Categories of Data Visualization

Data visualization

We briefly discuss **5 categories** of plots/charts here:

1. Comparison
2. Sequence
3. Distribution
4. Relationship
5. Part-whole

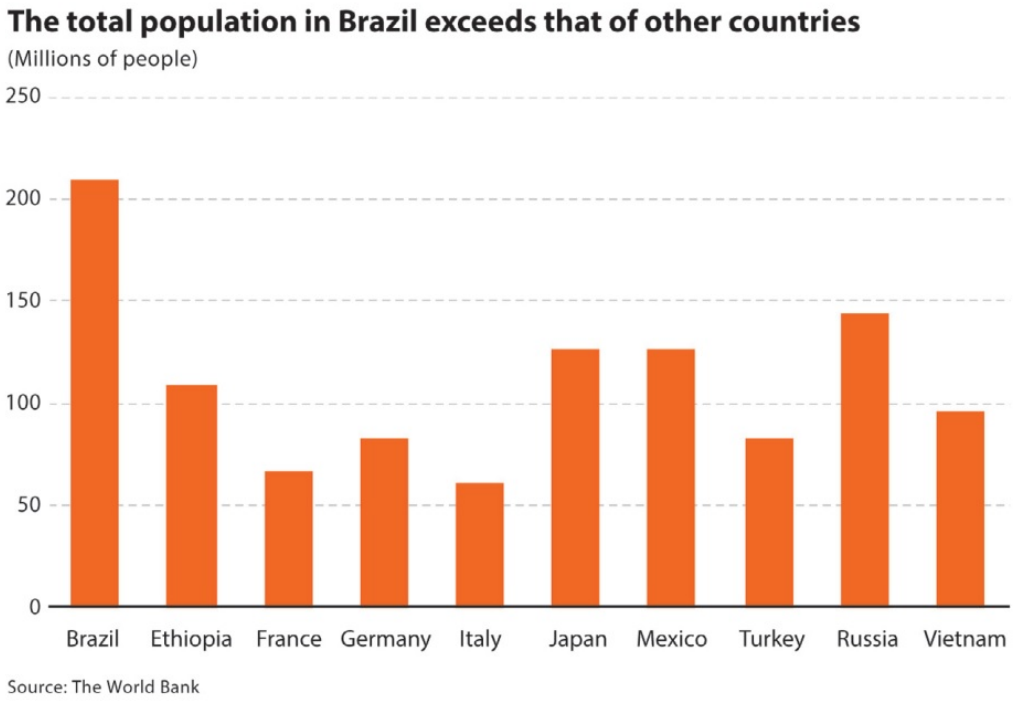
Comparison

Bar charts

- Consists of rectangles which can be arranged along the vertical axis so that the bars lie horizontally (often called a bar chart) or vertically on the horizontal axis (often called a column chart)
- The length or height of the rectangular bars in bar and column charts depict the value of your data.
- With rectangles sitting on the same straight axis, it's easy to compare the values quickly and accurately.

Data visualization

Comparison - Bar charts



The bar chart is a familiar chart that's easy to read and make. It sits at the top of the perceptual ranking matrix.

Data Source: The World Bank.

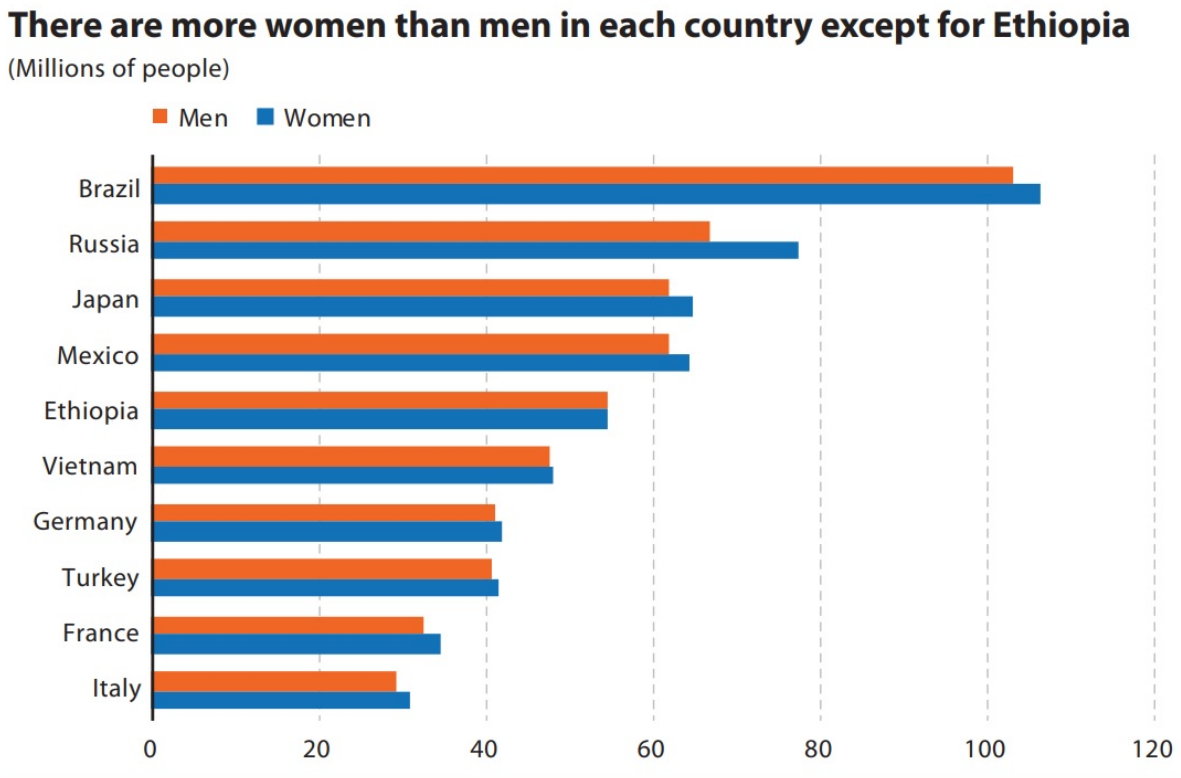
Comparison

Paired bar (a variation of bar charts)

- A simple bar chart is perfect for making comparisons across categories.
- If you want to show comparisons not just across but also within categories.

Data visualization

Comparison - Paired bar



A simple paired bar chart is familiar to most readers and easy to read.

Data Source: The World Bank.

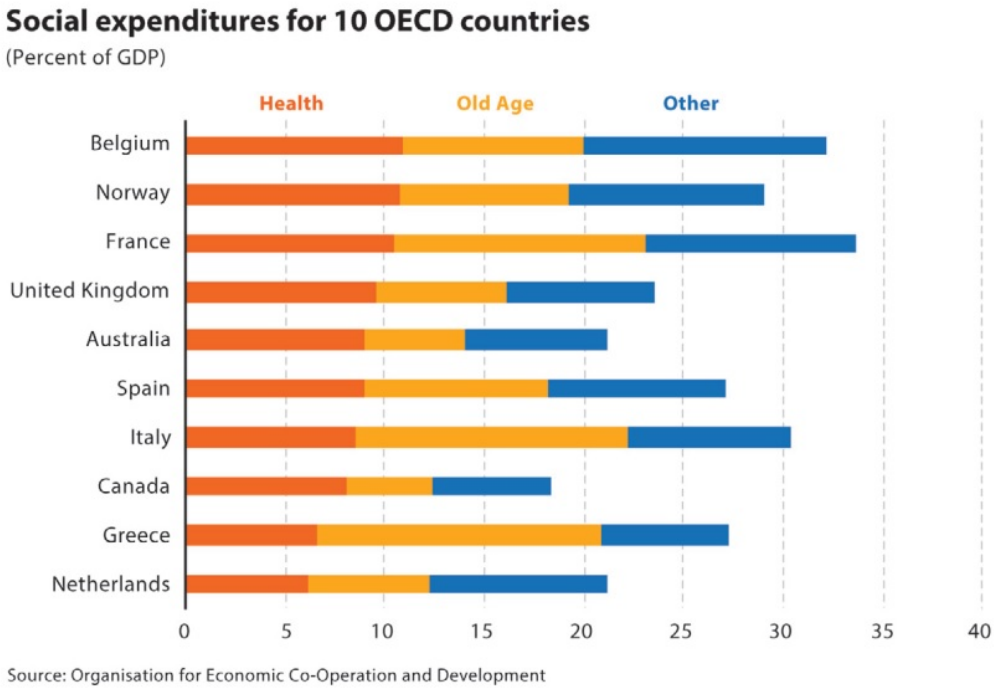
Comparison

Stacked bar (a variation of bar charts)

- While the paired bar chart shows two or more data values for each category, this chart subdivides the data within each category.
- For better comparison not just across but also within categories.
- The categories could sum to the same total, say, 100 percent, so that the total length of the bar is the same for every group. Or the totals may differ across the groups, in which case the total length of each bar may differ.
- Drawback: it can be difficult to compare the different values of the segments within the chart.

