

Geo Data Science

Exploratory Data Analysis

Prof. Dr. Martin Kada

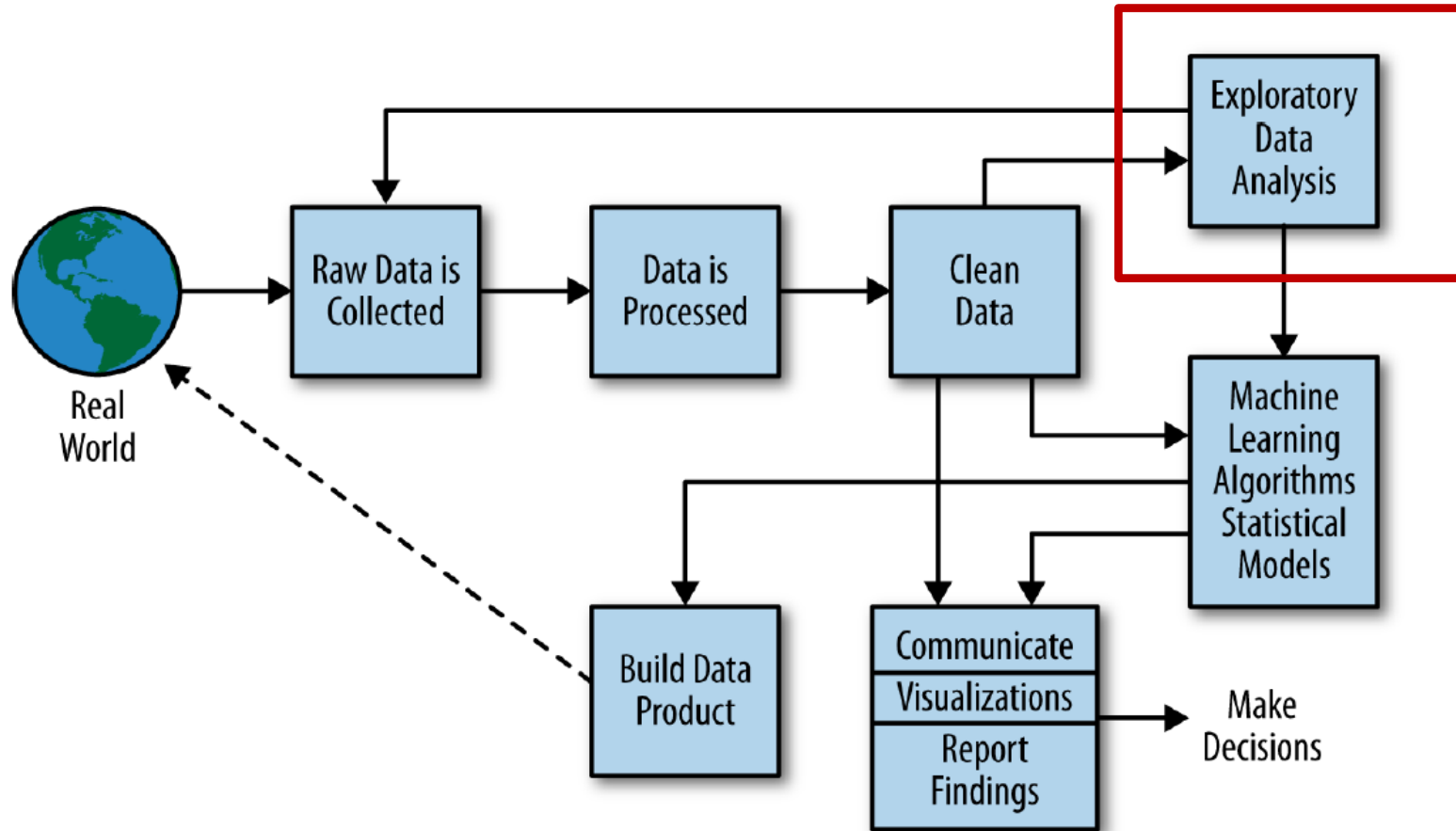
Chair Methods of Geoinformation Science (GIS)
Institute of Geodesy and Geoinformation Science

Copyright Notice

The teaching materials for this course and all elements contained therein are protected by international copyright laws. They may only be used for study purposes for the corresponding course.

