

## Data visualization

### *I Scientific and mathematical basics*

### *II Design principles*

## Basics:

One of the most important skills in today's business world is deriving insights from data and communicating them in a meaningful way. Presenting information and data using good visualizations requires two critical attributes.

- I Basic scientific and mathematical knowledge to provide the correct conclusions.
- II Design and artistic talent to present data in a vivid and compelling manner.

## Anscombe-Quartet

The Anscombe quartet consists of four sets of data points that have the same statistical properties. Each set in turn consists of eleven (x,y) points.

The Anscombe quartet is best suited to explain the most important statistical operations in a clear way.

I		II		III		IV	
x	y	x	y	x	y	x	y
10,0	8,04	10,0	9,14	10,0	7,46	8,0	6,58
8,0	6,95	8,0	8,14	8,0	6,77	8,0	5,76
13,0	7,58	13,0	8,74	13,0	12,74	8,0	7,71
9,0	8,81	9,0	8,77	9,0	7,11	8,0	8,84
11,0	8,33	11,0	9,26	11,0	7,81	8,0	8,47
14,0	9,96	14,0	8,10	14,0	8,84	8,0	7,04
6,0	7,24	6,0	6,13	6,0	6,08	8,0	5,25
4,0	4,26	4,0	3,10	4,0	5,39	19,0	12,5
12,0	10,84	12,0	9,13	12,0	8,15	8,0	5,56
7,0	4,82	7,0	7,26	7,0	6,42	8,0	7,91
5,0	5,68	5,0	4,74	5,0	5,83	8,0	6,89

If we look at the five most important and most frequently used statistical operations (arithmetic mean, variance, standard deviation, correlation coefficient and linear regression), we can see from the example of the Anscombe quartet that all four quantities lead to the same statistical results.

### Arithmetic mean

The arithmetic mean (also called the average or mean) is one of the most commonly used arithmetic operations in statistics. Using the Anscombe Quartet as an example, the arithmetic mean for all x-axes = 9 and for all y-axes = 7.5, respectively.

$$\bar{x}_{arithm} = \frac{1}{n} \sum_{i=1}^n x_i = \frac{x_1 + x_2 + \dots + x_n}{n}$$

### Variance

The variance represents the spread of values and describes the mean square deviation of the individual values from the mean or the arithmetic mean. In statistics, this is referred to as the spread around the mean. In relation to the Anscombe quartet, the variance is 11.0 on the x-axis and 4.13 on the y-axis.

$$s^2 = \frac{1}{n} \sum_i^n (x_i - \bar{x})^2$$

### Standard deviation

Unlike the variance, the standard deviation shows the average distance of all values from the arithmetic mean. In simplified terms, the standard deviation is the root of the variance.

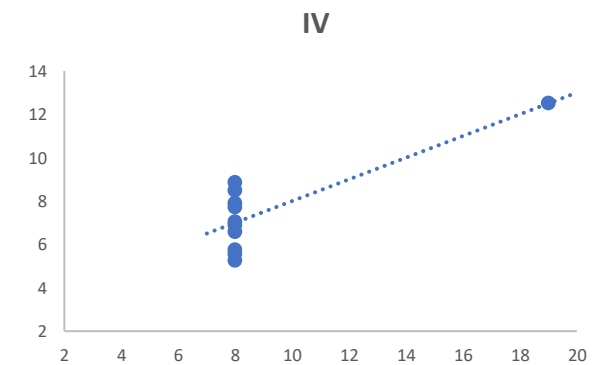
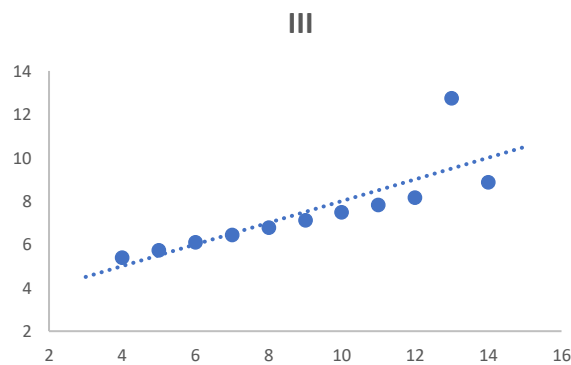
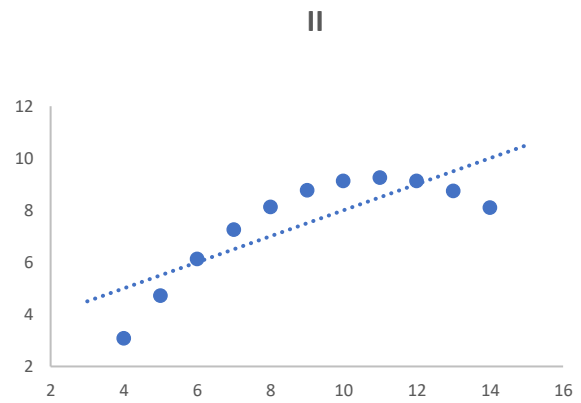
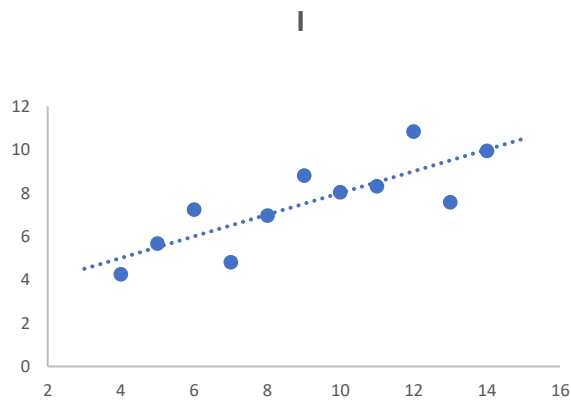
$$\sigma_x = \sqrt{Var(X)}$$