Data visualization

Comparison - Stacked bar



The stacked bar charts shows how different categories sum to a total. The interior series in the chart, however, are harder to compare with one another because they do not sit on the same baseline.

Data Source: Organisation for Economic Co-Operation and Development.

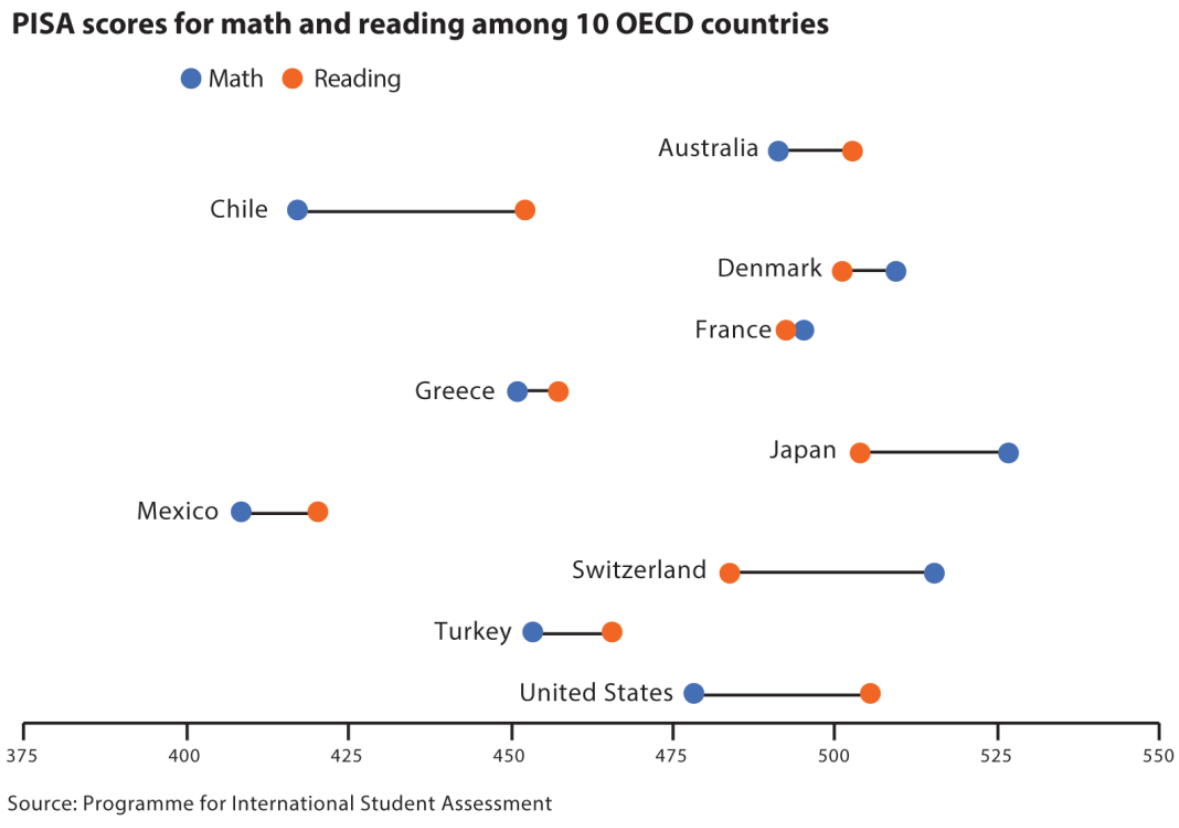
Comparison

Dot plot (a variation of bar charts)

- The dot plot is an easy way to compare categories (especially many categories) when paired bars add too much clutter.
- In the dot plot each data value is connected by a line to show the range or difference.
- Drawback: It is not entirely obvious when the direction of the "difference" change

Data visualization

Comparison - Dot plot



The basic dot plot places a dot for each data point and connects them with a line. Notice how more white space lightens the visualization.

Comparison

Heatmap

- Heatmaps use colors and color saturations to represent data values.
- Simply put, a heatmap is a table with color-coded cells. They are often used to visualize high-frequency data or when seeing general patterns is more important than exact values.
- Heatmaps can facilitate comparison both across and within categories.
- Heatmaps can be used for purposes other than comparison as well.

Data visualization

Comparison - Heatmap

Heatmaps use colors and color saturations to represent data values and can focus the reader's attention along the columns or across the rows.

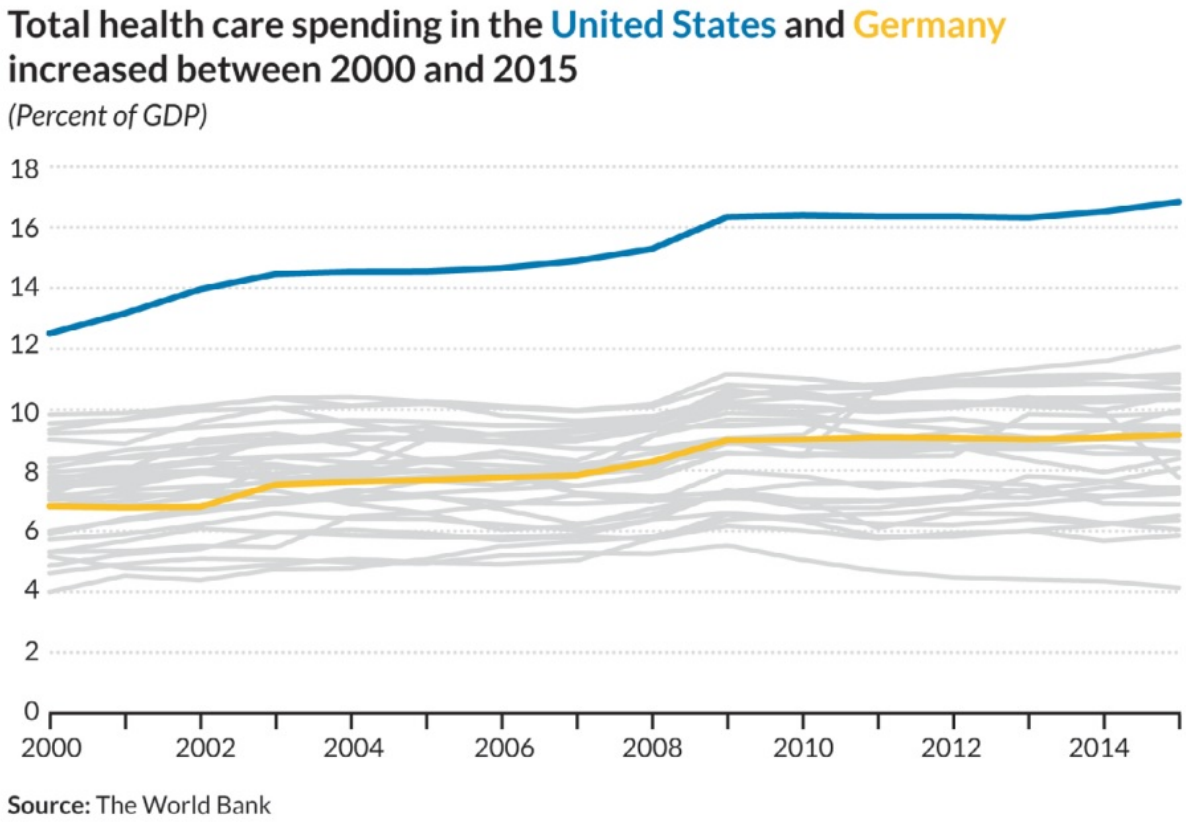
Sequence

Line chart

- Data values are connected by lines to show values in a sequence.
- Line charts help with the detection of trends and patterns.
- There is no hard rule to dictate the number of sequences you can include in a single line chart so far as the final chart is comprehensible.
- So, We might also take the line graph and break it into multiple graphs.

Data visualization

Sequence - Line chart



There is no hard rule to dictate the number of series you can include in a single line graph.