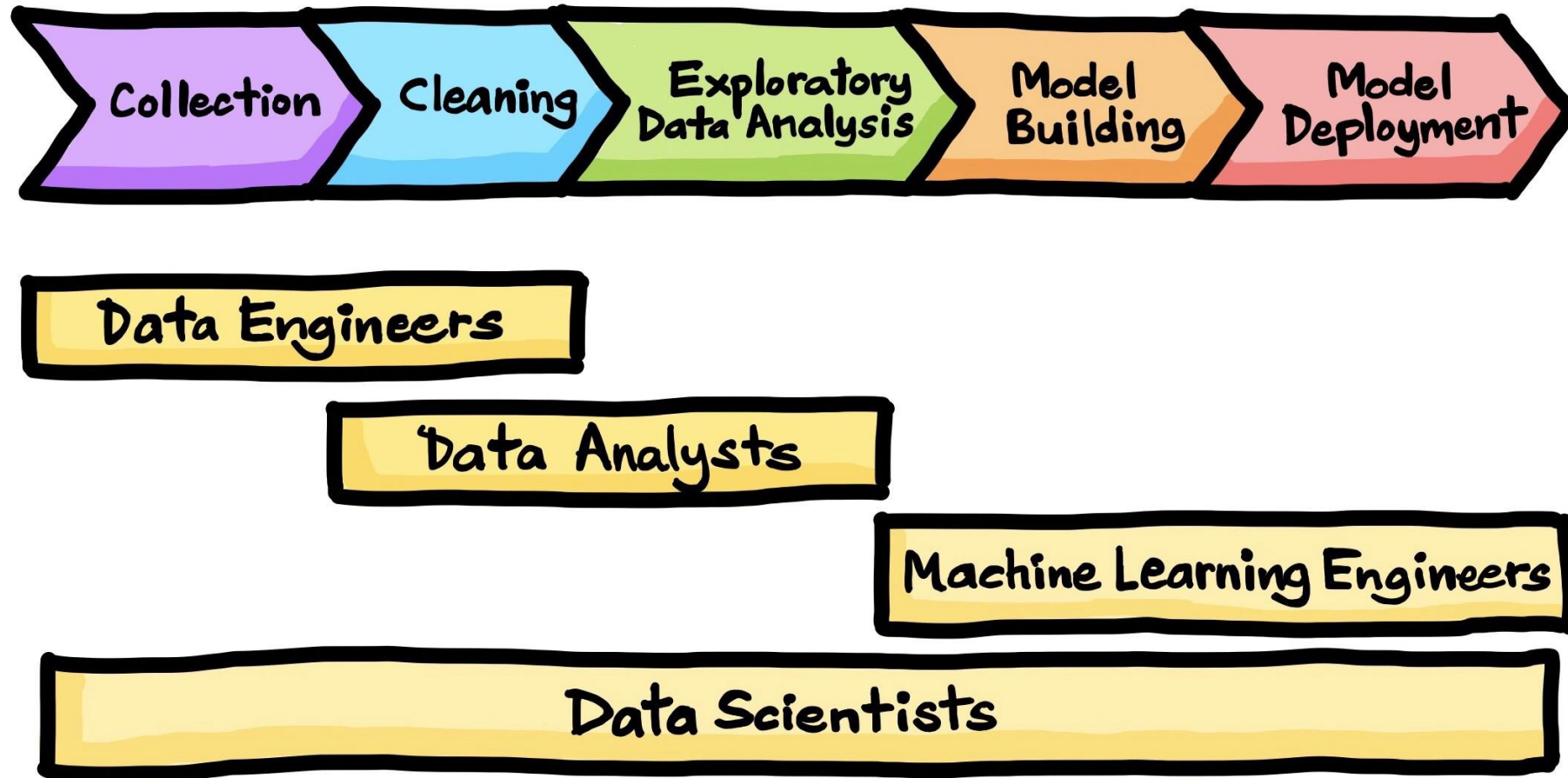
Any reproduction and redistribution of the course materials without written permission is prohibited, other than the following: You may print or download them for your own personal use while attending the course.