### Correlation coefficient

The correlation coefficient describes a relationship between several values. A distinction is made between a strong, moderate and weak correlation. The output value can be between -1 and +1.

$$r = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2} \cdot \sqrt{\sum_{i=1}^n (y_i - \bar{y})^2}}$$

The example of the Anscombe quartet shows the importance of the correlation coefficient. Despite very different plots, each plot has the same correlation coefficient of 0.81 (strong relationship).



## Linear regression

Linear regression describes the linear relationship between several values. By means of this procedure a trend line can be determined. Related to the Anscombe quartet, the following linear function  $y = 3 + 0.5x$  results.

$$\sum (f(x_i) - y_i)^2$$

## Diagram types

The visualization of data is highly dependent on which chart types are used for the data set. The choice of the right chart type depends on the existing file type as well as on the number of data sets that are available. Derived from this, different choices of chart type will also result in different conclusions or different statements that can be made.

## Univariat

Assuming that only one column of data is considered, in mathematics we speak of univariate, in contrast to bivariate or multivariate.

Column
#, #
#, #
#, #
#, #

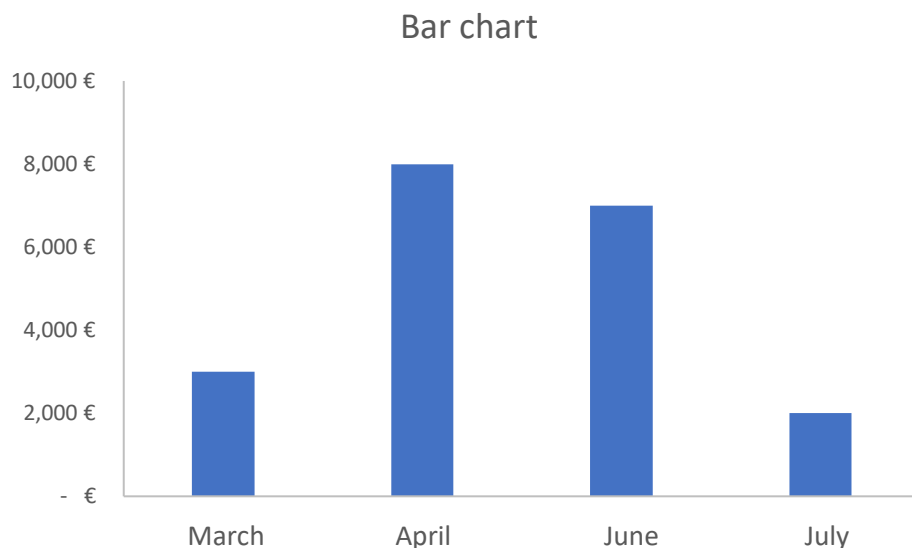
### Qualitative- und quantitative-data

Furthermore, a distinction is made between **qualitative** and **quantitative** univariate data. **Qualitative** data are data that can be grouped into categories but do not have a numerical origin. In contrast, **quantitative** data have a numerical origin. Using a dog as an example, the following qualitative and quantitative data can be collected.

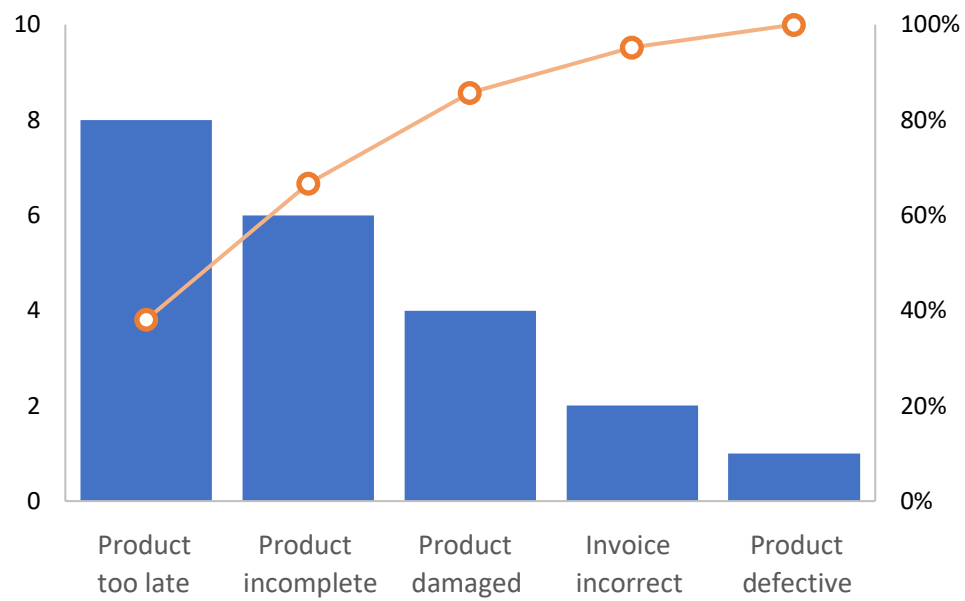
Example: dog	
Qualitative data	Quantitative data
<b>little dog</b>	<b>4 feet</b>
<b>soft fur</b>	<b>2 ears</b>