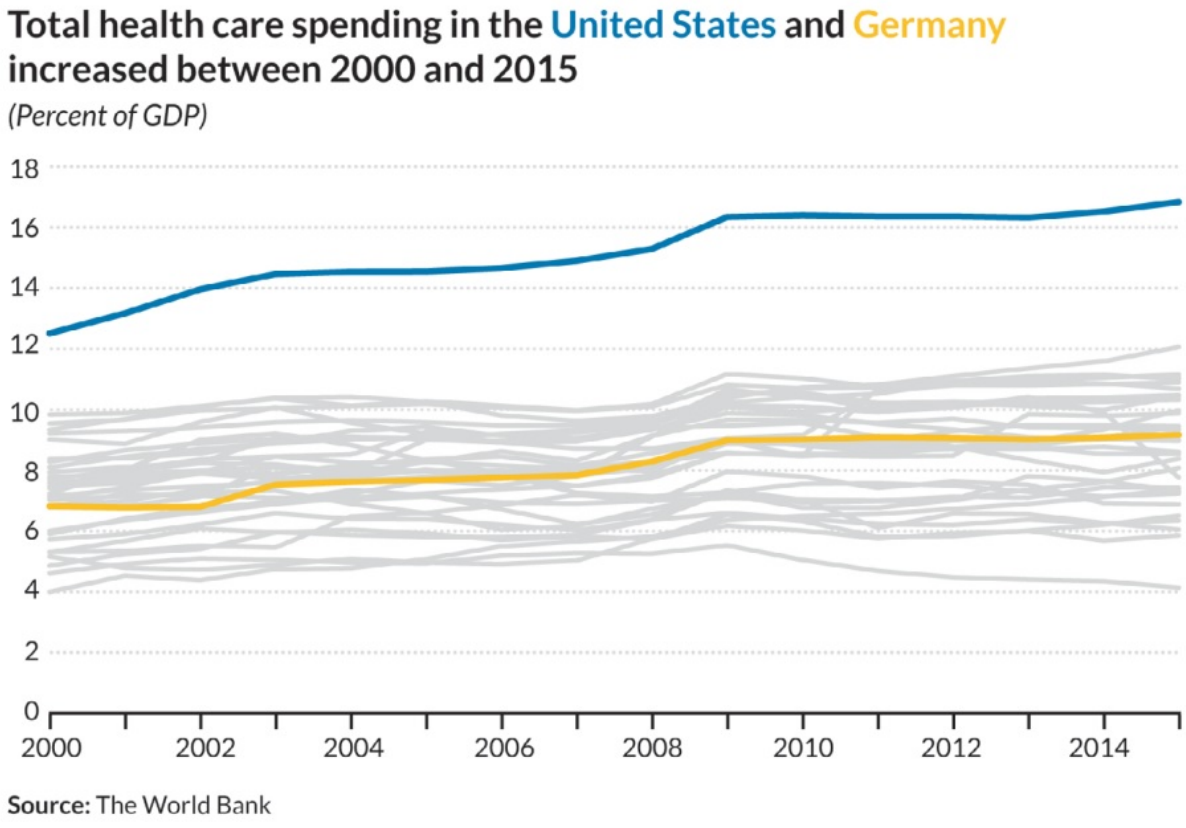
Sequence

Line chart

- Data values are connected by lines to show values in a sequence.
- Line charts help with the detection of trends and patterns.
- There is no hard rule to dictate the number of sequences you can include in a single line chart so far as the final chart is comprehensible.
- So, We might also take the line graph and break it into multiple graph.

Data visualization

Sequence - Line chart



There is no hard rule to dictate the number of series you can include in a single line graph.

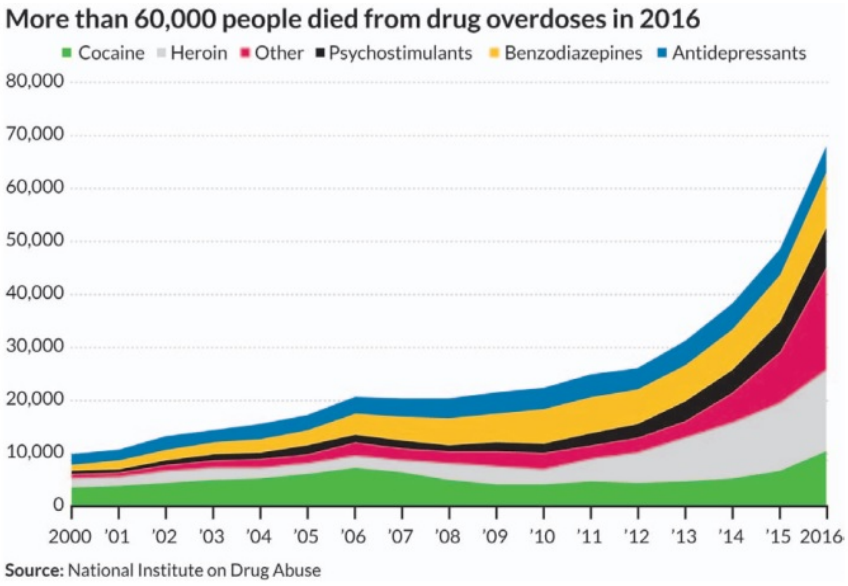
Sequence

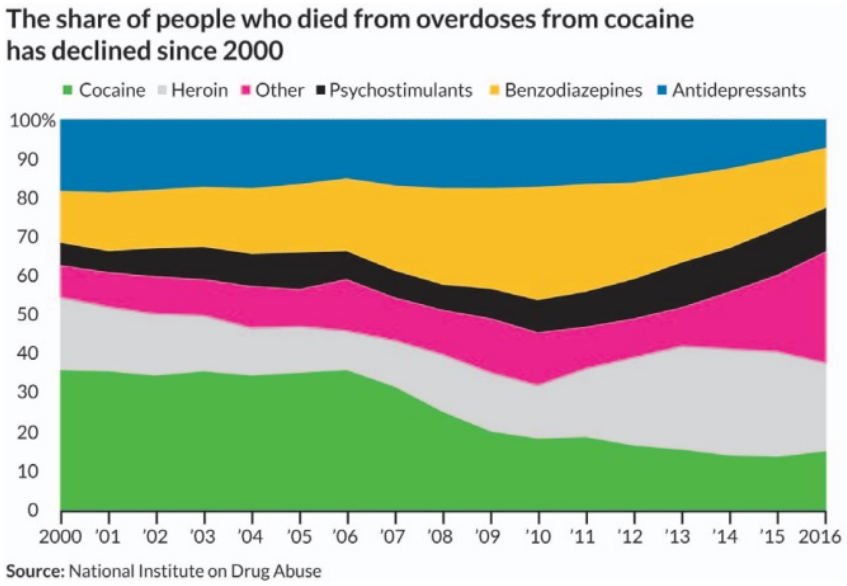
Area chart & Stacked area chart

- Area charts are line graphs with the area below the line filled in, giving the series more visual weight.
- Stacked area charts build on the typical area chart by showing multiple data series simultaneously.
- Instead of sitting independently of one another as in the previous chart, the data in a stacked area chart sum to a total or a percentage.

Data visualization

Sequence - Area chart & Stacked area chart





Stacked area charts build on the typical area chart by showing multiple data series simultaneously and sum to a total, often 100 percent.

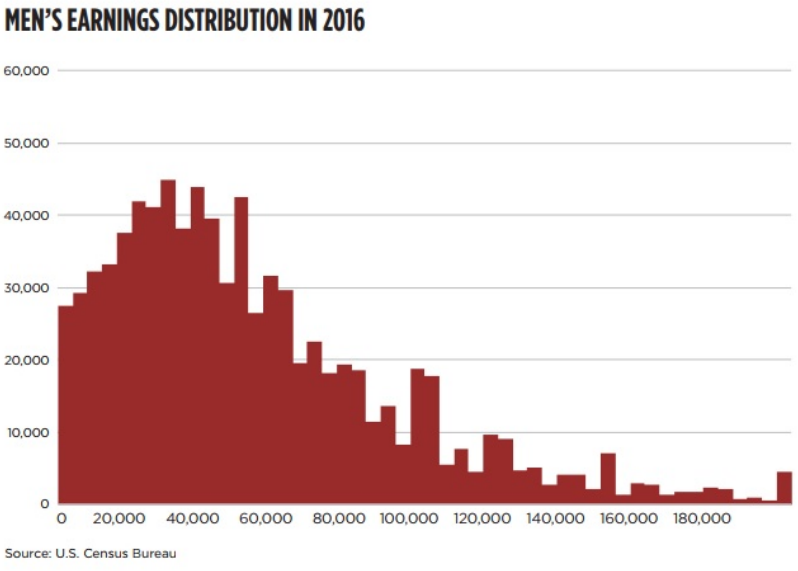
Distribution

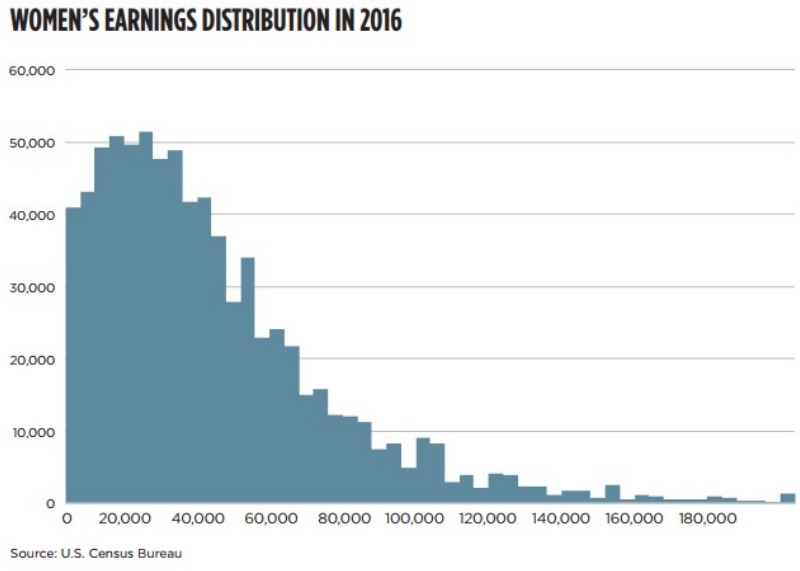
Histogram

- The histogram is the most basic graph type for visualizing a distribution.
- It is a specific kind of bar chart that presents the tabulated frequency of data over distinct intervals, called bins, that sum to the total distribution.
- The entire sample is divided into these bins, and the height of each bar shows the number of observations within each interval.
- Histograms can show where values are concentrated within a distribution, where extreme values are, and whether there are any gaps or unusual values.