The Data Science Process



Source: Schutt & O'Neil, 2014

The Data Science Process



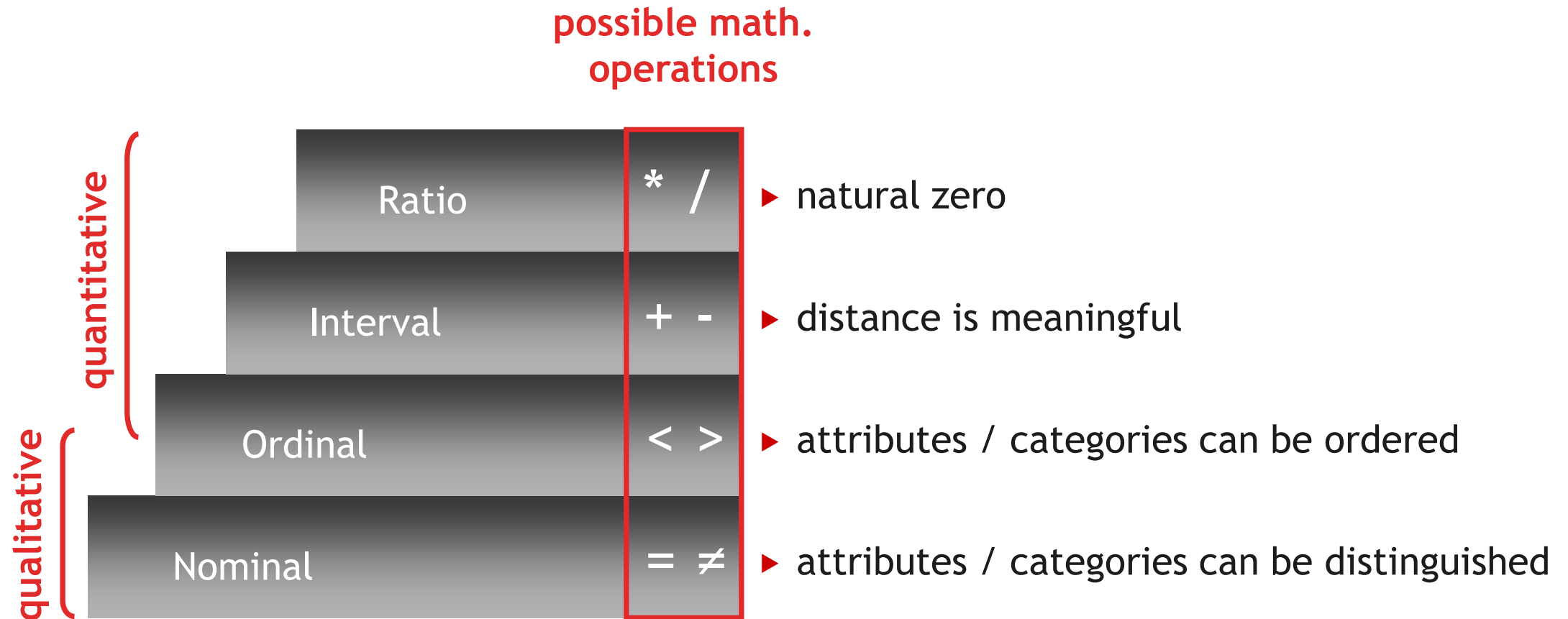
- “Exploratory data analysis is an attitude, a state of flexibility, a willingness to look for those things that we believe are not there, as well as those we believe to be there”
(John Tukey, founder of Exploratory Data Analysis)
- The first step towards a data science project / building a machine learning model
- This critical part in the data science process, often involves more than 70% of the time invested in the project and includes:
 - **Data manipulation:**
 - Solve various origin and scale problems in space and time
 - Data filtering, cleaning, outlier detection
 - Data transformations (reshaping)
 - Import/export to common formats