### Diagram types

For the presentation of univariate-**qualitative** data, the most common chart types are bar chart, pie chart or the Pareto chart.

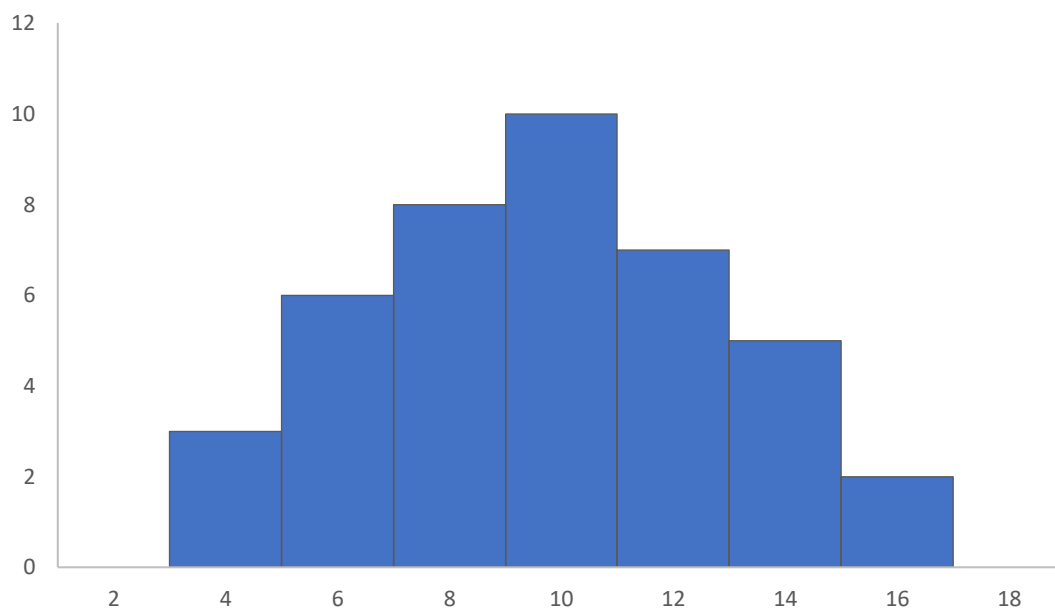


A particularly useful representation method in the area of quality management is the Pareto chart. This is very similar to the bar chart, but the bars are listed in descending order. By means of this representation, one can quickly find the type of error (20%) that has the greatest impact (80%).



For displaying univariate **quantitative** data, the chart types histogram, box chart, stem-and-leaf chart and quantile-quantile chart are suitable.

Histogram




## Scatter plot

In order to compare two quantitative variables (height and weight or sales and price), the most common way of presentation is a scatter plot. This representation can show the relationship/correlation between the measured values as well as the direction or trend.


### Correlation coefficient

As already mentioned above, the correlation coefficient describes the relationship of several values to each other. A distinction is made in the respective relationship strength of the values to each other.


A **strong correlation** exists if the coefficient is between 0.7 and 1.0 or -0.7 and -1.0.


$$0,7 \leq |r| < 1,0$$

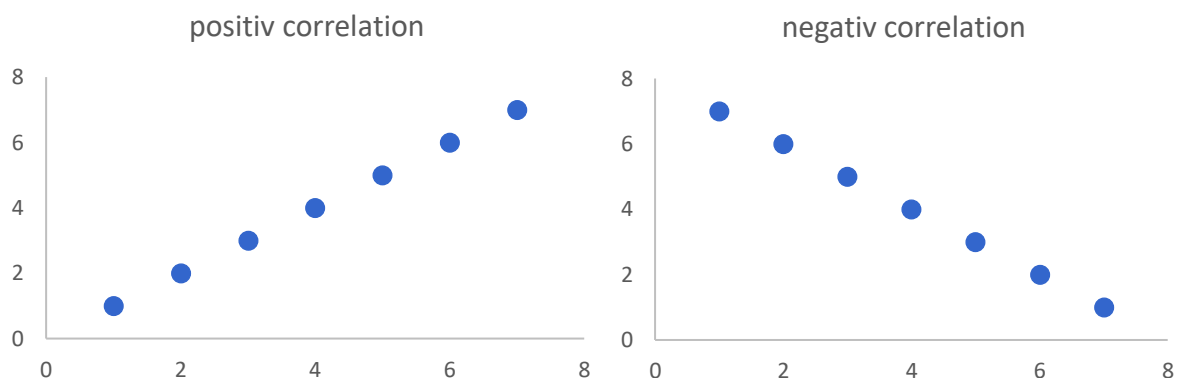
A **moderate correlation** exists if the coefficient is between 0.3 and 0.7 or -0.3 and -0.7.