Data visualization

Distribution - Histogram

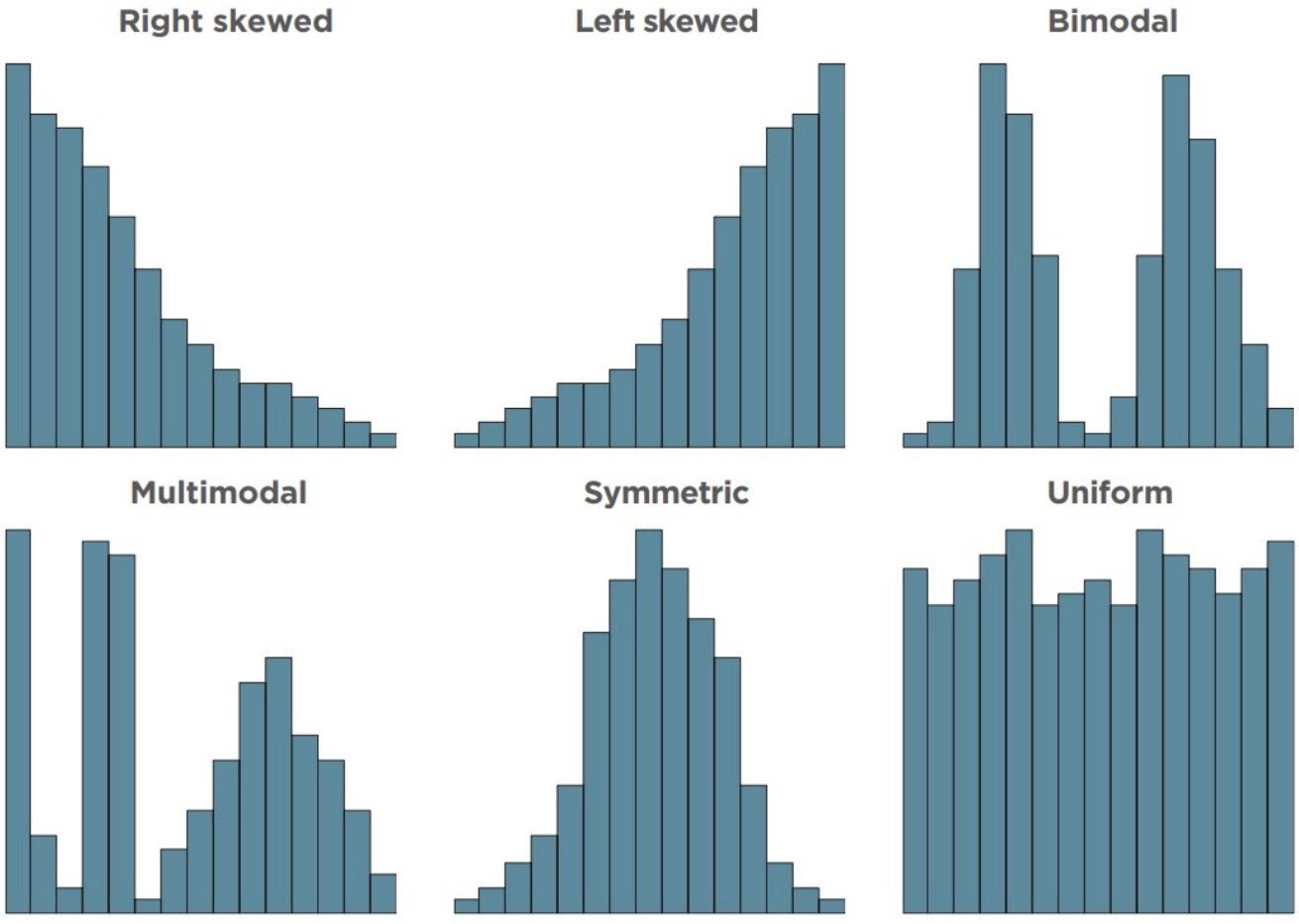




Histograms divide the entire sample into intervals (also called "bins"). The height of the bin shows the number of observations within it.

Data visualization

Distribution - Histogram (forms)



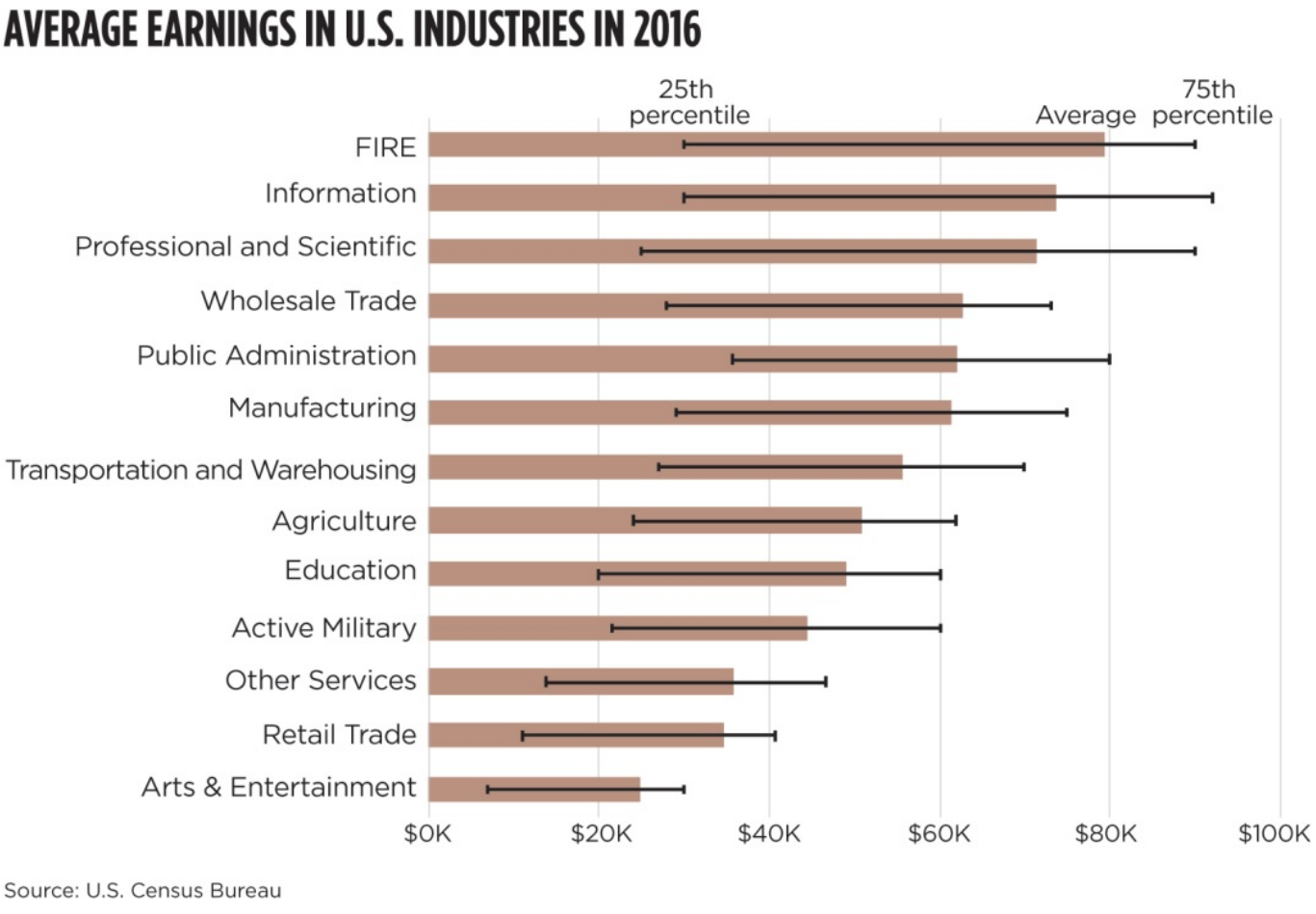
Distribution

Histogram with error bars

- The simplest and most common way to visualize uncertainty is to use error bars: small markers that denote the error margin or confidence interval.
- Error bars are an addition to other charts, often bar or line charts.
- The ends of the error bars can correspond to any value you choose: percentiles, the standard error, the 95-percent confidence interval, or even a fixed number.

