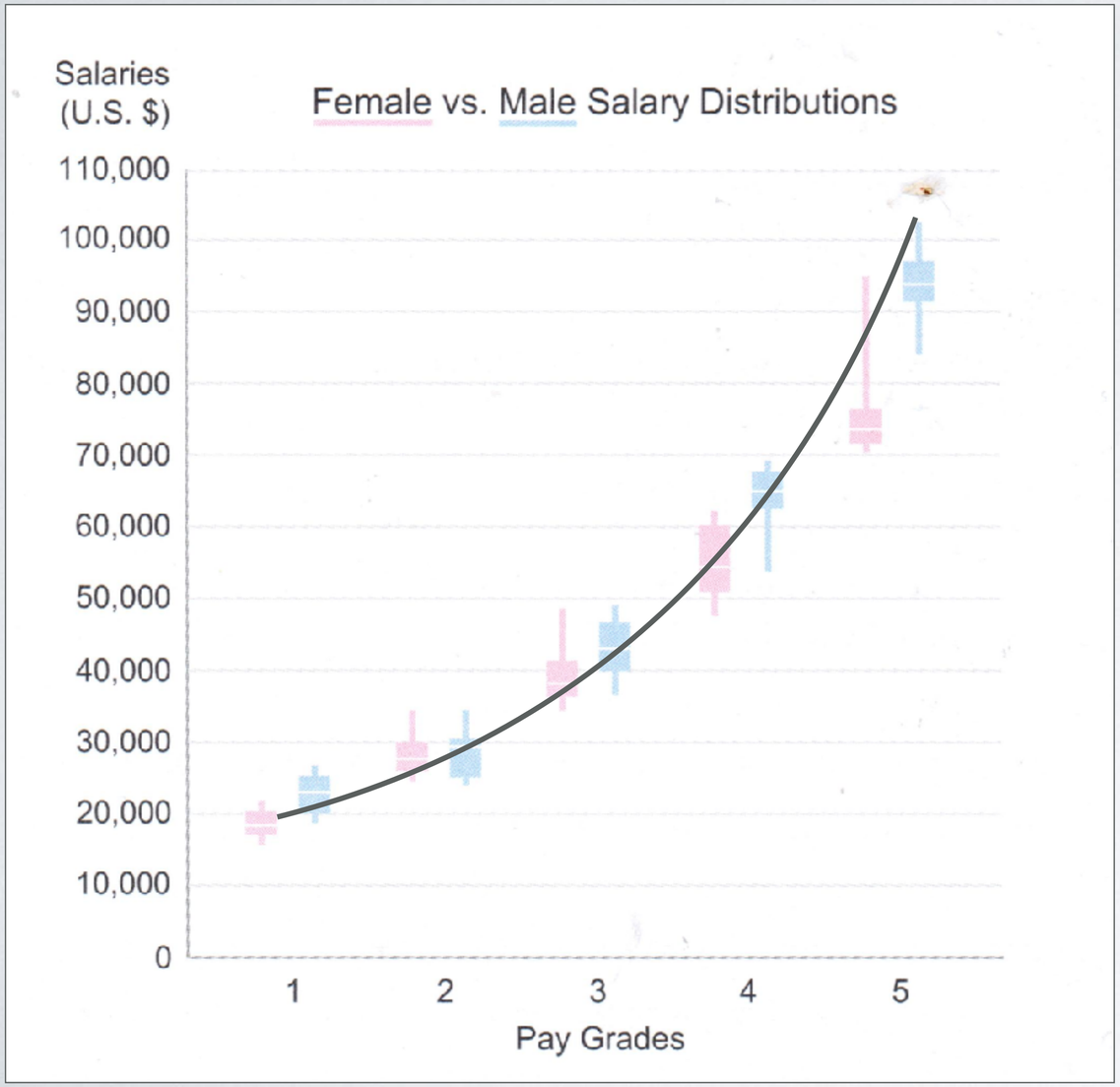
BOX PLOT EXAMPLES



BOX PLOT EXAMPLES

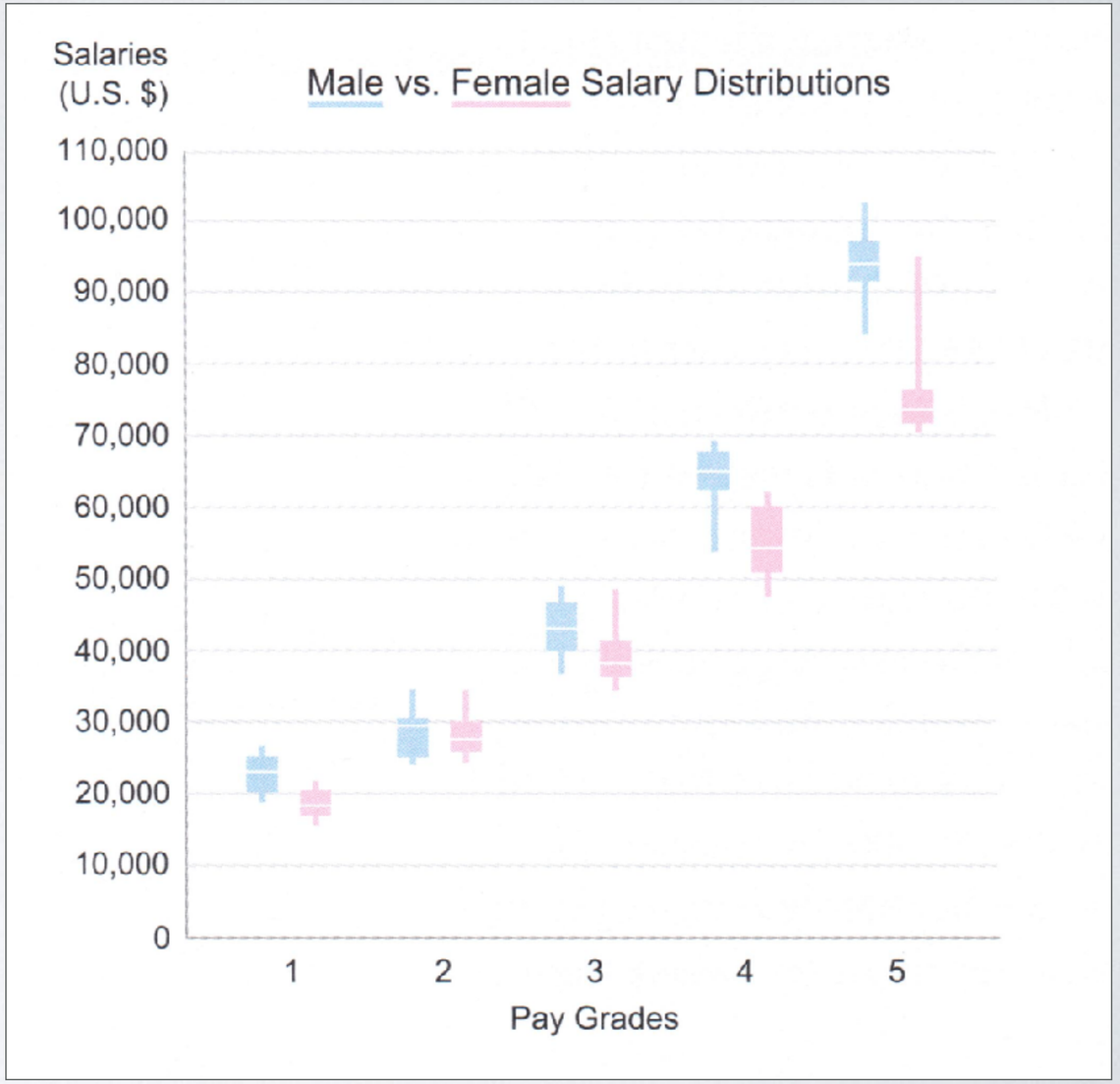


BOX PLOT EXAMPLES



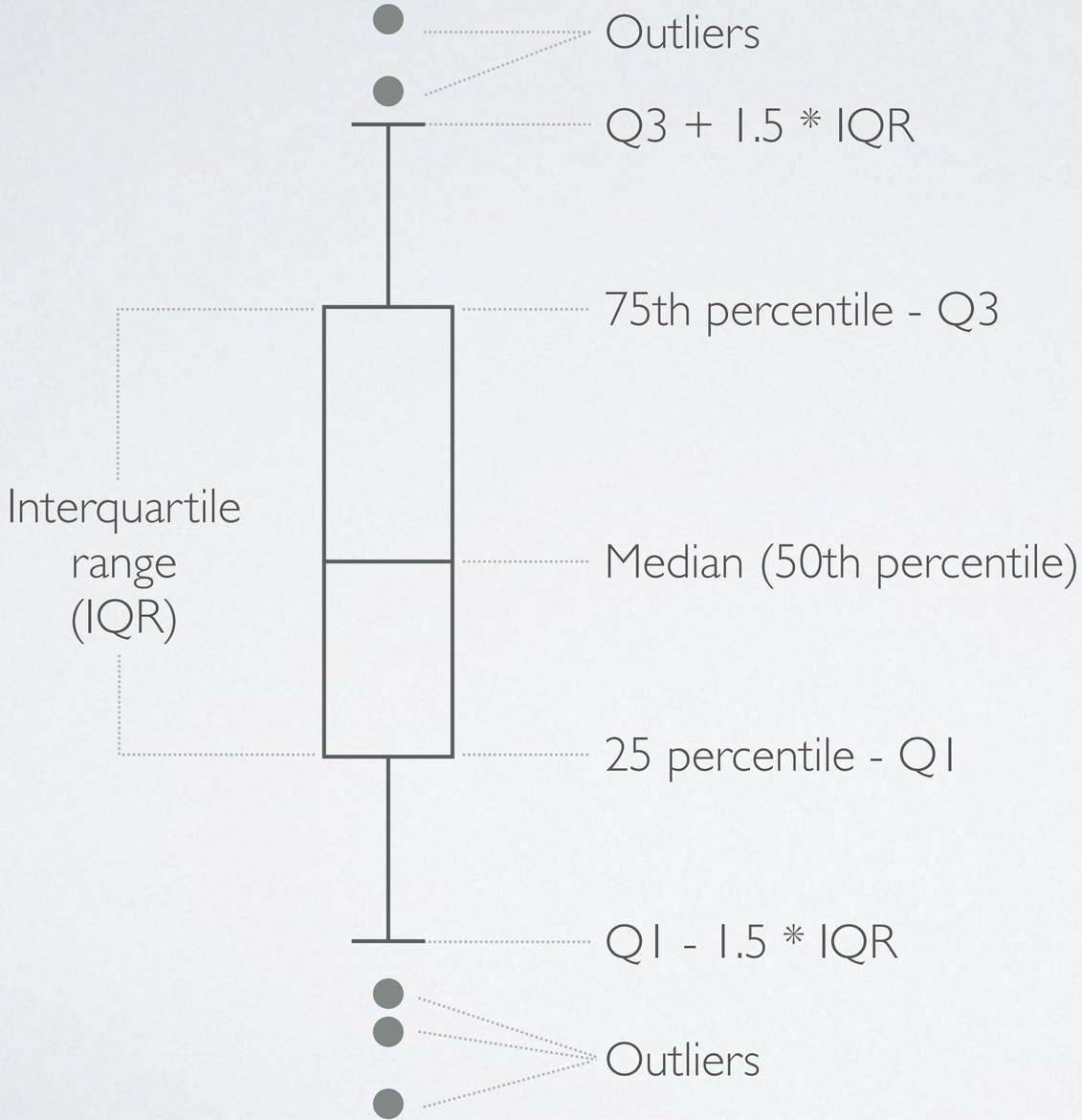
BOX PLOT EXAMPLES

- On average women are paid less than men in all salary grades