$$0,3 \leq |r| < 0,7$$

A **weak correlation** exists if the coefficient is between 0.0 and 0.3 or -0.3 and -0.0.

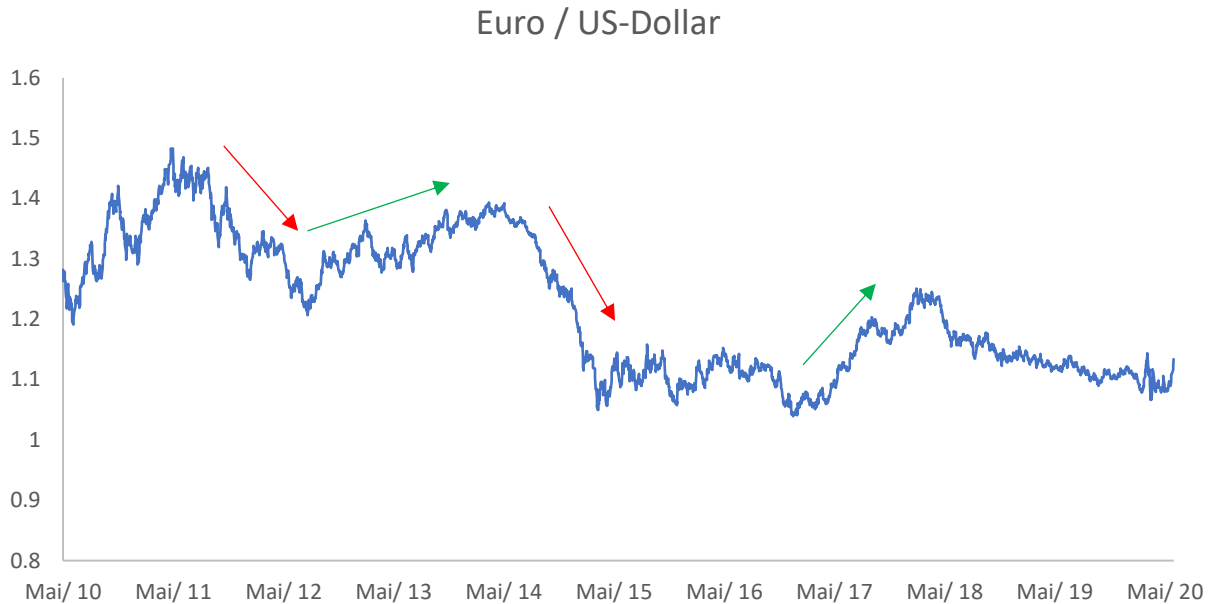

$$0,0 \leq |r| < 0,3$$

**Note:** A negative correlation coefficient does not mean that the correlation is weak but indicates an opposite relationship or a negative trend line. That is, the more of variable a, the less of variable b. The opposite is true for a positive correlation.