- **Data plotting:**
 - Create different plots of the data for visualizations
 - Explore spatial data using (carto-)graphic (or other visual) representations
- **Descriptive statistics:**
 - Summarizes the main properties of a set of values
 - Main measures are those of central tendency (location) and spread (variability)
 - It can also quantify the (linear) dependency of two related data sets, e.g., temperature and solar radiation at the same time in one location

Ways to Analyze Data













- **Univariate:**
 - Each variable is analyzed separately:
data distribution, central value, and data spread / uncertainty
- **Bivariate:**
 - Two variables are analyzed together to look for correlation or separation of data
- **Multivariate:**
 - More than two variables are analyzed together
 - Generally difficult to visualize the data and results

Types of Variables



Nominal Data (=, ≠)

- Classification according to type or quality
- Often labeled with numbers or letters, but no ranking implied!

Point	airport 	town 	mine 	capital 
Line	river 	road 	boundary 	pipeline 
Area	orchard 	desert 	forest 	water 

Representation of nominal data

<http://www.geog.okstate.edu/users/Larson/home.htm>

Ordinal Data (<, >)

- Information about rank or hierarchy
- Possible to describe one item as smaller or larger than another
- Not possible to measure differences, because there are no specific values attached

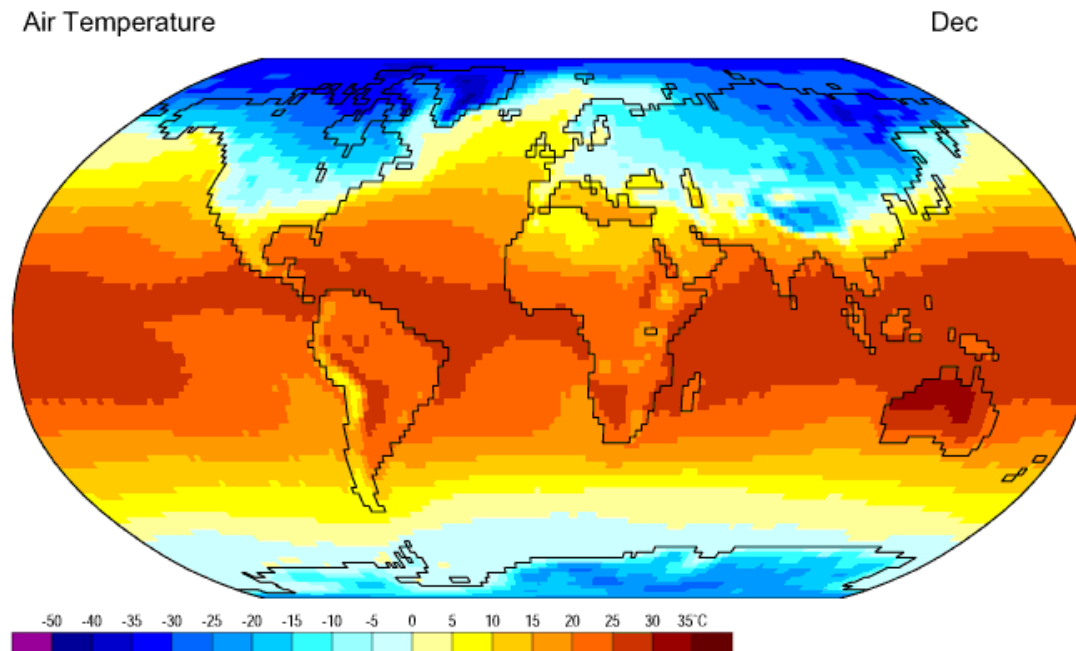
Point	Airports ✂ international ✂ national ✂ regional	Oil well production ■ high ■ medium ■ low	Populated places ● large ● medium ● small
	Roads expressway major local	Drainage river stream creek	Boundaries international provincial county
	Soil quality ■ good ■ fair ■ poor	Cost of living ■ high ■ medium ■ low	Industrial regions ■ major ■ minor

Representation of ordinal data

<http://www.geog.okstate.edu/users/Larson/home.htm>

Interval Data (+, -)