- The disparity in salaries between men and women becomes increasingly greater as one's salary increases
- Salaries vary the most for women in the higher salary grades

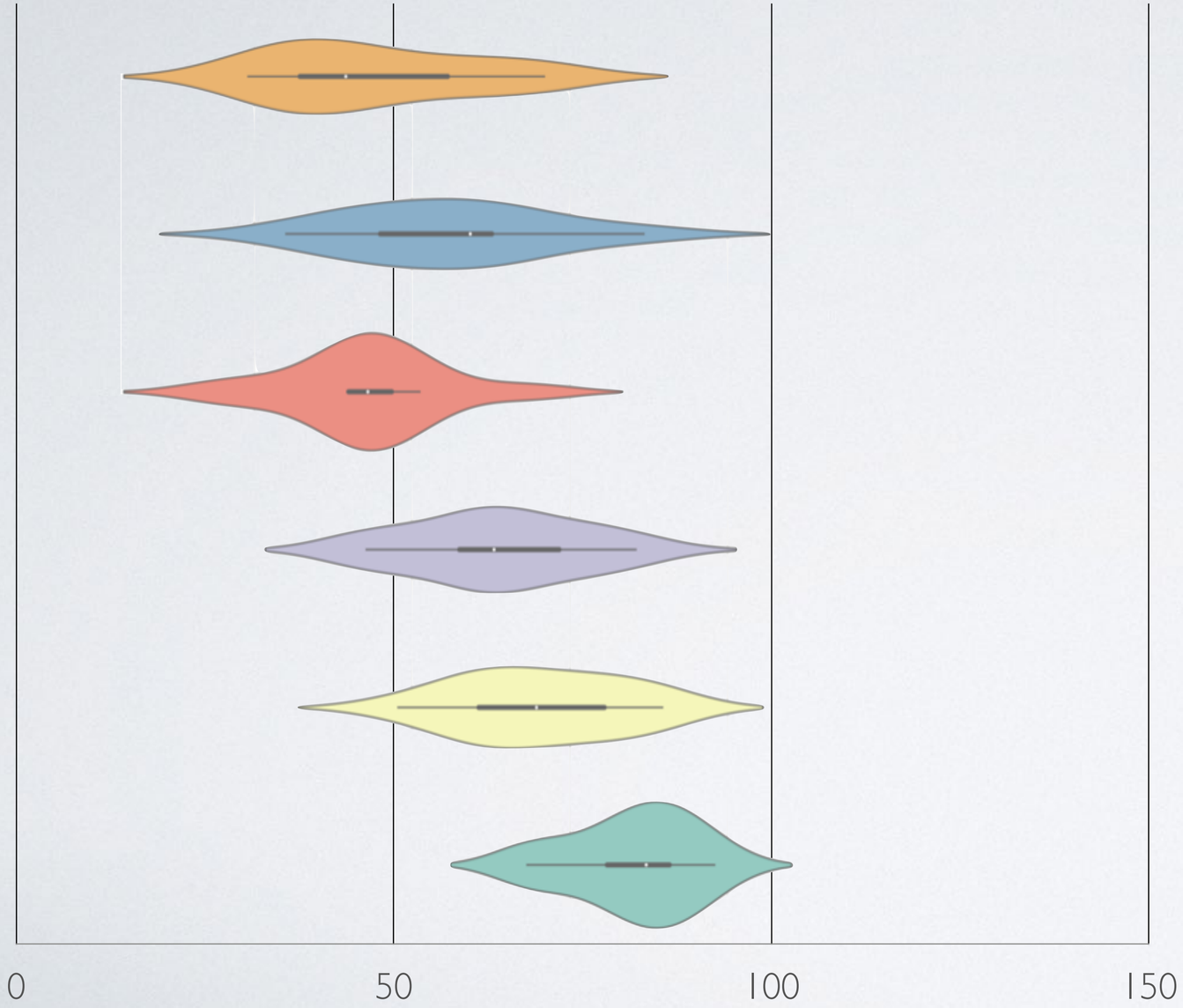


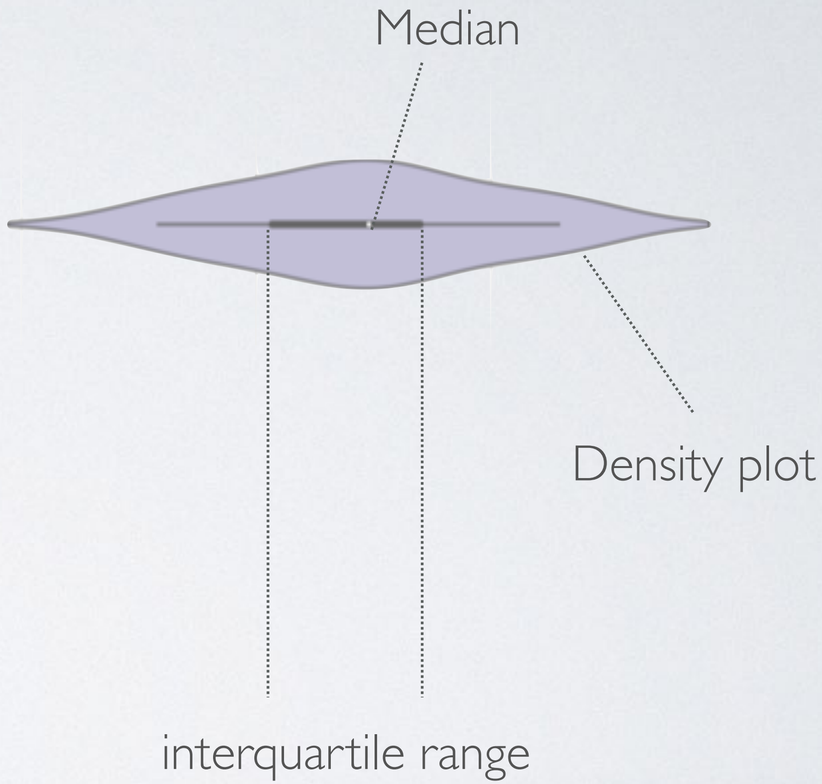
BOX PLOT EXAMPLES

ANOTHER BOXPLOT (FOR EXAMPLE IN SEABORN)