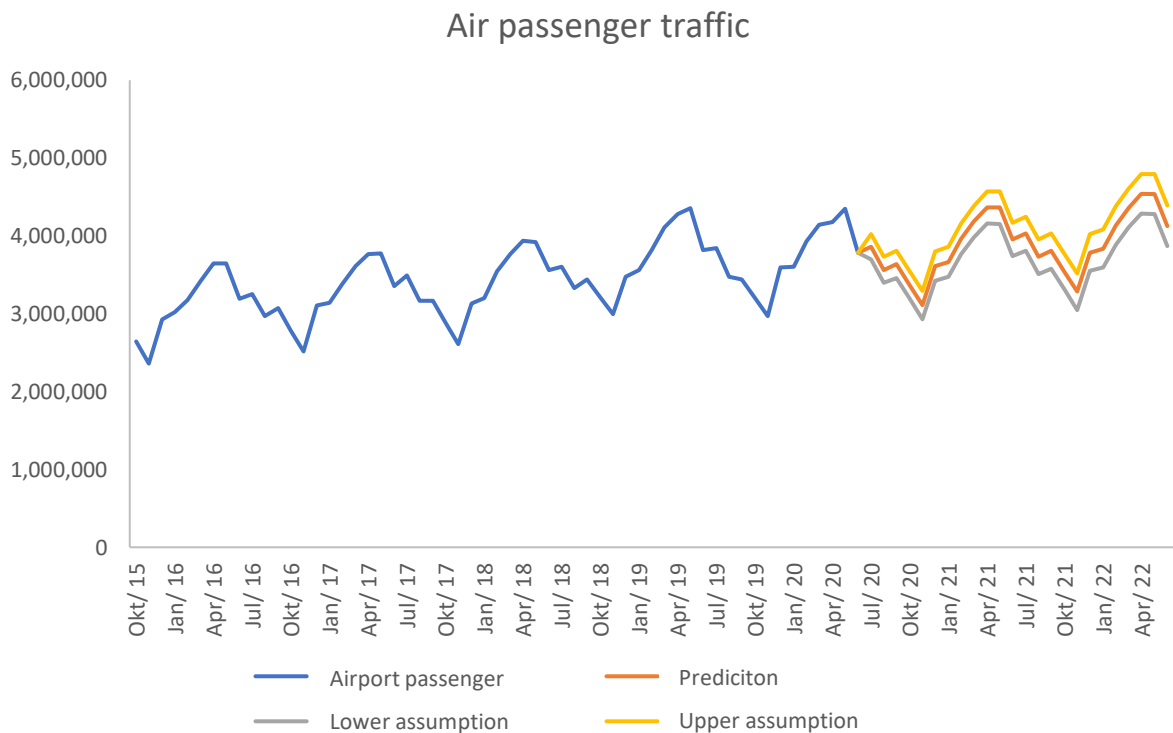


## Line chart

If a quantitative variable is present over a longer period of time, which is associated with a date, then it is convenient to use a line chart. This representation is very widespread and offers a good possibility to show trends. One can immediately see at a glance whether the trend is increasing or decreasing.



The line chart can also be used to identify seasonal effects or patterns and derive predictions based on them.

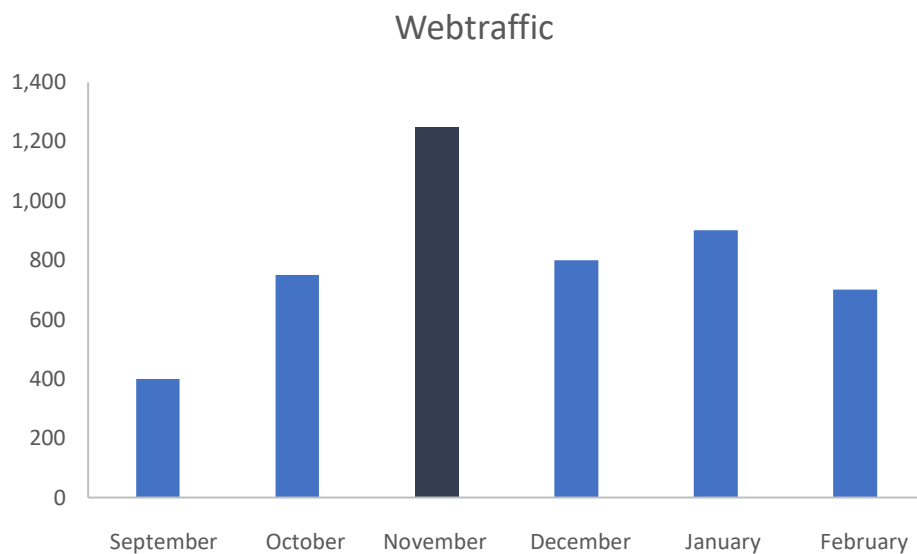


### Significance (Which question should be answered?)

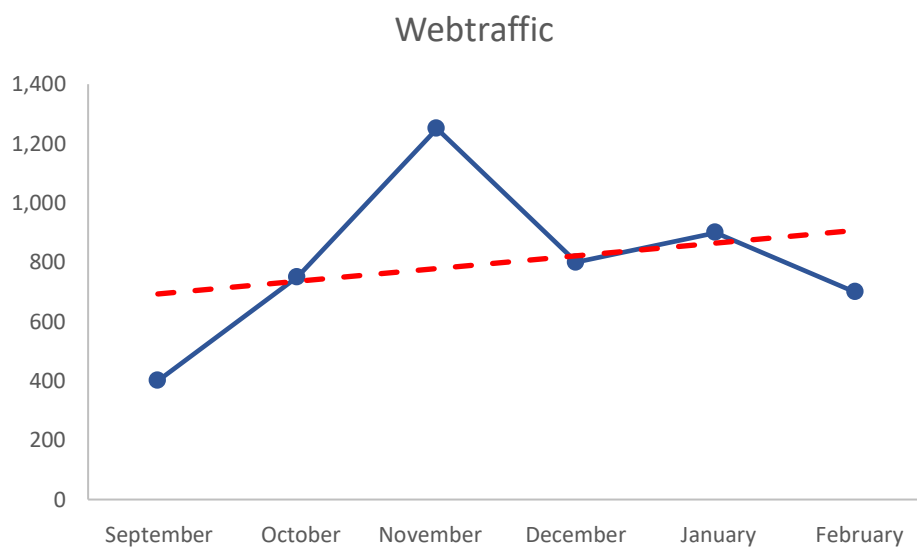
The right choice of data visualization strongly depends on the question to be asked or answered. Let's say you want to analyze the web traffic from a website and the following questions are asked.