Data visualization

Distribution - Histogram with error bars



The simplest and most common way to visualize uncertainty or distributions is to use error

Distribution

Box-and-whisker plot

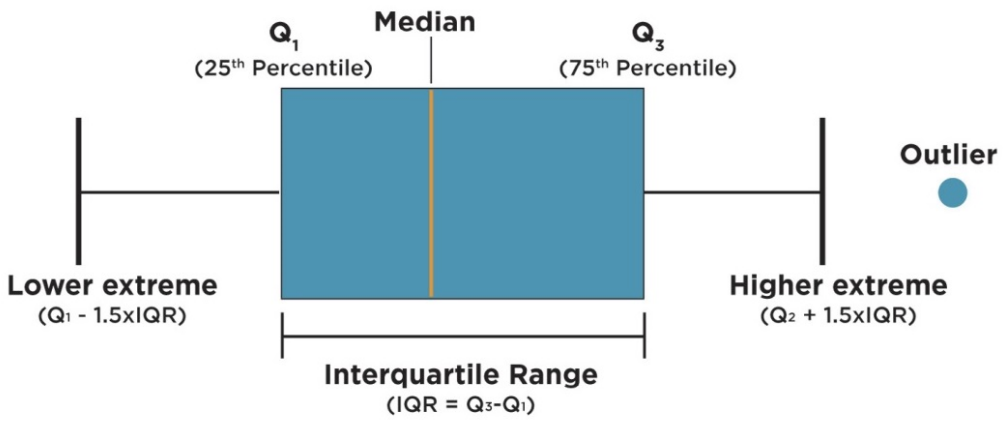
- When you visualize the distribution of your data, you can show the **entire distribution** or just **specific points** within it.
- The **box-and-whisker plot** (or boxplot), uses a box and line **markers** to show specific **percentile** values within a distribution.
- You can also add **markers** to show **outliers** or other **interesting data points** or values.
- It is a **compact summary** of the data **distribution**, though it displays **less detail** than a histogram or violin chart.

Data visualization

Distribution - Box-and-whisker plot

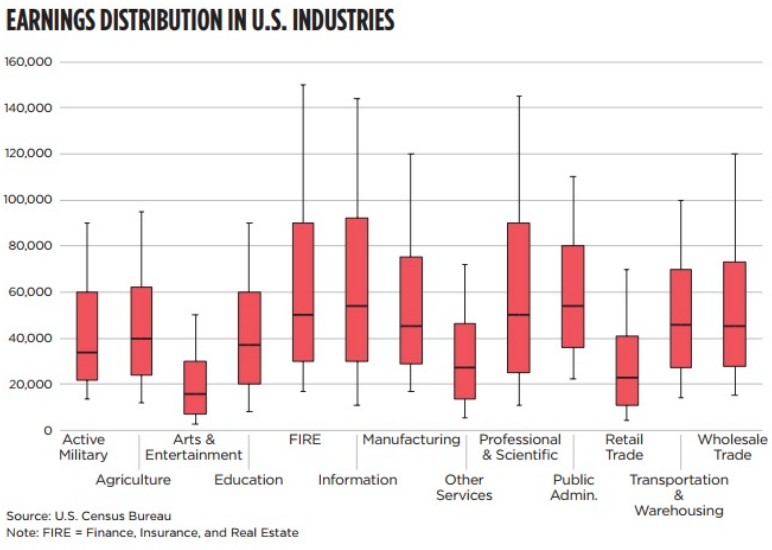
Most standard **box-and-whisker plots** have five major components:

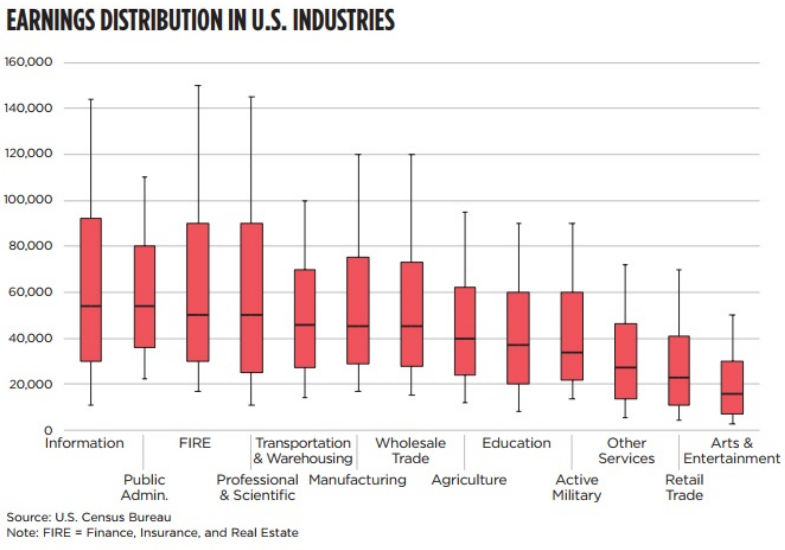
1. The median, encoded by a single horizontal line inside the box.
2. Two hinges, which are the upper and lower edges of the box signify the IRQ.
3. The higher and lower extremes (sometimes the maximum and minimum) are placed at a position 1.5 times the IQR.
4. Two whiskers (the lines) connect the hinges to a specific observation (for example a defined extreme) or percentile.
5. Outliers are individual data points that are further away from the median than the edges of the whiskers.



Data visualization

Distribution - Box-and-whisker plot





These charts show the distribution of earnings in thirteen industries either sorted alphabetically (left) or by median value (right). The edges of the box show the 25th and 75th percentiles and the whiskers show the 10th and 90th percentiles.

Data visualization

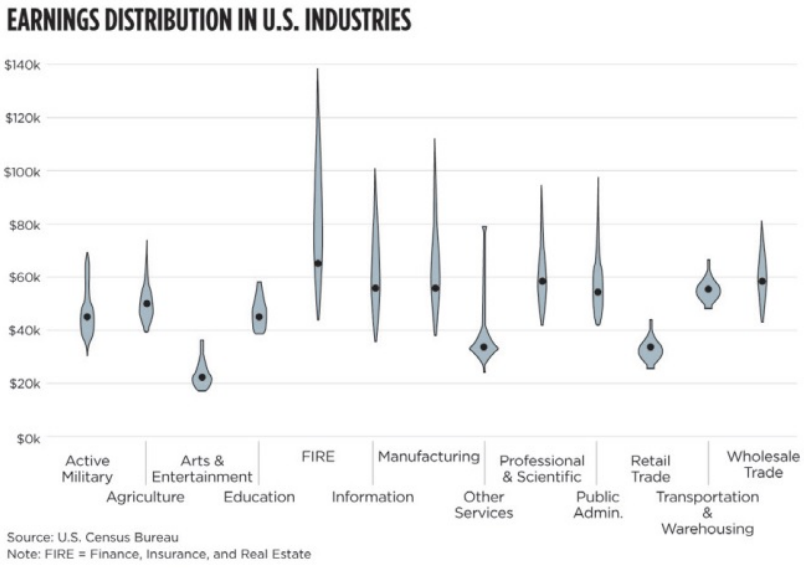
Distribution

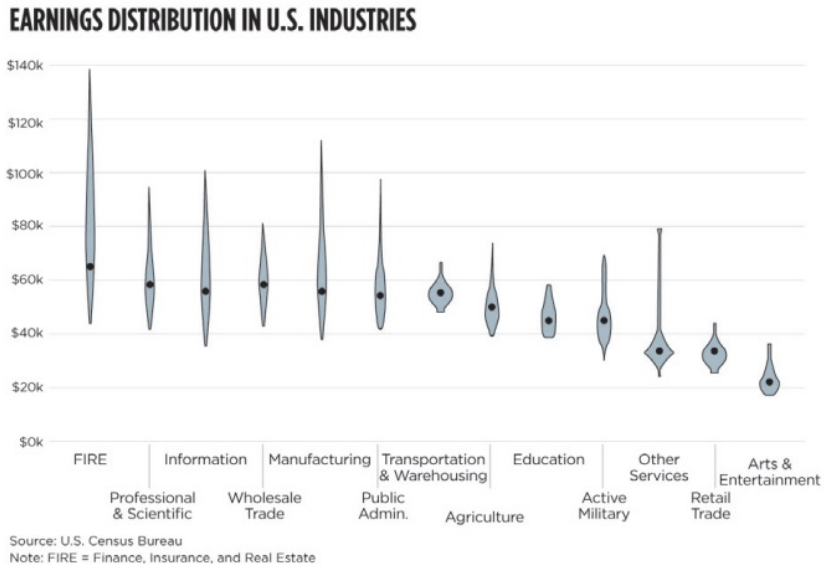
Violin chart

- Unlike the box-and-whisker plot, in which we choose specific points in the distribution, or the histogram in which values are grouped together into intervals, the violin chart shows the shape of the whole distribution.
- One consideration in creating this chart type is that it requires estimating what is called the kernel density of each distribution. Kernel densities are a way to estimate the distribution of a variable—akin to a histogram—but can be smoothed or made to look more continuous using different algorithms. For the violin plot, those density estimates are plotted to mirror each other around an invisible central line.
- One can say A histogram plots a summary view of a distribution along a single axis. The violin plot mirrors a smoothed version of the histogram on either side of that single axis.