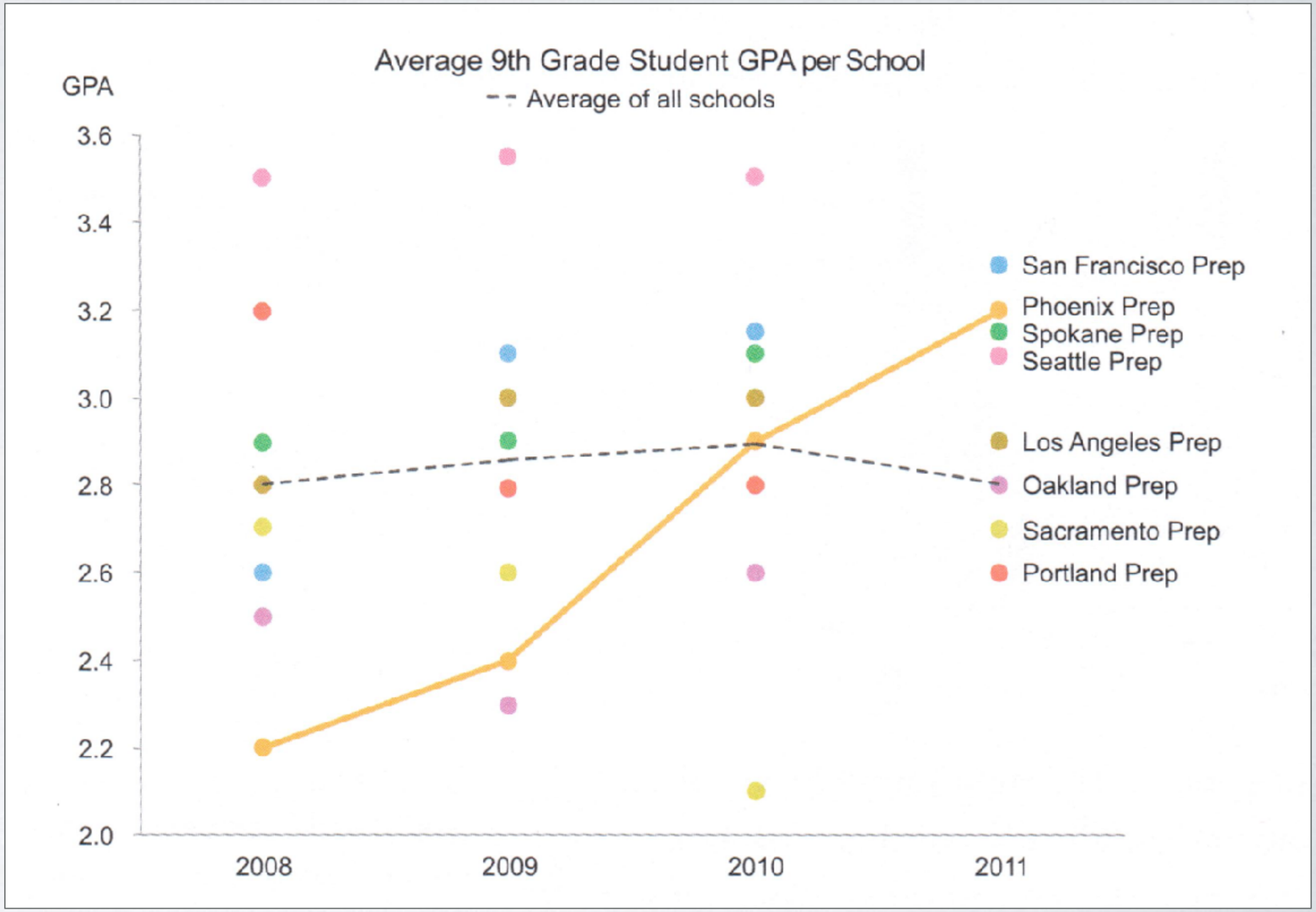
VIOLIN PLOT





STRIP PLOT FOR MULTIPLE DISTRIBUTIONS

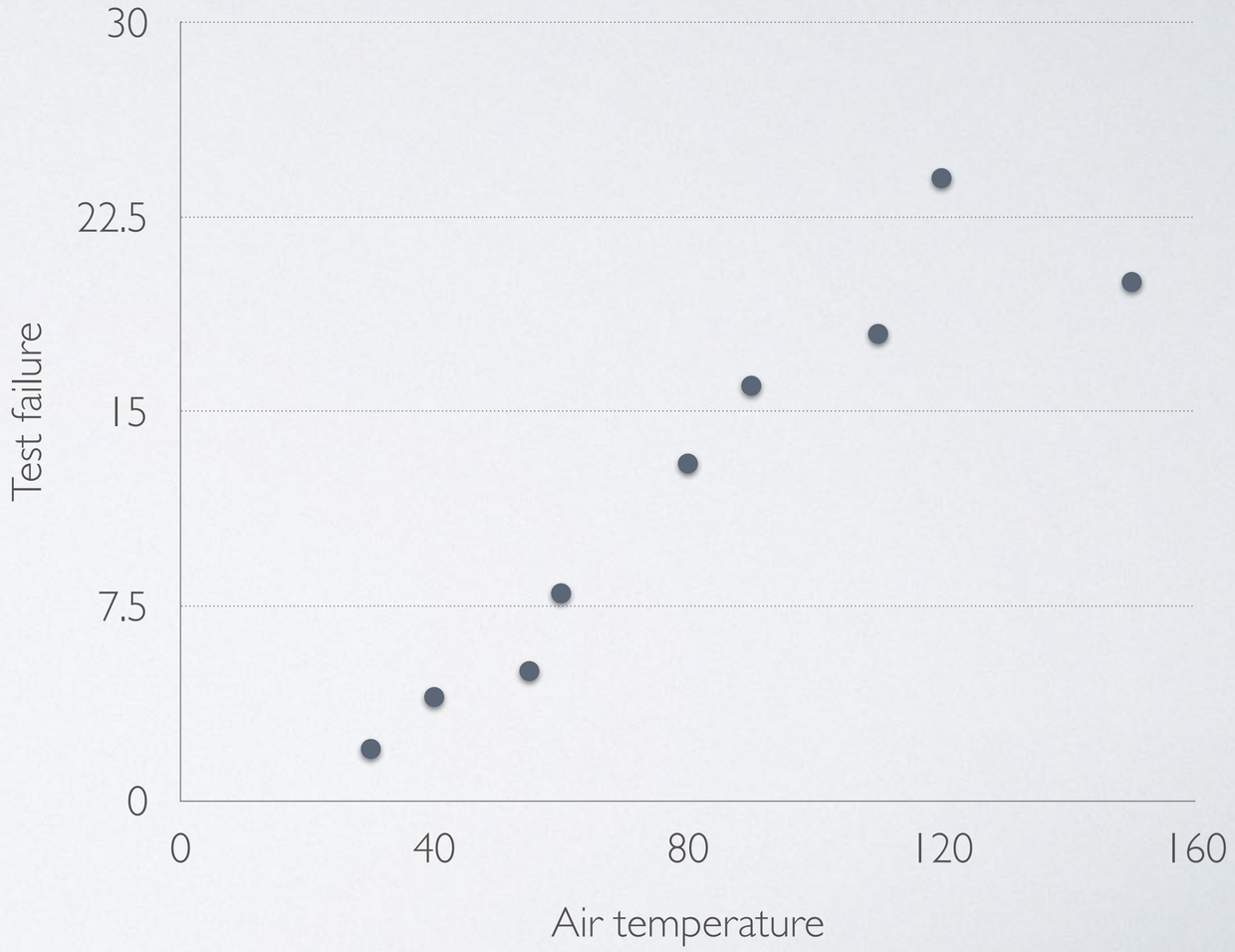
- Multiple distributions are displayed as several rows or columns
- They are useful when there are relatively few values in the dataset and you have a reason to display each value individually



CORRELATION

- Whether two paired sets of quantitative values vary in relation to each other, in which direction, and to what degree
- A common numerical measure of correlation is the linear correlation coefficient
- Since the relationship we want to show is between 2 sets of quantitative values, both x and y axes must provide quantitative scale
- Points work perfectly to encode this relationship