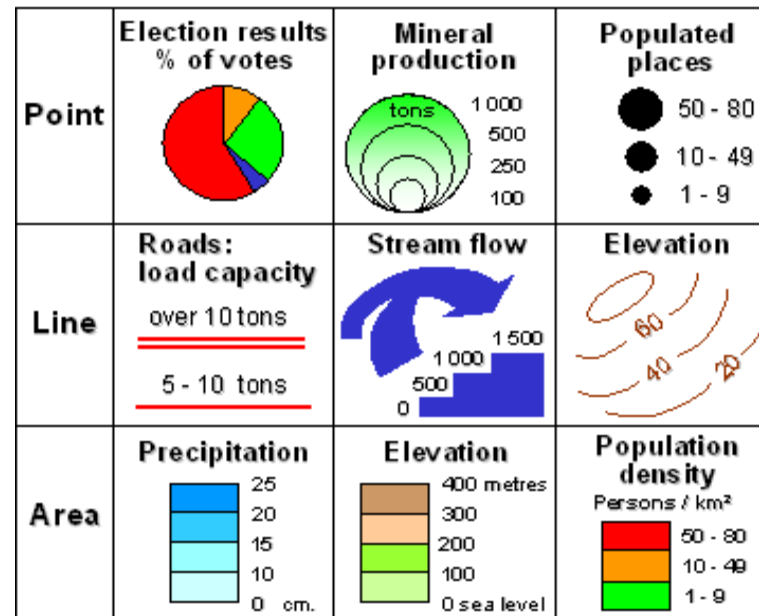
- Include numerical values
- Information can be arranged along a scale → distance/difference can be calculated



Data: NCEP/NCAR Reanalysis Project, 1959-1997 Climatologies
Animation: Department of Geography, University of Oregon, March 2000

Ratio Data (*, /)

- Like interval data, but there is a natural zero → data can be expressed as ratios



Representation of interval and ratio data

<http://www.geog.okstate.edu/users/Larson/home.htm>

Types of Variables

- **Categorical data (qualitative data):**
 - **Nominal:** can be named, e.g. soil types
 - Can only be separated, but not (uniquely) ranked (no intrinsic order): colors, names
 - Can be coded as numbers (0, 1, 2, ...), but many numerical operations do not make sense
 - **Ordinal:** can be ordered (and named), e.g. seismic scale, grades, sizes
- **Numerical data (quantitative data):**
 - **Interval:** can be subtracted (and ordered and named) → difference, e.g., integers in equally spaced intervals
 - **Ratio:** can be divided (and subtracted, ordered and named), e.g. amount of money you have in your pocket right now
 - the most informative scale

Categorization by Dimension

- One-Dimensional (18, 20, 43, 32, ...)
- Multi-dimensional ((1, 4), (5, 3), (6.2, 10.4), ...)
- High-dimensional (e.g. Time Series Data)
- No-dimensional (e.g. Protein Folding Structures)

- Measures of **Central Tendency**:
 - Estimating the **location** of a typical value
 - mean, median, mode, ...
- Measures of **Spread**:
 - Estimating the **variability** within the data
 - mean absolute deviation, variance, standard deviation, ...

Measures of Central Tendency

- **Mean:**

- Arithmetic average of the values in a data set

$$\text{mean} = \bar{x} = \frac{\sum_{i=1}^n x_i}{n}$$

- Relevant for interval / ratio data, but disputable for ordinal
- Expressed in the units of the data

Measures of Central Tendency

- **Trimmed mean:**

- Arithmetic average of the values in a data set, for which a fixed number of the lowest and the highest p values are dropped beforehand

$$\text{trimmed mean} = \bar{x} = \frac{\sum_{i=p+1}^{n-p} x_{(i)}}{n - 2p}$$

sorted values



- Eliminates the influence of extreme values

Measures of Central Tendency

- **Weighted mean:**

- Summation of the multiplication of each data value x_i with a (user-specified) weight w_i divided by the sum of the weights

$$\text{weighted mean} = \bar{x}_w = \frac{\sum_{i=1}^n w_i x_i}{\sum_{i=1}^n w_i}$$