Data visualization

Distribution - Violin chart





Instead of showing select points (percentiles) in a data distribution, the violin chart shows the estimated shape of the entire distribution using kernel densities.

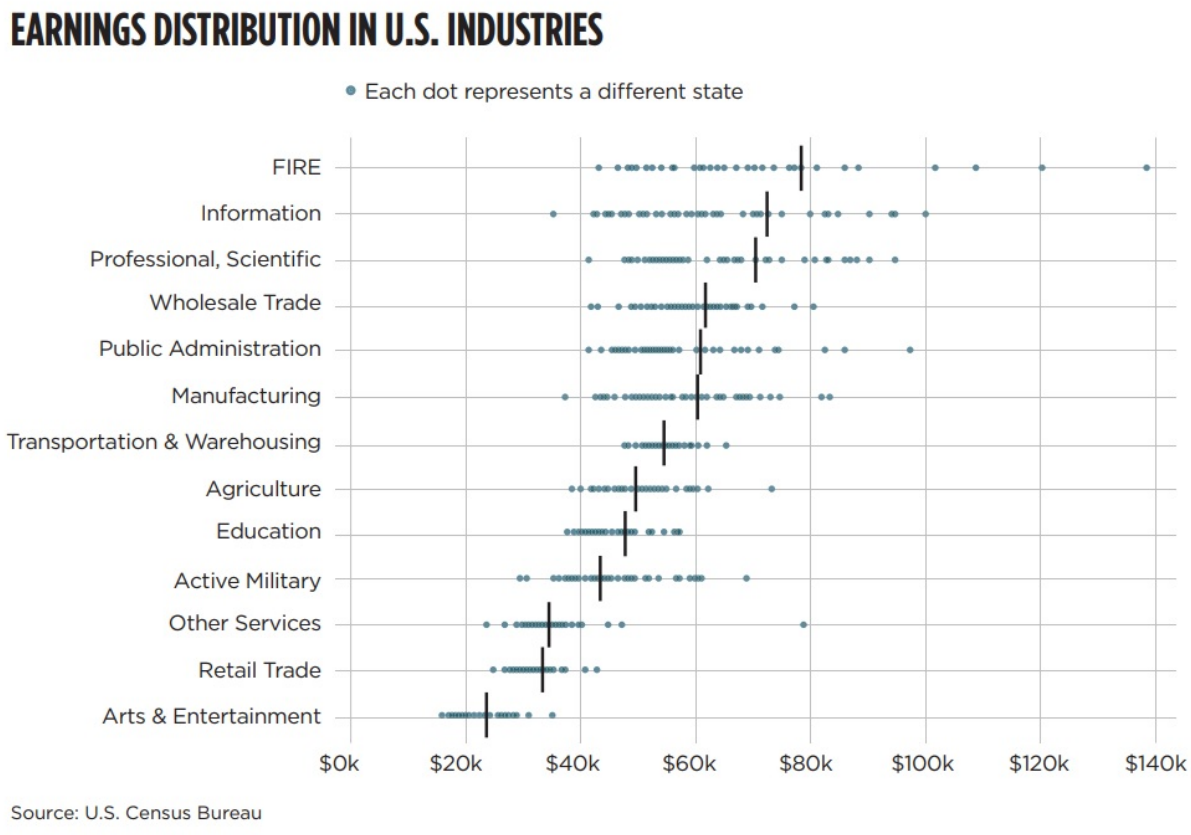
Distribution

Strip plot

- In this graph type, the data points are plotted along a single horizontal or vertical axis.
- Some data-points in strip plots can become obscured. But, especially by virtue of the overlapping transparent colors, the patterns emerge as dark gatherings.
- There's no rule for how many data points are too many, but as you plot your data, you can always tell when you've passed that threshold.
- One way to make the data more visible is to use a technique called jittering. This is when we alter the placing of individual data-points so that they spread. The resultant plot is usually called beeswarm plot.

Data visualization

Distribution - Strip plot

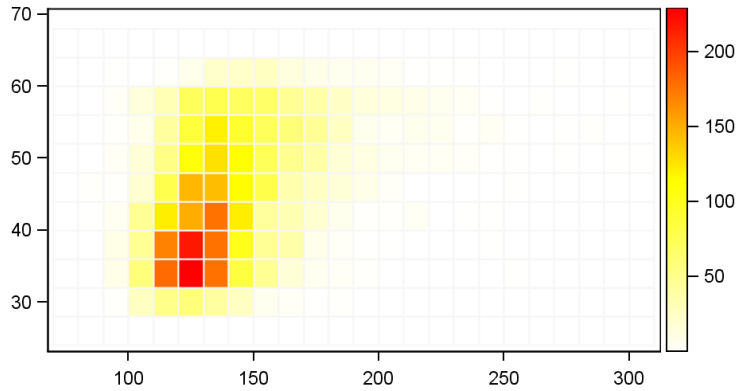


In a strip plot, data points are plotted along a single horizontal or vertical axis. This strip plot encodes the data with circles, but small lines are also often used.

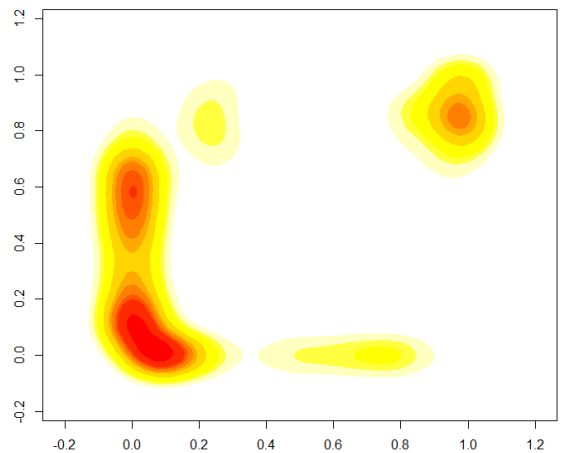
Data visualization

Distribution - Higher dimensions

The heatmap idea (color density) can be used to visualize 2 dimensional distributions. Such heatmaps consists of a grid of beans in which the color density of each bin represents the the number of observations in that bin (also called 2d histogram).



To get a smooth 2 dimensional distribution, One can employ the idea of kernel density estimation and color densities (note that the image on the right is not depicting the same distribution as above).



Note. Visualizing distribution of higher dimensions is also conceivable, albeit not the focus of this course.

Data visualization

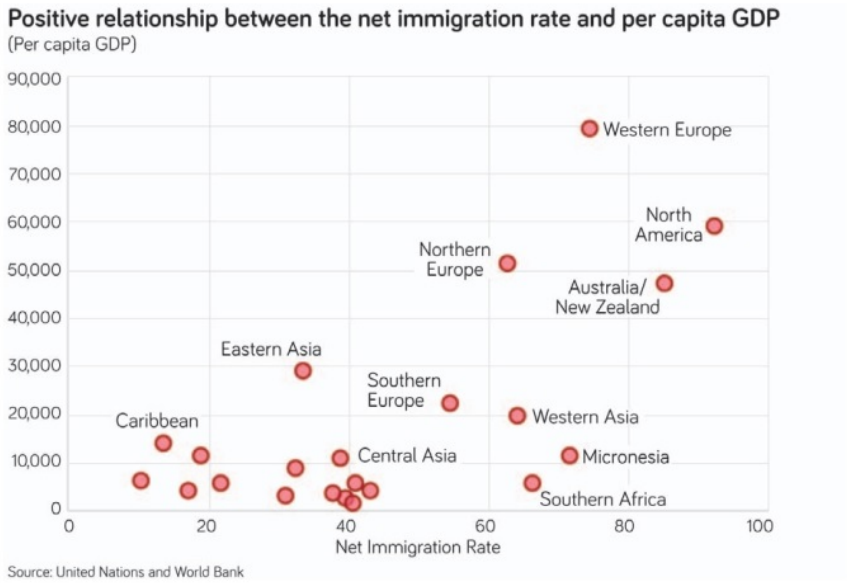
Relationship

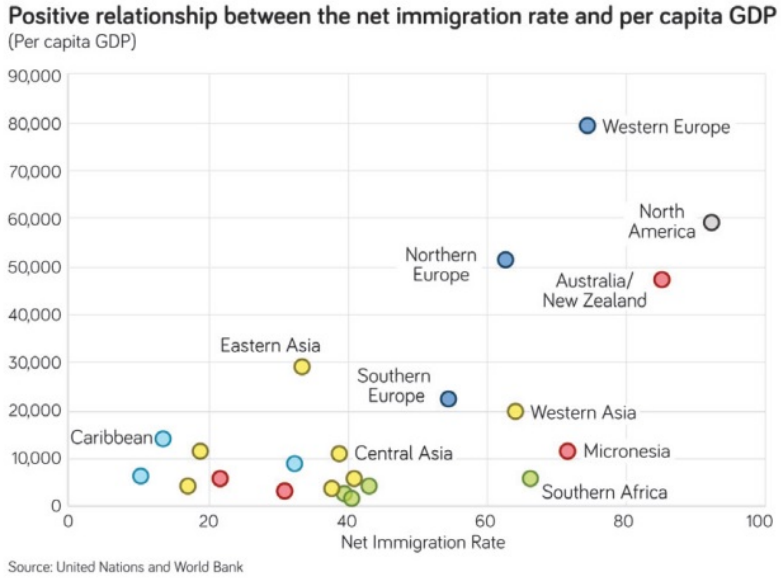
Scatter plot

- The scatter plot is perhaps the most common visualization to illustrate correlations (or lack thereof) between two variables.
- One variable is plotted along a horizontal axis, and the other along a vertical axis. The specific observations (not binning) are plotted in the created space.
- Different types of secondary visual elements or annotation can accompany a scatter plot (for example, line of best fit, node coloring, etc.).