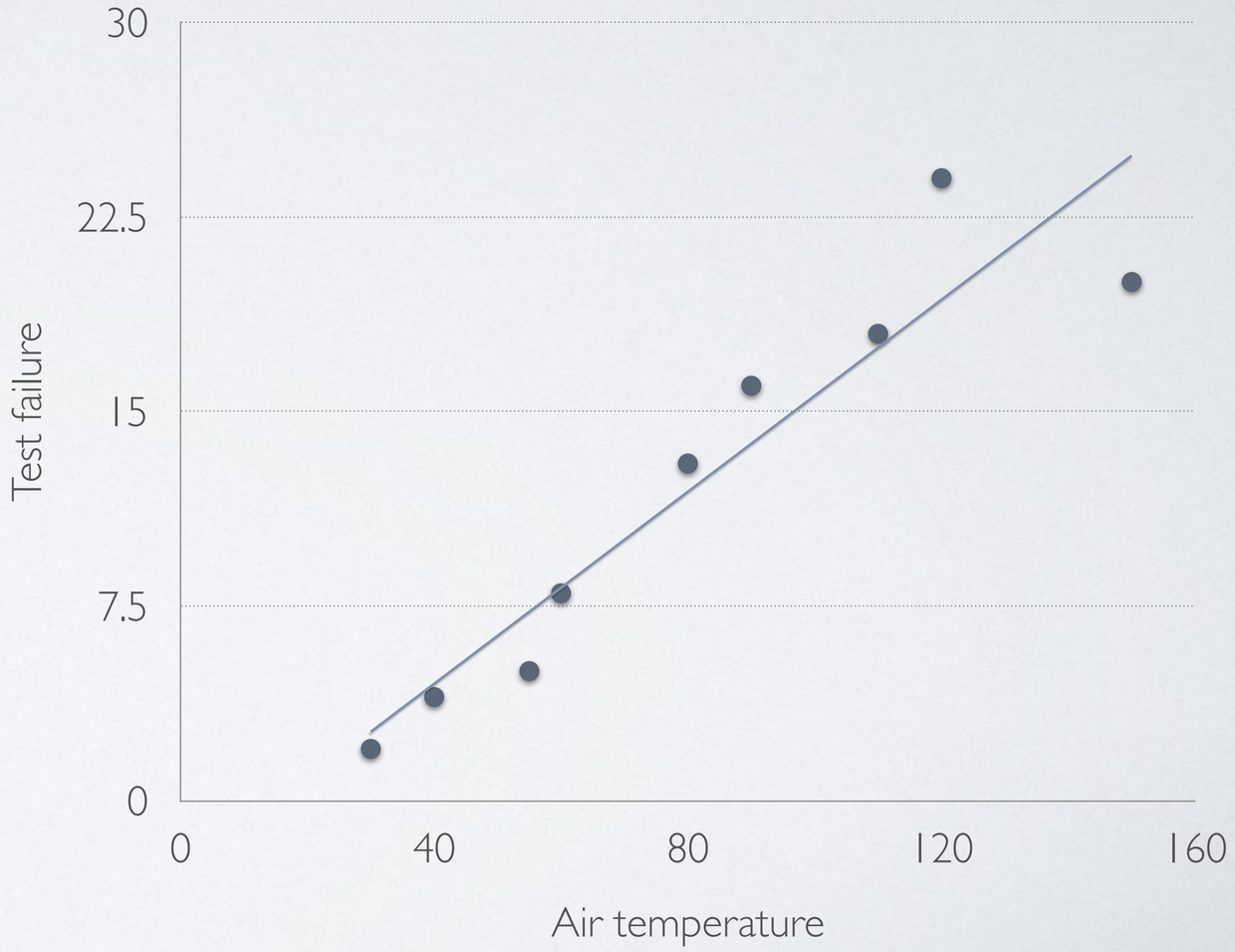
Scatter plot: how test failure correlates with air temperature



CORRELATION

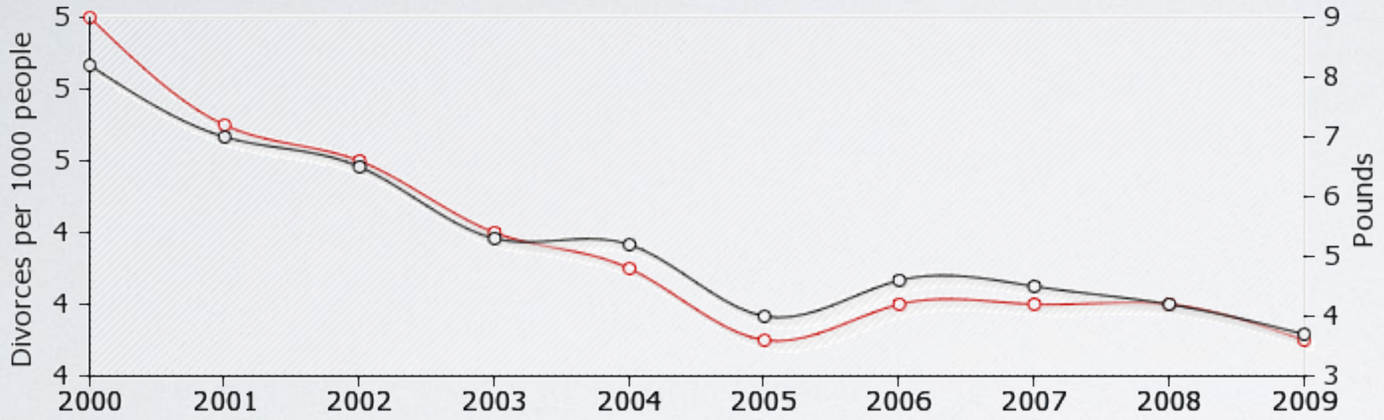
- Whether two paired sets of quantitative values vary in relation to each other, in which direction, and to what degree
- A common numerical measure of correlation is the linear correlation coefficient
- Since the relationship we want to show is between 2 sets of quantitative values, both x and y axes must provide quantitative scale
- Points work perfectly to encode this relationship
- To make a potential correlation more visible we can add a trend line

Scatter plot: how test failure correlates with air temperature



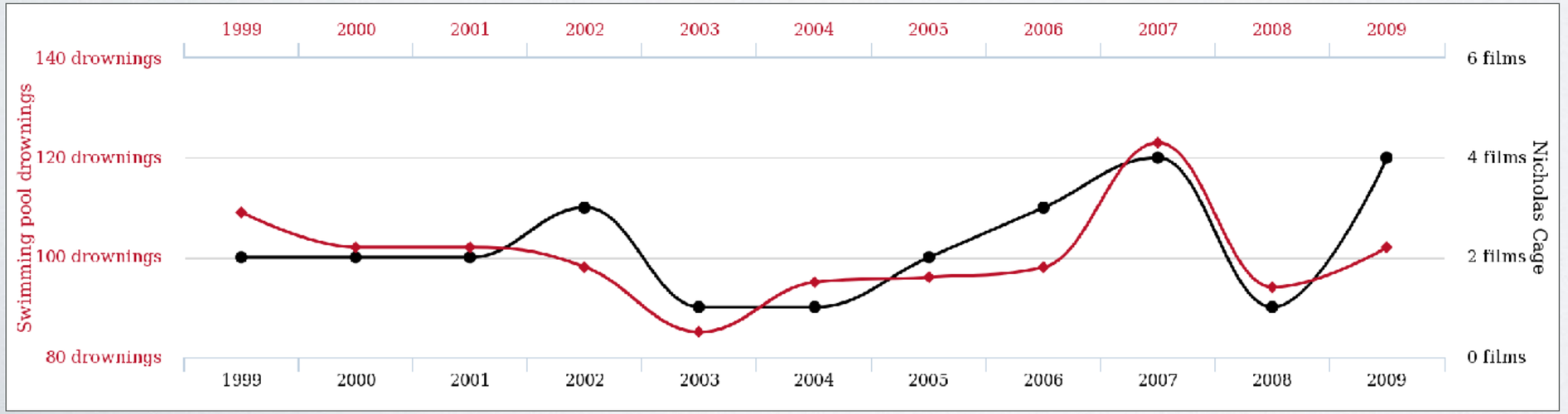
CORRELATION ≠ CAUSATION

Be careful with inferring theories out of correlations



Divorce rate in Maine

Per capita consumption of margarine (US)



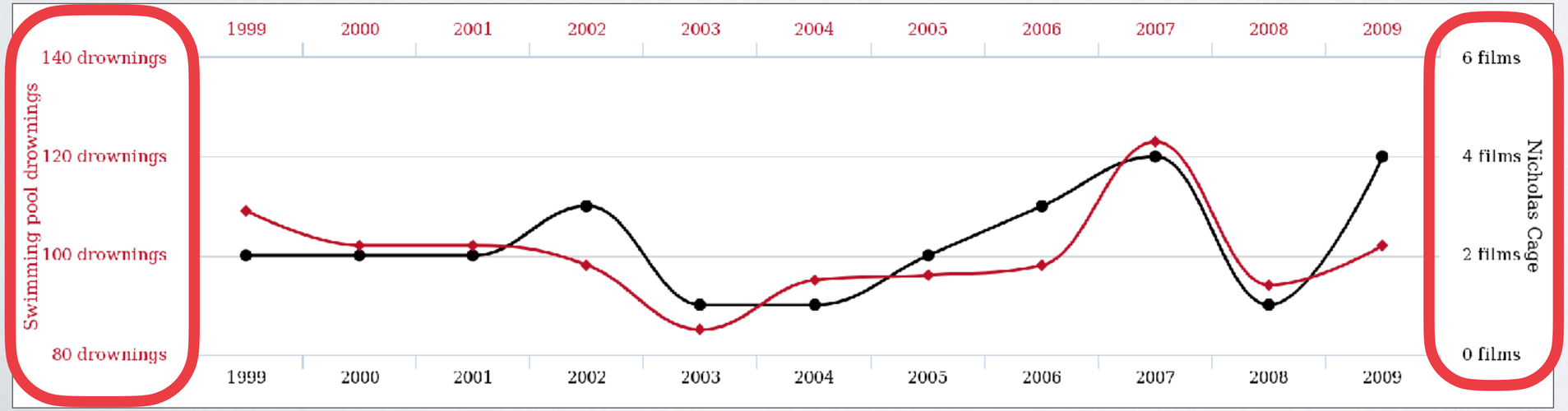
Number of people who drowned by falling into a pool

Films Nicolas Cage appeared in

CORRELATION ≠ CAUSATION

Be careful with inferring theories out of correlations

- Divorce rate in Maine
- Per capita consumption of margarine (US)



- Number of people who drowned by falling into a pool
- Films Nicolas Cage appeared in