- Highly variable observations can be given a lower weight
- Give more weights to underrepresented categories

Measures of Central Tendency

- **Mid-range:**

- Arithmetic average of the highest and the lowest value in the data set

$$\text{mid-range} = \frac{\max(x_i) + \min(x_i)}{2}$$

- **Median:**

- Value that divides the data set into two equally sized halves
- Also defined as the 50th percentile: $x_{0.5}$
- Relevant for interval / ratio and ordinal data

- **Mode:**


- Most frequently occurring value in a data set
- Relevant for data for which frequencies makes sense
- Robust estimator (not sensitive to extreme values)

Measures of Spread

- **Mean absolute deviation:**
 - Average of absolute deviations to the central value

$$\text{mean absolute deviation} = \frac{\sum_{i=1}^n |x_i - \bar{x}|}{n}$$

central value
(e.g. mean,
median, mode)




Measures of Spread

- **Variance:**

- Average of squared deviations to the mean

$$\text{variance} = s^2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n - 1}$$

mean value



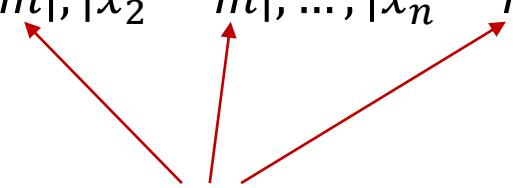
- **Standard deviation:**

- Square root of the variance

$$\text{standard deviation} = s = \sqrt{\text{variance}}$$

Measures of Spread

- Mean absolute deviation, variance, and standard deviation are not robust to outliers
- Median absolute deviation from the median (MAD):

$$\text{MAD} = \text{median}(|x_1 - m|, |x_2 - m|, \dots, |x_n - m|)$$


median value

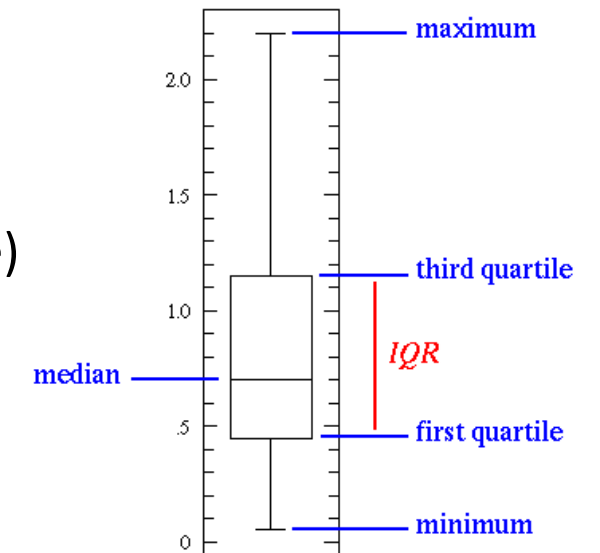
Measures of Spread

- Statistics based on sorted (ranked) data:
- Range:
 - Difference between the highest and the lowest value in the data set ($x_{\max} - x_{\min}$)

- **Interquartile range (IQR):**
 - The P-th percentile is the value such that at least P percent of the values in the data set are equal or lower than the percentile, and at least 100-P percent of the values are equal or higher than the percentile
 - Quantiles are essentially the same as percentiles, but indexed by fractions ($x_{0.9}$ is equal to 90th percentile)
 - Interquartile range (IQR) is the difference between upper and lower quartiles ($x_{0.75} - x_{0.25}$)
 - Summary statistics based on quantiles are robust to outliers