1. in which month did the most web visits take place?
2. how often did one have more than 2,000 visitors per month?
3. how often did you have a certain number of visitors?

To answer the first question quickly and easily, it is helpful to create a bar chart or Pareto chart.



However, in order to read off the trend quickly, a line chart would be more helpful. With the help of a trend line, the overall course can be made visible.

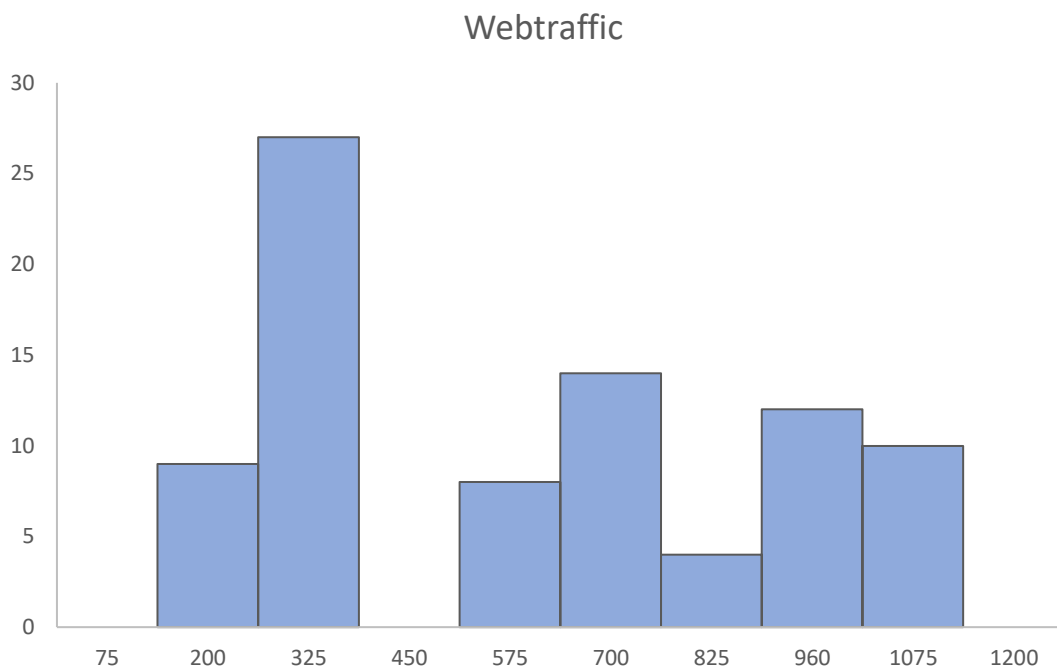




But to answer the second and third questions, a histogram is needed that acts independently of temporal aspects. On the x-axis, the web calls are listed in sections of 125 calls each. On the y-axis the number is shown, how often this number of visitors occurred per month. By means of this representation, one can quickly draw the following conclusions.

1. there were always at least 200 or at most 1,075 visitors per month on the website.
2. 450 visitors per month were never reached.
3. 325 visitors per month were most frequently on the website.

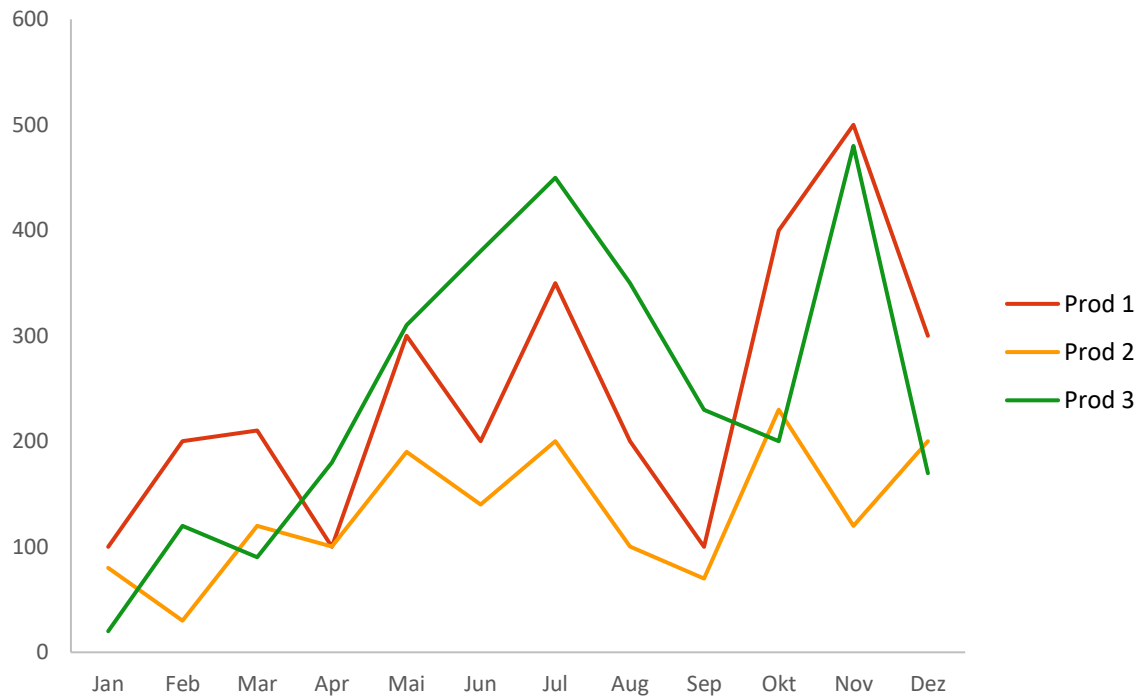
Only by means of a histogram can these questions be answered quickly. This would not have been possible so quickly and easily with a time bar chart or line chart.