CORRELATION ≠ CAUS

Be careful with inferring theor

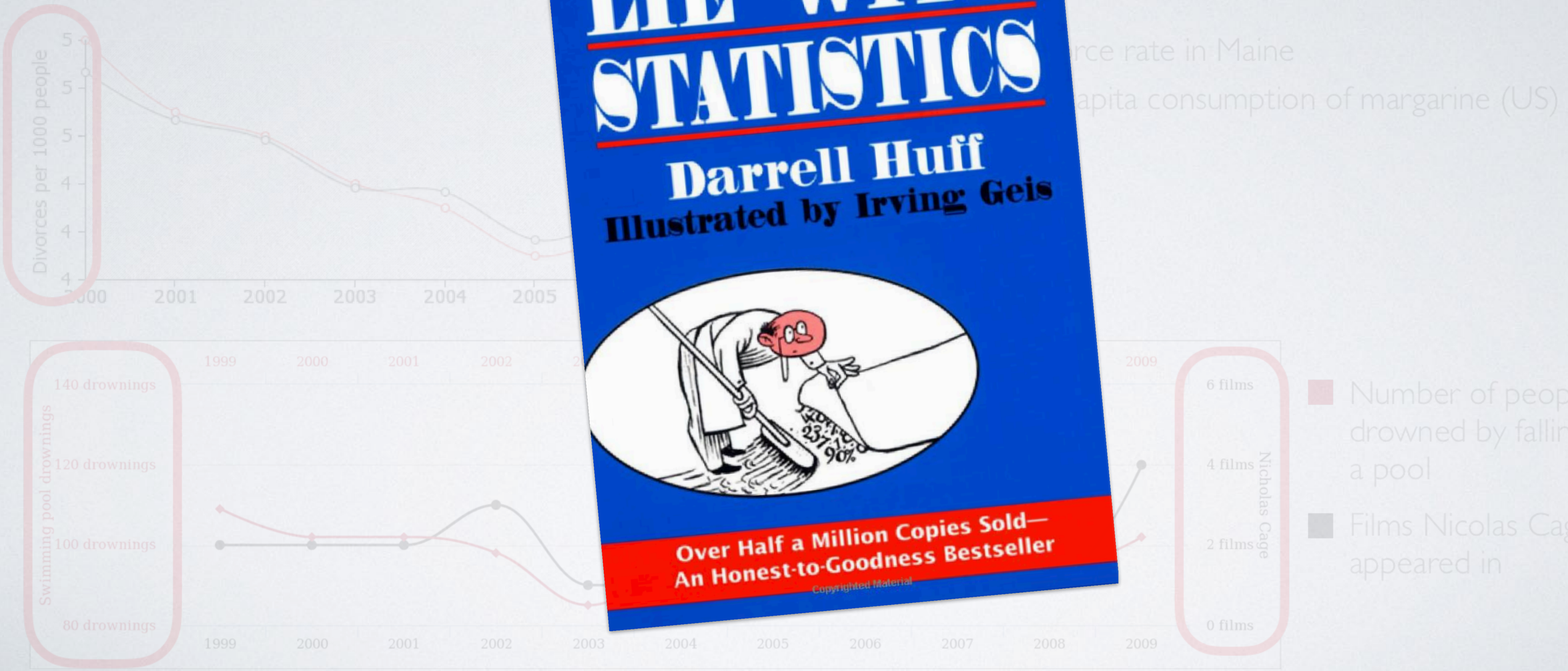
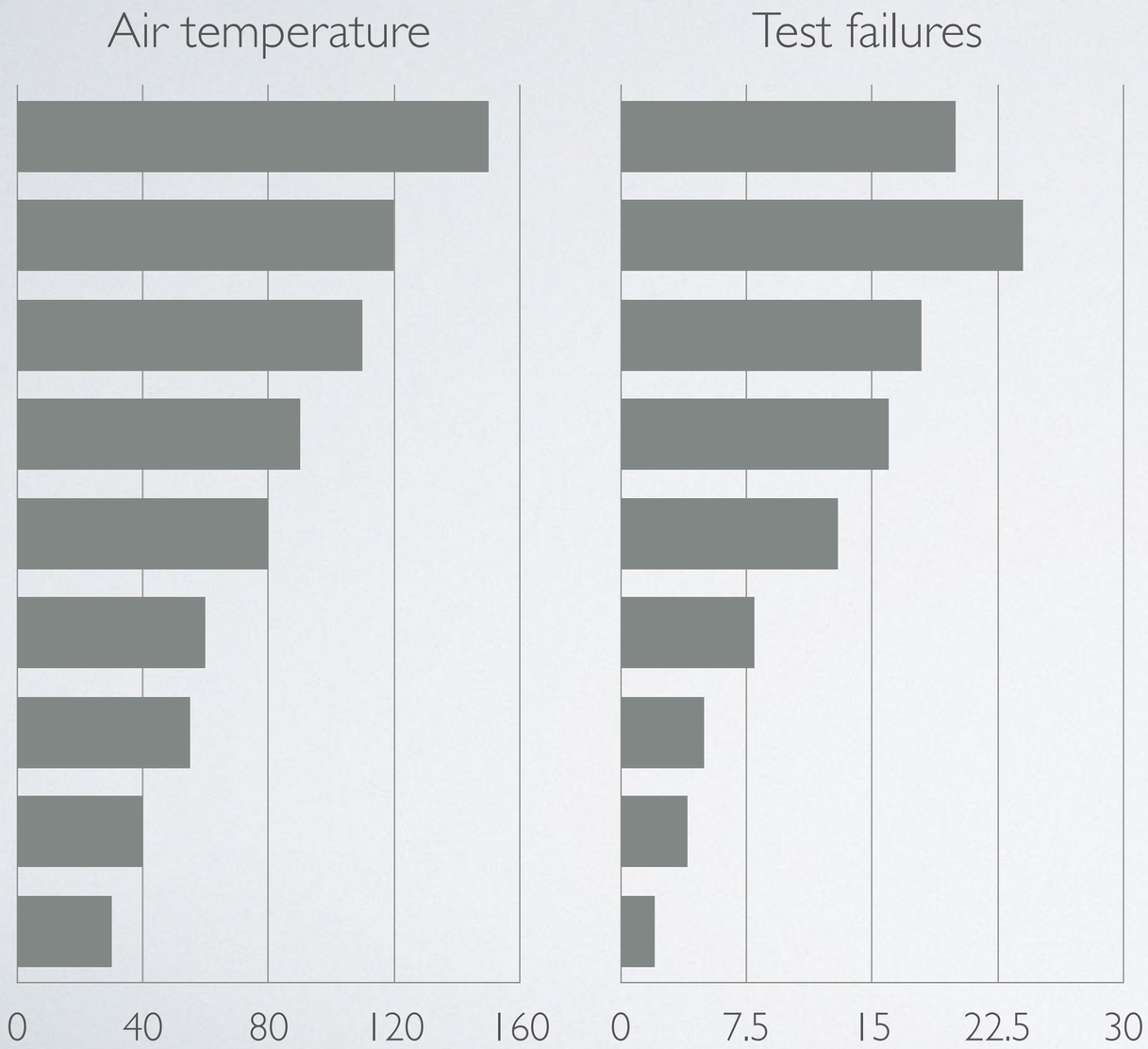