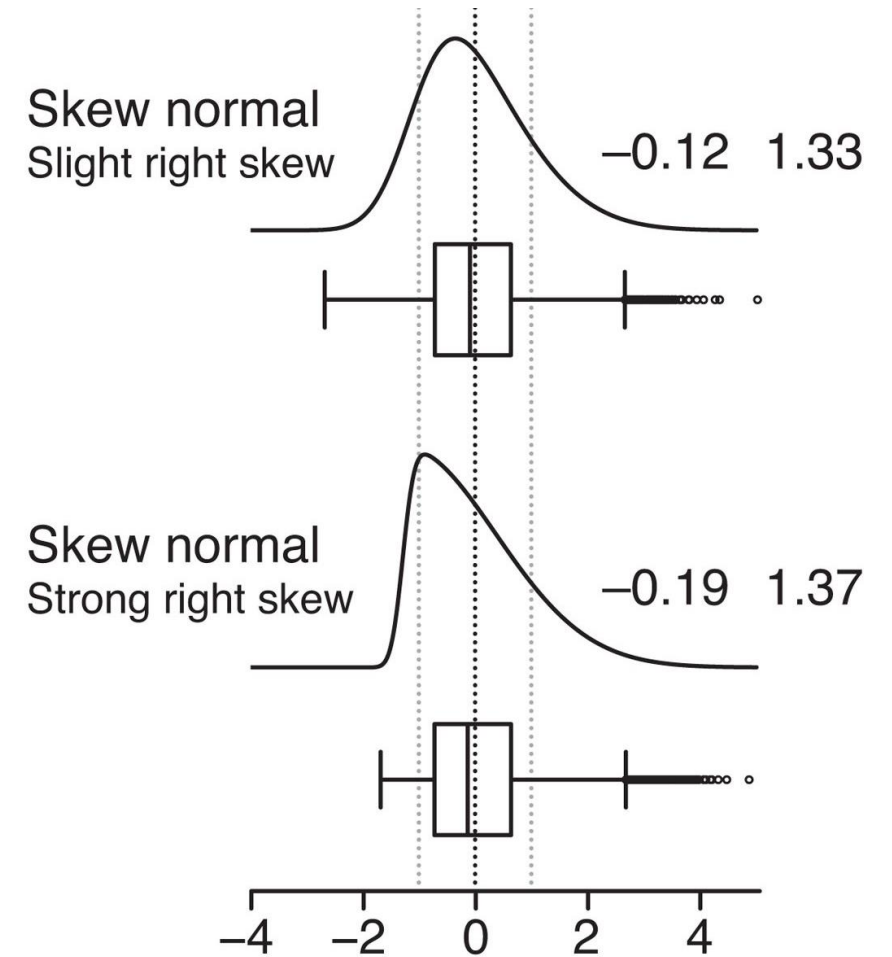
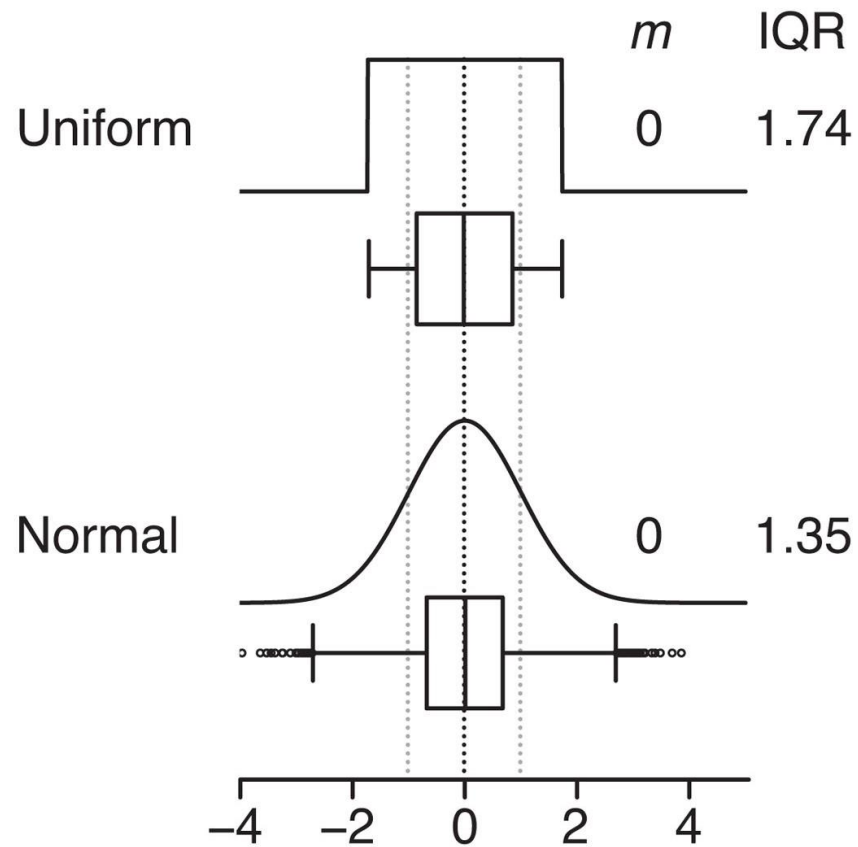
Boxplots