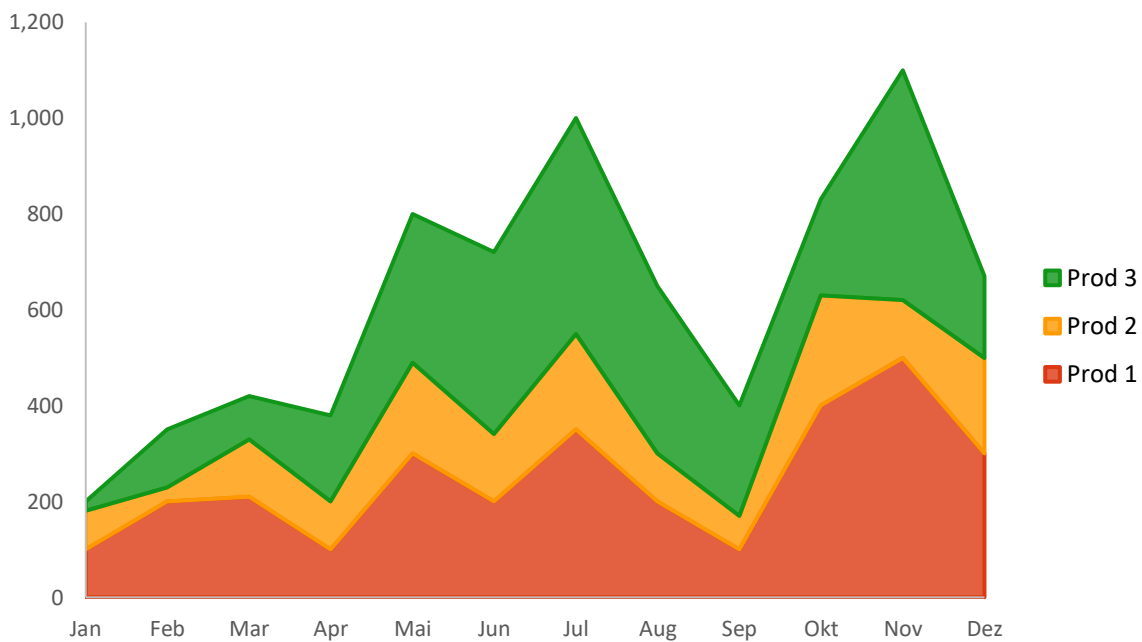
In conclusion, the same data basis but with the help of different representations allows to answer different questions faster and easier.

## Representation of multiple variables

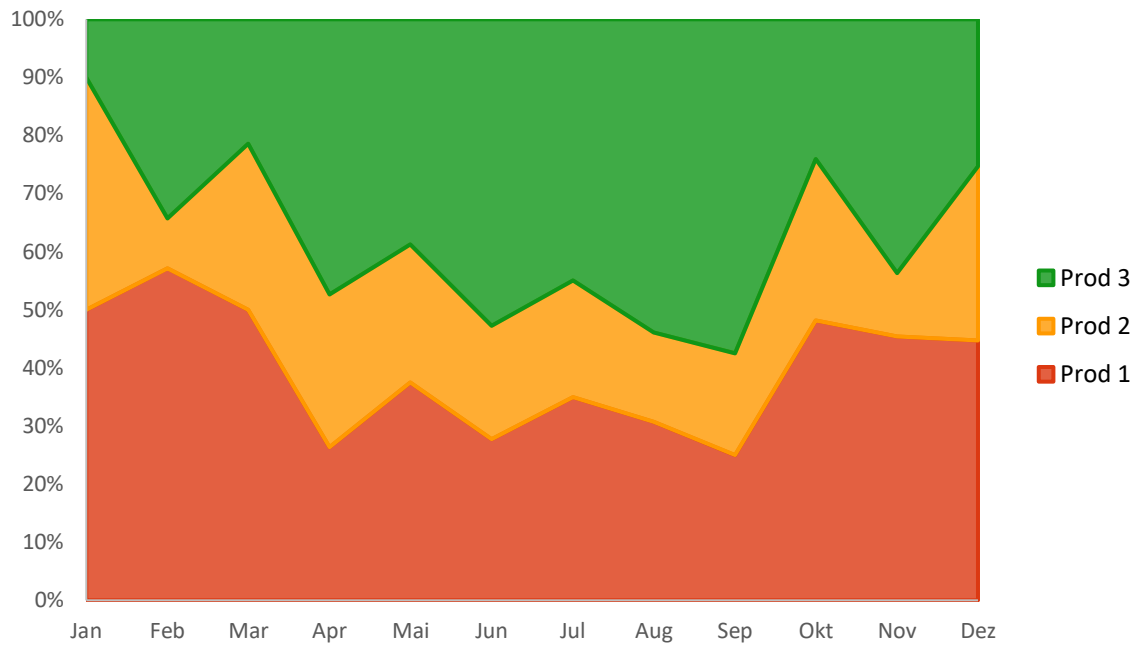
To compare two quantitative variables, you need more than just the x- and y-axis. Therefore, it is helpful to color the variables to make the distinction clearer. This also requires a legend to distinguish the individual values. Again, different chart types answer different questions.