TABLE LENS

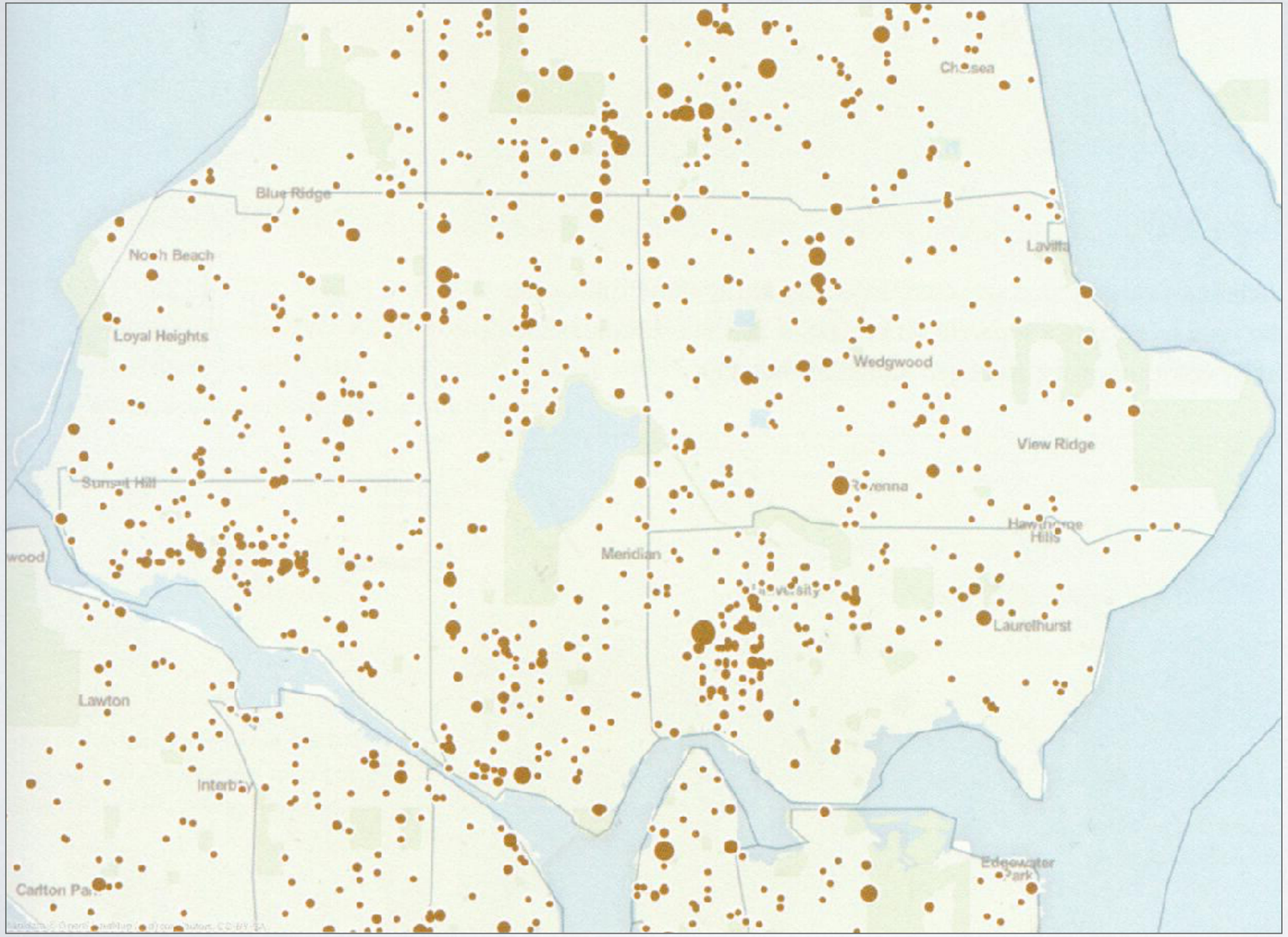


- Another way to display correlation
- Table lens vs scatter plot
  - Scatter plot overall superior
  - With table lens you can look at multiple variables
  - With tables lens you can look at a smaller set of values

GEOSPATIAL

- Geospatial displays feature the geographical location of values, positioning values on a map
- Encoding of quantitative values on geospatial displays:
    - 2D position can't be used because the position map already the geographical position
    - Objects length can't be used because object can't be aligned to share the same baseline
    - We have to use:
        - Points—typically circles—of varying size (to place values in precise locations)
        - Points or areas of varying color intensity (size is normally better for points, especially when small)
        - Color intensities applied to geographical regions (choropleth maps)
        - Lines of varying thickness or color intensity

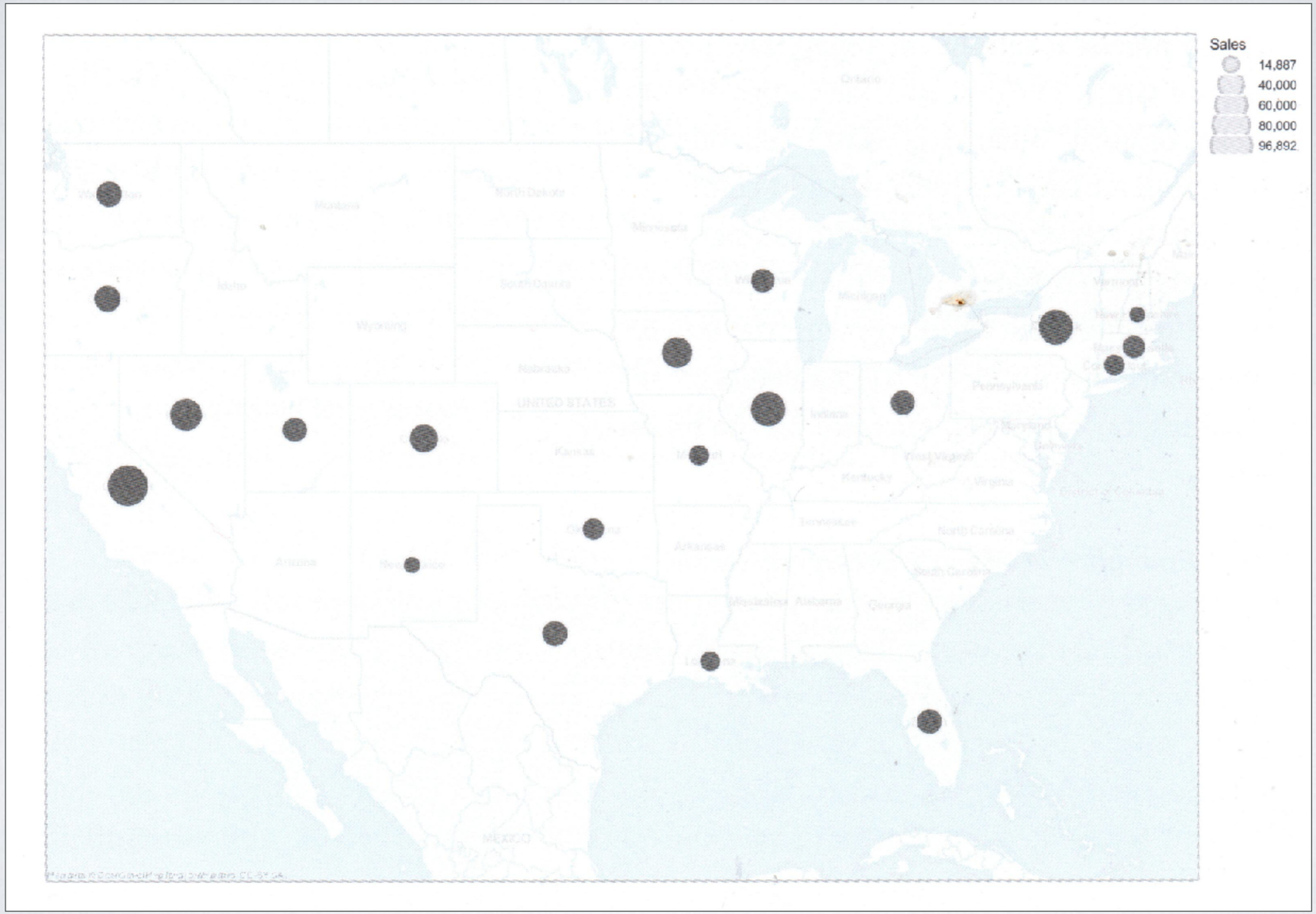
EXAMPLE: CIRCLE SIZE AND PRECISE LOCATION