Data visualization

Relationship - Scatter plot





Both scatterplots show the association between net immigration and per capita GDP, using either a single transparent color (left) or different colors for regions of the world (right).

Relationship

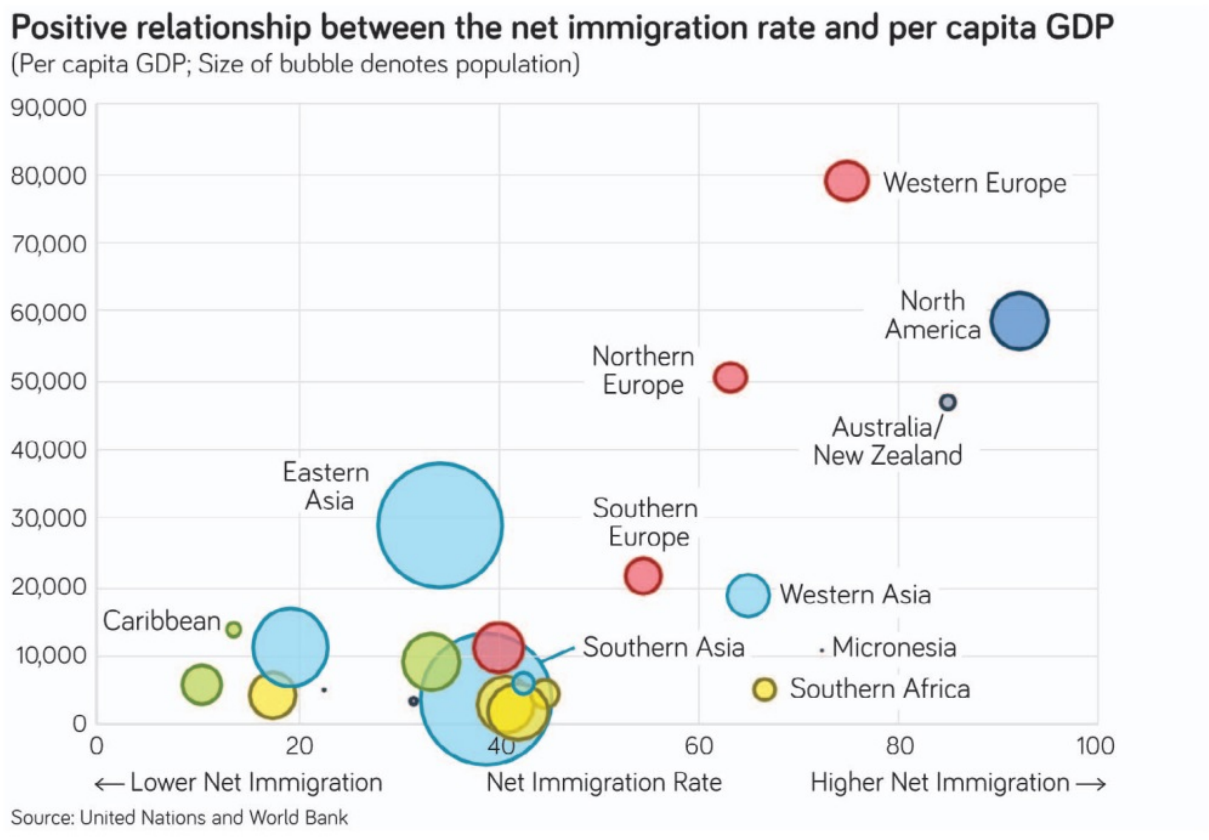
Bubble plot

- The scatter plot can be transformed into a bubble plot (or bubble scatter plot) by varying the sizes of the circles according to a third variable.
- Note that if there already exists a third dimension in the plot (for example the color/shape of the nodes), the size of the bubbles then could represent a fourth dimension.
- Also note that bubble on their own with underlying axes constitute another type of plot which falls under the comparison category.

Data visualization

Note. Here, the size of the circles corresponds to the population in each region.

Relationship - Bubble plot



As before, more colors can be added to denote another variable, such as region of the world.

Data visualization

Relationship

Correlation matrix

- A correlation matrix is a matrix with the variables listed along the horizontal and vertical axes.
- The numbers in the cells represent the strength of the correlation (Pearson's, Spearman's, etc.)
- For a large number of variables or groupings of the variables, a matrix consisting of numbers barely reveals the underlying patterns.
- A correlation matrix graph in which the stregnth of the correlation in each cell is represented by a color density (similar a heatmap) or a sized shape (usually circles), can reveal the underlying patterns.
- A correlation matrix (graph) can also accommodate other visual elements.