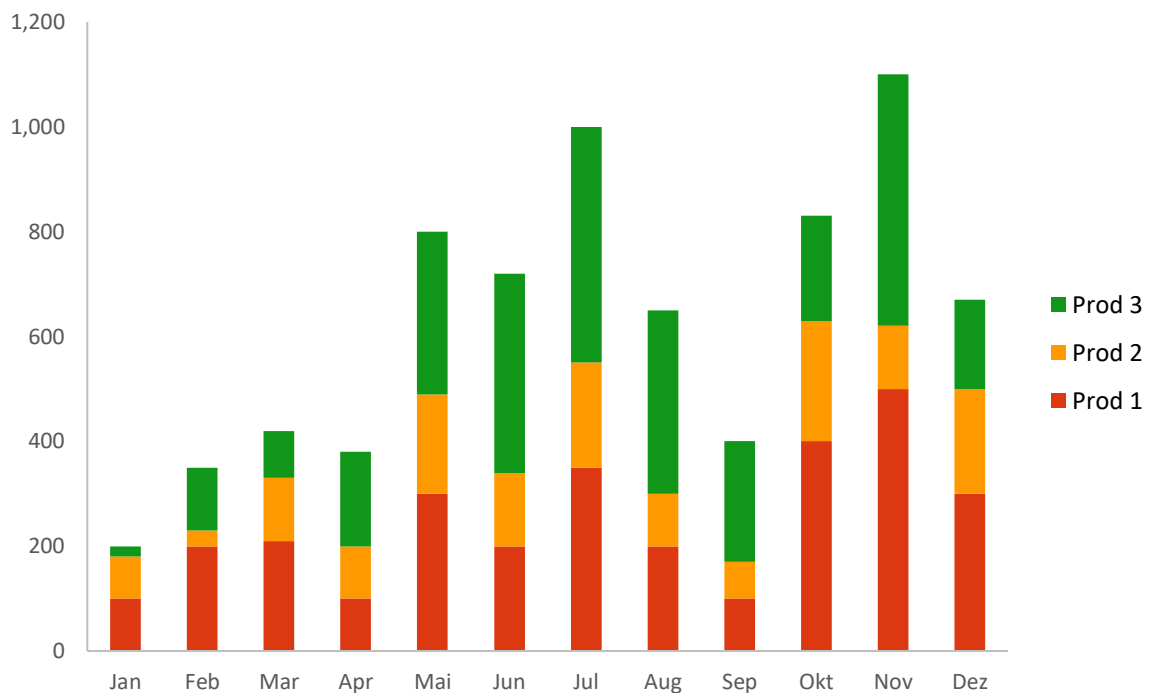
A stacked area chart contains the same information, yet it is easier to see the respective share of total sales in this chart.



To quickly and easily see the exact percentage distribution to 100%, it is useful to use an area chart that converts the y-axis to percent. Here, too, the same data basis is available but a different question is answered.



The stacked bar chart conveys the same message as the stacked area chart, but the speed of interpretation of the representation is always dependent on the particular person viewing the visualization.



## Dashboards

Assuming you want to display the aspects of sales, locations, products and the temporal development in a diagram, you quickly reach the limits of the possibilities that can be displayed in a diagram. In this case it is useful to transfer all this information into a dashboard in order to highlight the respective

questions or aspects. With the help of a dashboard and the data linked in the background, it is easy and fast to adapt the visualization to the respective question.

## Summary

In order to derive conclusions from data, it is necessary to visualize data. This requires that one knows the data type as well as the diagram types and knows how to apply them. Additionally, it is helpful to know the mathematical and statistical basics to support these conclusions.

## Outlook – Design principles

Visual Encoding describes the design principles of variables in diagrams. There are some important questions to answer in order to best convey the findings.

- Which variable is set on the x- or y-axis?
- Which colors are to be used?
- Should different shapes be used?