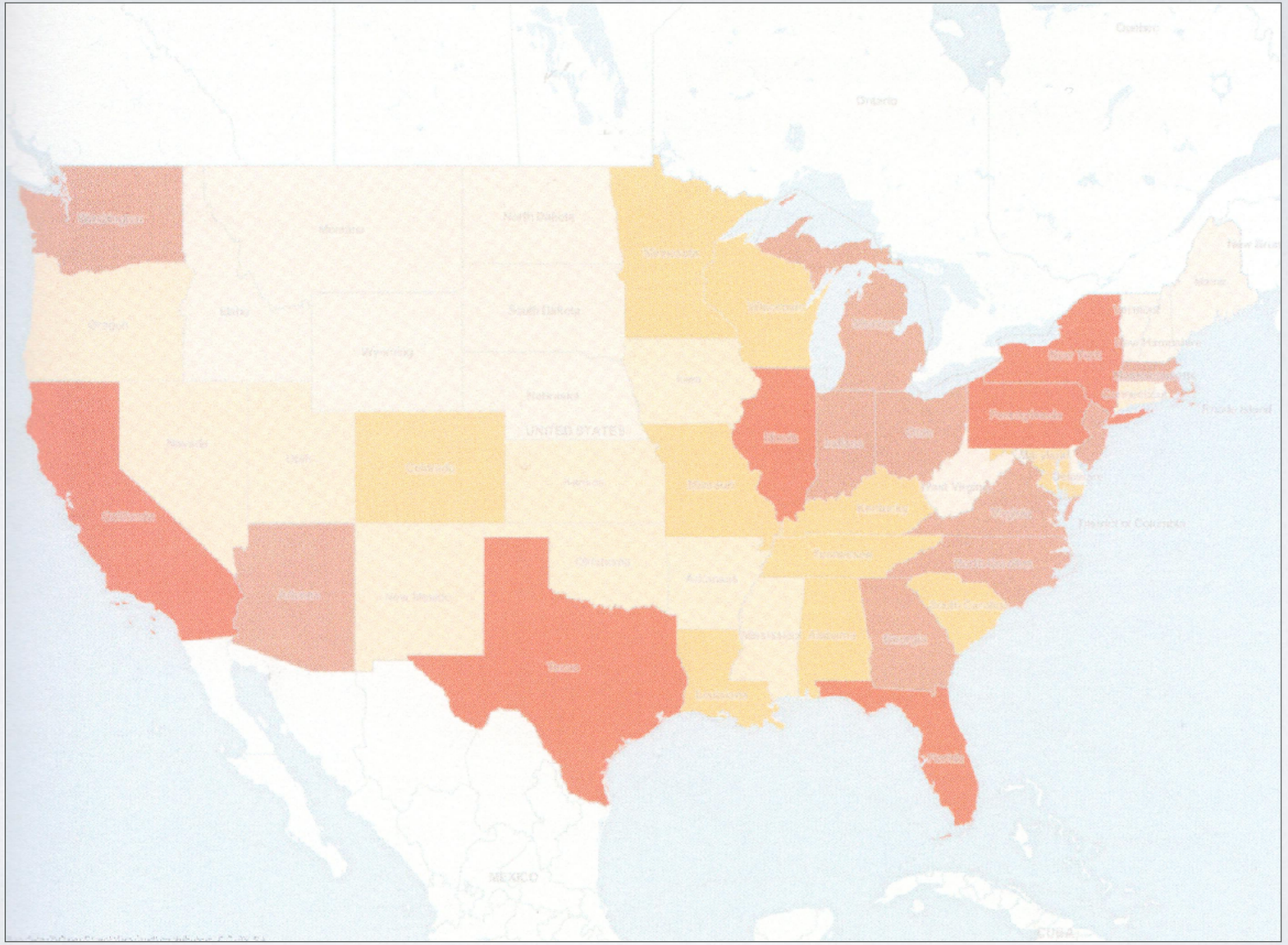
911 calls in Seattle

- 1
- 5
- 10
- 20

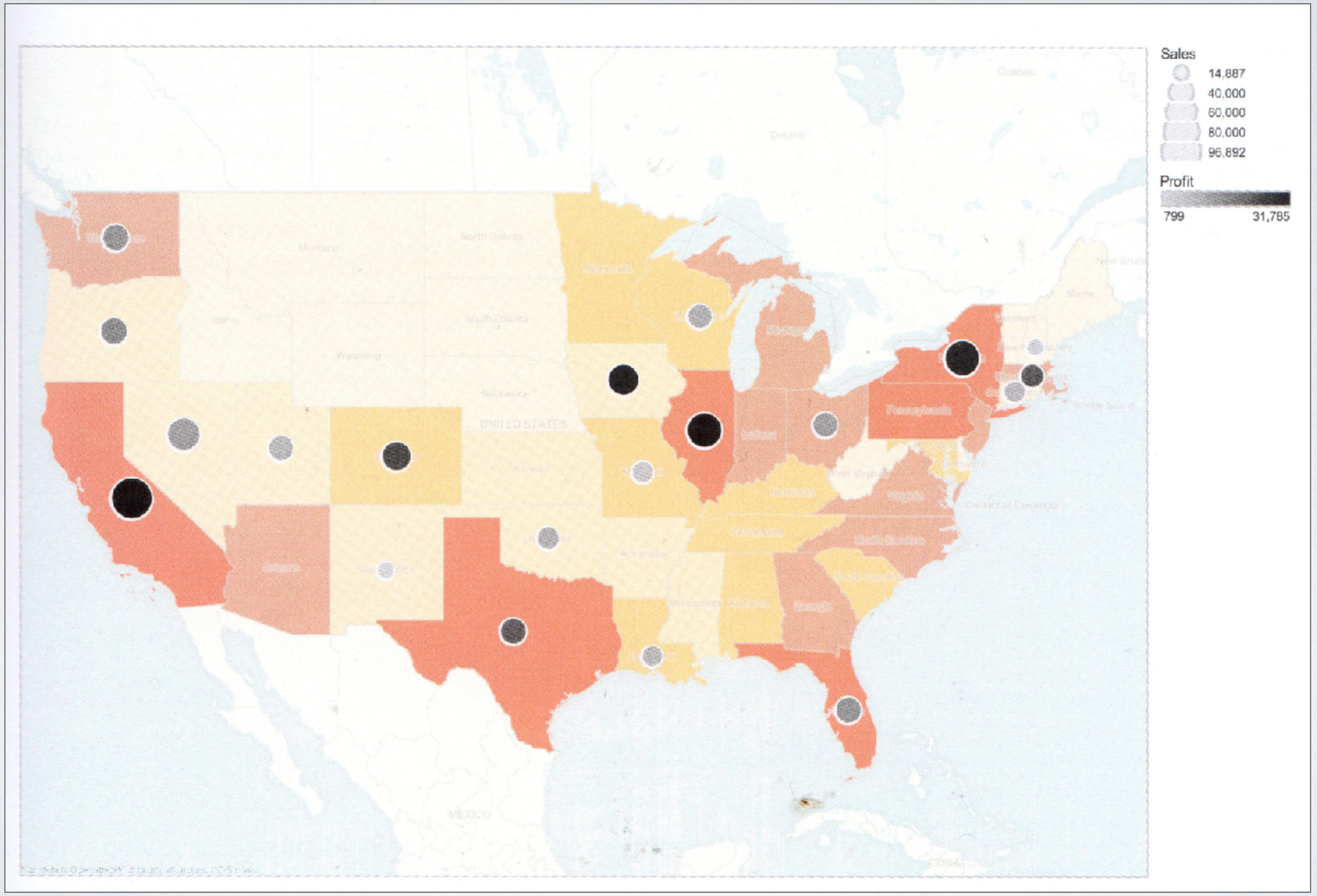
EXAMPLE: CIRCLE SIZE DISPLAYS VALUES FOR ENTIRE REGION