Data visualization

Relationship - Correlation matrix

World Migration

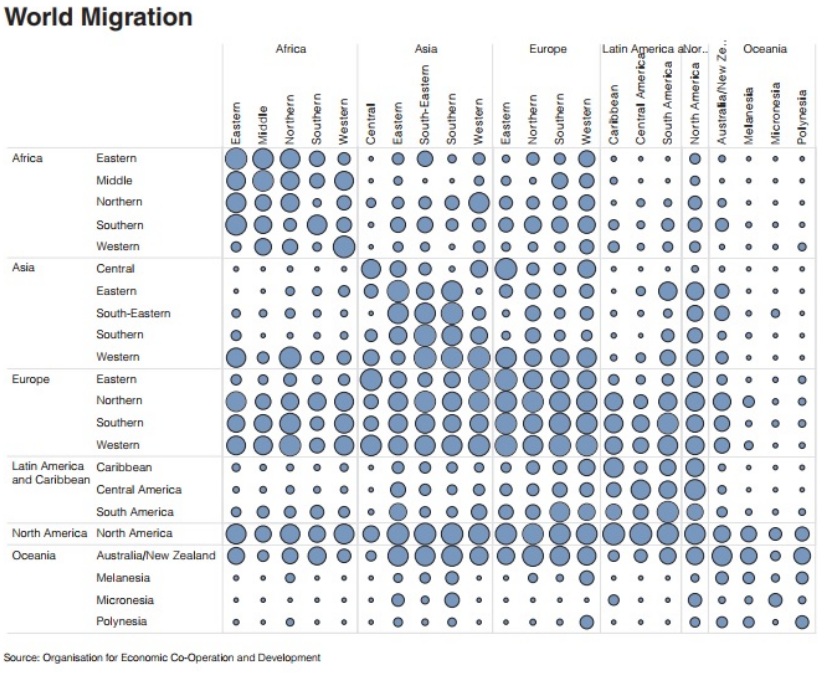
		Africa					Asia					Europe				Latin American and Caribbean			North America	Oceania			
		Eastern	Middle	Northern	Southern	Western	Central	Eastern	South-Eastern	Southern	Western	Eastern	Northern	Southern	Western	Caribbean	Central America	South America	North America	Australia/New Zealar	Melanesia	Micronesia	Polynesia
Africa	Eastern	49.0	10.0	7.1	0.6	0.1	0.0	0.1	0.6	0.0	0.1	0.0	0.1	0.1	1.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	Middle	3.9	17.0	3.7	0.5	4.3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.4	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	Northern	7.2	1.1	3.2	0.0	0.4	0.0	0.1	0.1	0.4	9.0	0.2	0.2	0.3	0.8	0.0	0.0	0.0	0.2	0.0	0.0	0.0	0.0
	Southern	14.0	2.0	0.2	7.1	0.5	0.0	0.5	0.8	0.1	0.2	0.4	1.7	1.1	1.8	0.0	0.0	0.1	0.3	0.2	0.0	0.0	0.0
	Western	0.0	1.5	0.6	0.0	58.0	0.0	0.0	0.0	0.0	0.1	0.0	0.0	0.0	0.4	0.1	0.0	0.0	0.1	0.0	0.0	0.0	0.0
Asia	Central	0.0	0.0	0.0	0.0	0.0	4.9	1.0	0.1	0.0	1.6	44.0	0.1	0.1	2.4	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	Eastern	0.0	0.0	0.0	0.0	0.1	0.3	53.0	1.8	12.0	0.0	0.2	0.6	0.1	0.3	0.0	0.0	3.6	2.2	0.4	0.0	0.0	0.0
	South-Eastern	0.0	0.0	0.0	0.0	0.0	0.0	9.2	13.0	68.0	0.2	0.0	0.6	0.1	0.2	0.0	0.0	0.0	0.9	0.4	0.0	0.0	0.0
	Southern	0.0	0.0	0.0	0.0	0.0	0.1	2.1	110.0	8.6	1.6	0.0	0.6	0.1	0.1	0.0	0.0	0.2	0.5	0.0	0.0	0.0	0.0
	Western	6.0	0.1	38.0	0.2	0.3	0.9	0.3	170.0	40.0	130.0	12.0	1.7	2.9	5.4	0.0	0.0	0.6	1.4	0.1	0.0	0.0	0.0
Europe	Eastern	0.0	0.0	0.2	0.0	0.1	56.0	1.5	0.4	1.0	21.0	100.0	4.7	2.5	3.9	0.0	0.0	0.1	0.5	0.0	0.0	0.0	0.0
	Northern	8.5	0.7	2.0	2.4	4.5	0.4	5.0	23.0	5.9	8.8	26.0	20.0	8.9	8.8	2.4	0.3	2.7	4.2	2.2	0.1	0.0	0.0
	Southern	2.0	2.4	15.0	0.3	5.3	0.5	4.1	6.2	2.3	3.2	32.0	5.8	31.0	16.0	3.8	1.7	25.0	1.9	0.6	0.0	0.0	0.0
	Western	4.2	4.0	34.0	0.5	6.5	12.0	4.8	8.7	8.5	31.0	60.0	7.7	49.0	29.0	2.0	0.7	7.3	3.6	0.6	0.0	0.0	0.0
Latin American and Caribbean	Caribbean	0.0	0.0	0.0	0.0	0.0	0.0	0.1	0.1	0.1	0.0	0.0	0.2	0.3	1.4	7.1	0.2	1.0	2.6	0.0	0.0	0.0	0.0
	Central America	0.0	0.0	0.0	0.0	0.0	0.0	0.5	0.1	0.0	0.1	0.1	0.1	0.5	0.4	0.5	6.5	2.1	9.8	0.0	0.0	0.0	0.0
	South America	0.0	0.1	0.1	0.2	0.1	0.0	1.8	0.1	0.1	0.6	0.5	0.5	8.0	1.6	0.8	0.5	42.0	1.3	0.0	0.0	0.0	0.0
North America	North America	8.2	1.3	6.7	1.5	8.1	1.2	53.0	47.0	53.0	17.0	23.0	18.0	19.0	15.0	66.0	160.0	34.0	12.0	1.6	0.7	0.2	0.4
Oceania	Australia/New Zealand	1.6	0.1	0.8	2.5	0.2	0.0	8.8	8.3	10.0	3.1	1.9	18.0	6.8	3.7	0.1	0.2	1.3	2.1	7.4	1.4	0.0	1.5
	Melanesia	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.2	0.0	0.0	0.0	0.0	0.4	0.0	0.0	0.0	0.0	0.2	0.1	0.0	0.1
	Micronesia	0.0	0.0	0.0	0.0	0.0	0.0	0.1	0.0	0.5	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.2	0.0	0.0	0.2	0.0
	Polynesia	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.2	0.0	0.0	0.0	0.0	0.1	0.1	0.0	0.2

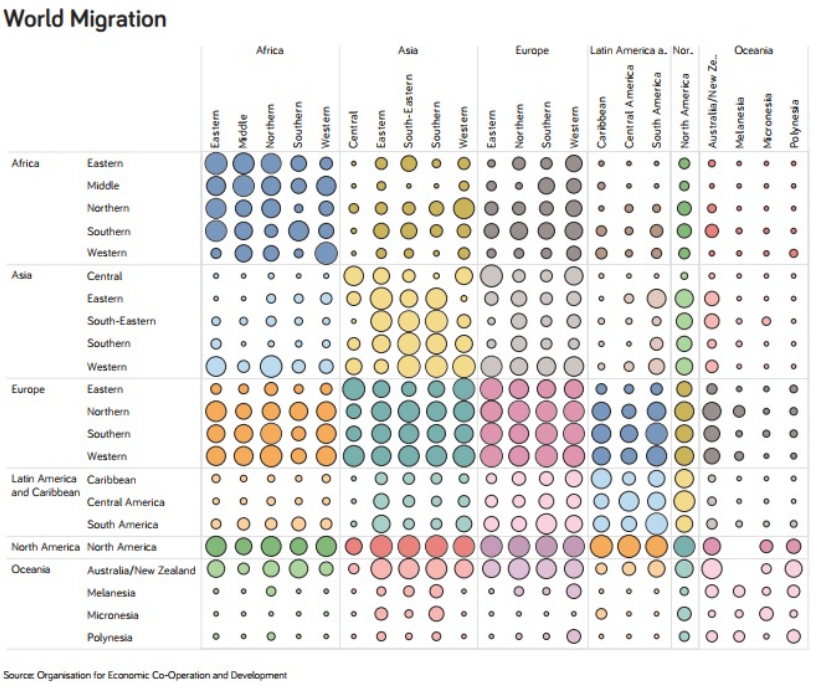
Source: Organisation for Economic Co-Operation and Development
Note: Data limited to a minimum of 200,000 immigrants or emigrants

The basic correlation matrix is a table with numbers that show the strength of the relationship between observations.

Data visualization

Relationship - Correlation matrix





An alternative to the correlation matrix table is to use circles or other shapes, to which color can be added to visually organize the space.

Data visualization

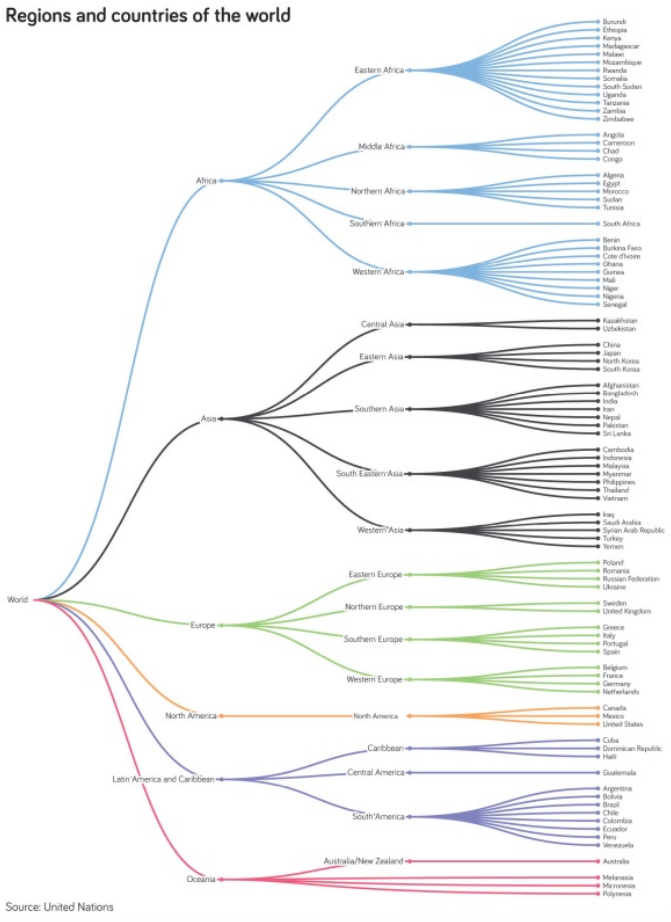
Relationship

Tree diagrams

- Tree diagrams show levels of a hierarchy.
- Nodes branch outward from an initial node connected by lines called links, link lines, or branches.
- The initial node is called the root and is the parent to all other nodes, some of which have child nodes of their own. Nodes who are not parent nodes are called leaf nodes.
- Sometimes tree diagrams are call dendrograms

Data visualization

Relationship - Tree diagrams



A simple tree diagram that shows the breakdown of regions into countries.

Part-to-whole

Pie charts

- Pie charts indicate how a whole consists of parts through slices of a pie.
- The most important rule for pie charts is that the slices must sum to 100 percent or at least some sort of total.
- Pie charts comes in different variations where annotation is also a common practice.
- Pie charts consisting of many slices can be confusing.
- The rotation of pie charts may improve their comprehensibility.

Data visualization

Part-to-whole - Pie charts

Pie charts show part-to-whole relationships. These two show the distribution of imported goods to the United States, either in dollars or percentages.

Part-to-whole

Treemap

- A treemap divides sections of a square or rectangles into groups to illustrate a hierarchy or part-to-whole relationship.
- In other words, the treemap is a squarified version of a pie chart.
- Treemap is more comprehensible than a pie chart with many slices.
- Coloring of the rectangles can add another dimension to a treemap.

Data visualization

Part-to-whole - Treemap

Color can add another dimension to a treemap. In this case, the change in imports between 1996 and 2016.