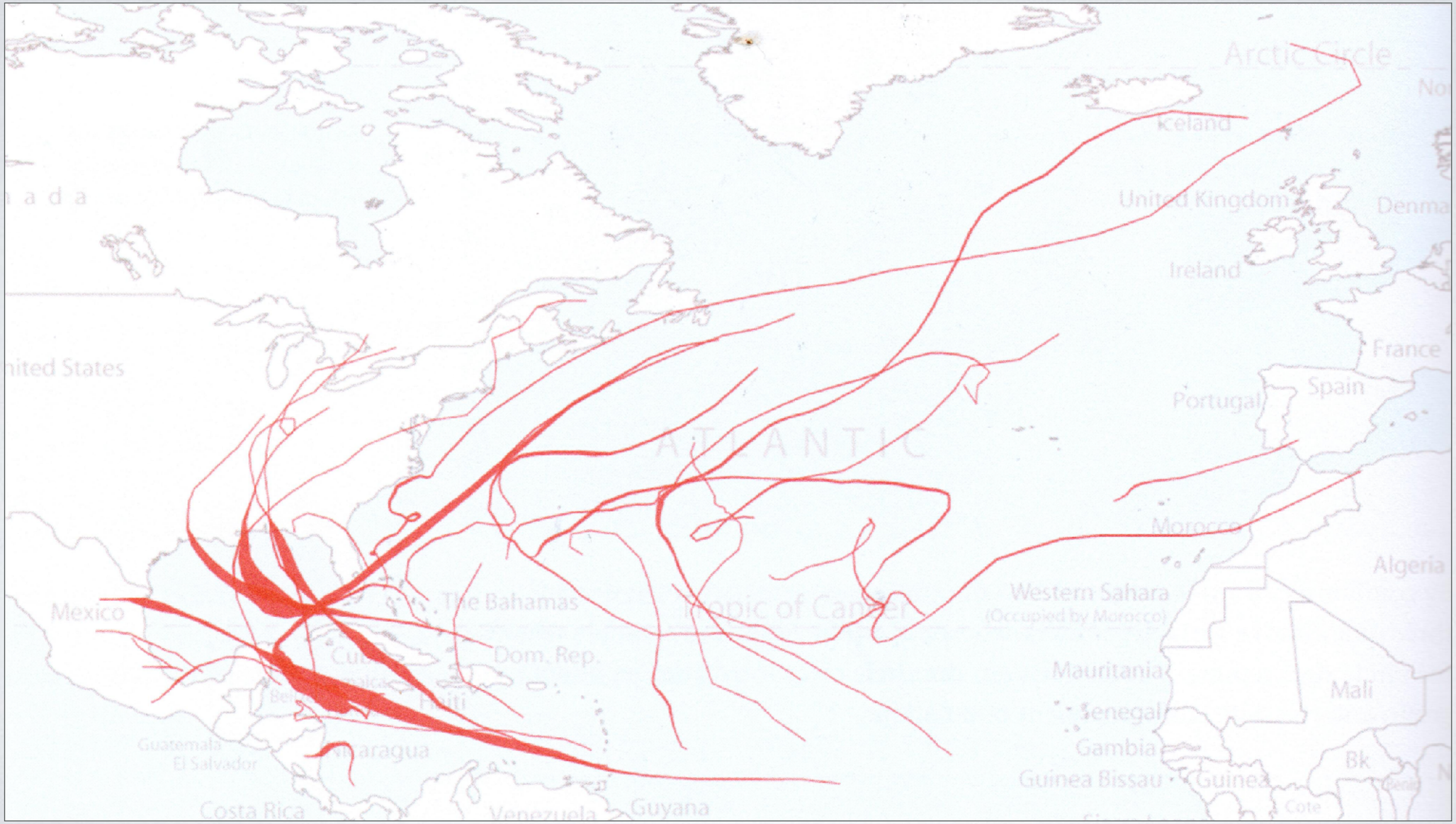
EXAMPLE: CHOROPLETH



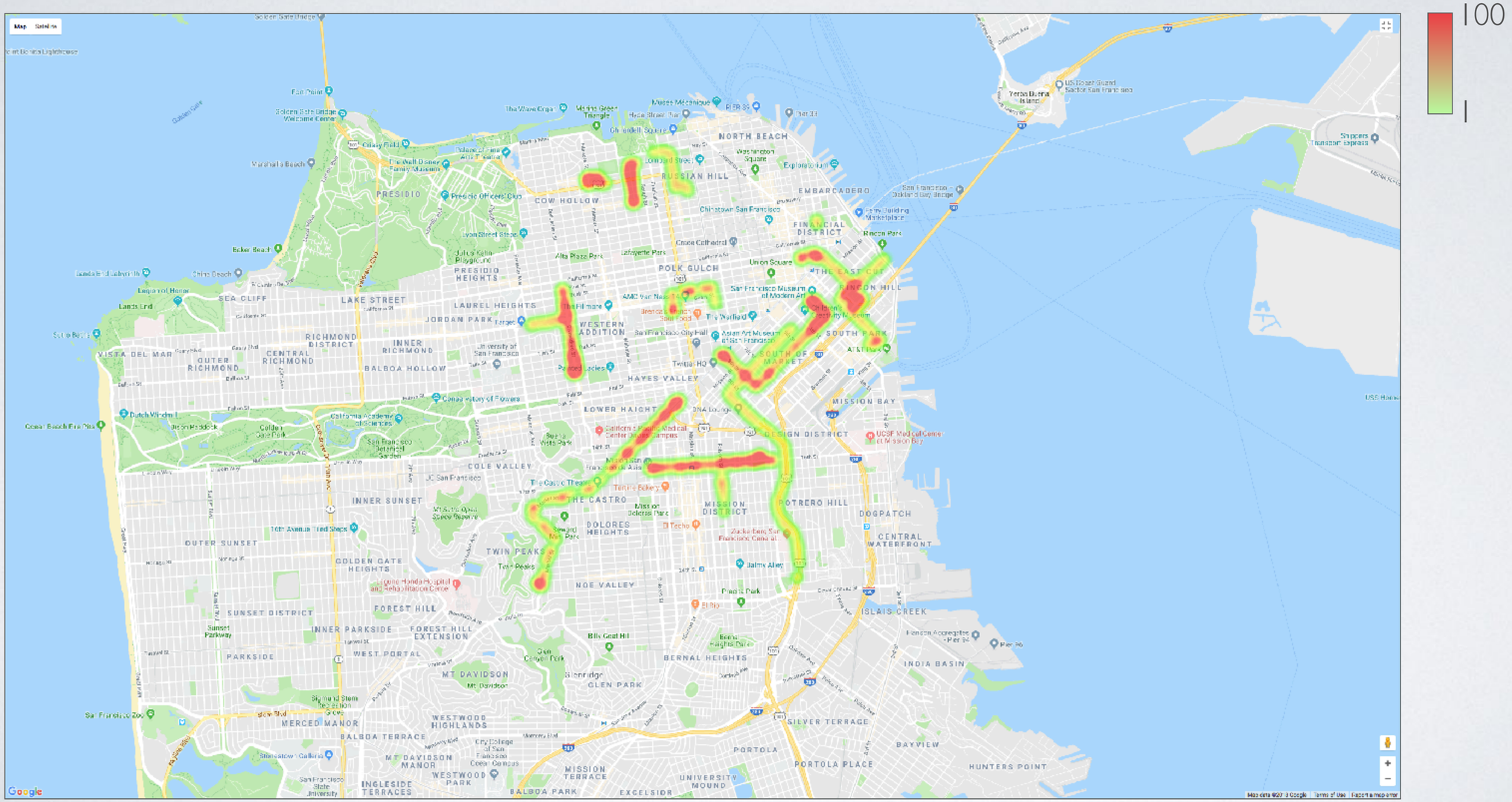
EXAMPLE: CHOROPLETH + CIRCLE SIZE + COLOR INTENSITY



EXAMPLE: LINE WIDTH



MAPPING QUANTITATIVE VALUES ON GRADIENT: HEATMAP



SUMMARY OF GRAPH VARIATIONS

	Value-encoding objects			
Relationship	Points	Lines	Bars	Boxes
Nominal comparison				
Time series				
Ranking				
Part-to-whole				

SUMMARY OF GRAPH VARIATIONS

	Value-encoding objects			
Relationship	**Points**	**Lines**	**Bars**	**Boxes**
Nominal comparison	Dot plot when bars can't be used because scale doesn't start at zero	No	Horizontal or vertical	No
Time series				
Ranking				
Part-to-whole				

SUMMARY OF GRAPH VARIATIONS

	Value-encoding objects			
Relationship	**Points**	**Lines**	**Bars**	**Boxes**
Nominal comparison	Dot plot when bars can't be used because scale doesn't start at zero	No	Horizontal or vertical	No
Time series	Dot plot, but only when values are not collected at consistent interval of time	Emphasis on overall pattern; categorical on x, quantitative on y	Emphasis on individual values; categorical on x, quantitative on y	Only when showing how distribution changes over time; categorical on x, quantitative on y
Ranking				
Part-to-whole				

SUMMARY OF GRAPH VARIATIONS

	Value-encoding objects			
Relationship	**Points**	**Lines**	**Bars**	**Boxes**
Nominal comparison	Dot plot when bars can't be used because scale doesn't start at zero	No	Horizontal or vertical	No
Time series	Dot plot, but only when values are not collected at consistent interval of time	Emphasis on overall pattern; categorical on x, quantitative on y	Emphasis on individual values; categorical on x, quantitative on y	Only when showing how distribution changes over time; categorical on x, quantitative on y
Ranking	Dot plot when bars can't be used because scale doesn't start at zero	No	Horizontal or vertical	Only when ranking multiple distributions; horizontal or vertical
Part-to-whole				

SUMMARY OF GRAPH VARIATIONS

	Value-encoding objects			
Relationship	**Points**	**Lines**	**Bars**	**Boxes**
Nominal comparison	Dot plot when bars can't be used because scale doesn't start at zero	No	Horizontal or vertical	No
Time series	Dot plot, but only when values are not collected at consistent interval of time	Emphasis on overall pattern; categorical on x, quantitative on y	Emphasis on individual values; categorical on x, quantitative on y	Only when showing how distribution changes over time; categorical on x, quantitative on y
Ranking	Dot plot when bars can't be used because scale doesn't start at zero	No	Horizontal or vertical	Only when ranking multiple distributions; horizontal or vertical
Part-to-whole	No	To display how parts of a whole changed over time	Horizontal or vertical	No

SUMMARY OF GRAPH VARIATIONS

	Value-encoding objects			
Relationship	Points	Lines	Bars	Boxes
Deviation				
Distribution				
Single				
Multiple				
Correlation				
Geospatial				

SUMMARY OF GRAPH VARIATIONS

	Value-encoding objects			
Relationship	**Points**	**Lines**	**Bars**	**Boxes**
Deviation	Dot plot when quantitative scale doesn't start at zero	Useful when combined with a time series	Horizontal or vertical, but always vertical when combined with time series	No
Distribution				
Single				
Multiple				
Correlation				
Geospatial				

SUMMARY OF GRAPH VARIATIONS

	Value-encoding objects			
Relationship	**Points**	**Lines**	**Bars**	**Boxes**
Deviation	Dot plot when quantitative scale doesn't start at zero	Useful when combined with a time series	Horizontal or vertical, but always vertical when combined with time series	No
Distribution				
Single	Strip plot, emphasis on individual values	Frequency polygon, emphasis on overall pattern	Histogram, emphasis on individual interval	No
Multiple				
Correlation				
Geospatial				

SUMMARY OF GRAPH VARIATIONS

	Value-encoding objects			
Relationship	**Points**	**Lines**	**Bars**	**Boxes**
Deviation	Dot plot when quantitative scale doesn't start at zero	Useful when combined with a time series	Horizontal or vertical, but always vertical when combined with time series	No
Distribution				
Single	Strip plot, emphasis on individual values	Frequency polygon, emphasis on overall pattern	Histogram, emphasis on individual interval	No
Multiple	Strip plot, emphasis on individual values	Frequency polygon, limit to a few lines	No	Box plot
Correlation				
Geospatial				

SUMMARY OF GRAPH VARIATIONS

	Value-encoding objects			
Relationship	**Points**	**Lines**	**Bars**	**Boxes**
Deviation	Dot plot when quantitative scale doesn't start at zero	Useful when combined with a time series	Horizontal or vertical, but always vertical when combined with time series	No
Distribution				
Single	Strip plot, emphasis on individual values	Frequency polygon, emphasis on overall pattern	Histogram, emphasis on individual interval	No
Multiple	Strip plot, emphasis on individual values	Frequency polygon, limit to a few lines	No	Box plot
Correlation	Scatter plot	No	Horizontal or vertical, in the form of a table lens	No
Geospatial				

SUMMARY OF GRAPH VARIATIONS

	Value-encoding objects			
Relationship	**Points**	**Lines**	**Bars**	**Boxes**
Deviation	Dot plot when quantitative scale doesn't start at zero	Useful when combined with a time series	Horizontal or vertical, but always vertical when combined with time series	No
Distribution				
Single	Strip plot, emphasis on individual values	Frequency polygon, emphasis on overall pattern	Histogram, emphasis on individual interval	No
Multiple	Strip plot, emphasis on individual values	Frequency polygon, limit to a few lines	No	Box plot
Correlation	Scatter plot	No	Horizontal or vertical, in the form of a table lens	No
Geospatial	Vary points size to encode values	To mark routes	No	No