- Plot data points for a first overview of the data
- **Boxplot:** shows characteristic values of the data:
 - Box covers the central half of the values (1st quartile to 3rd quartile)
 - Thick line is the median
 - Whiskers reach out to the minimum and maximum, unless these are very extreme, then they are shown as outliers



- Delete outliers to get more robust results, but you may lose important information

Boxplots



Frequency Tables and Histograms

- Consider the 10 hypothetical sample values:

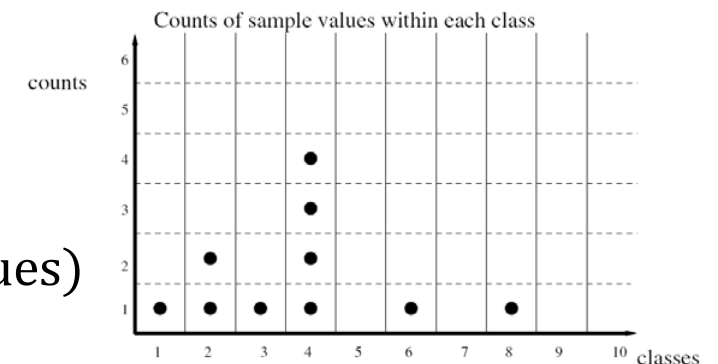
4	1	3	8	4	4	2	4	6	2
---	---	---	---	---	---	---	---	---	---

- Estimated relative frequency table:

$$\hat{f}_k = (\# \text{ of data in } k\text{-th class}) / (\text{total } \# \text{ of data})$$

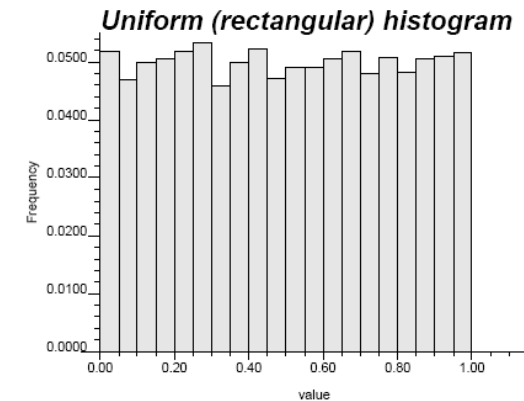
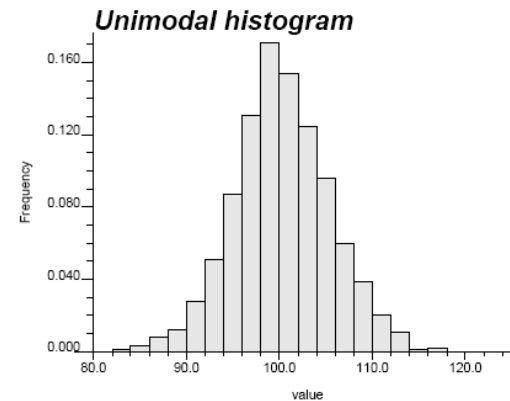
x_k	1	2	3	4	5	6	7	8	9
\hat{f}_k	0.1	0.2	0.1	0.4	0.0	0.1	0.0	0.1	0.0

- Histogram shape depends on number and width of bins (“classification”):
 - Use non-overlapping equal intervals with simple bounds
 - Rule of thumb for number of classes: $5 \times \log_{10}(\# \text{ of data values})$
 - For a density histogram, total area of bars = 1

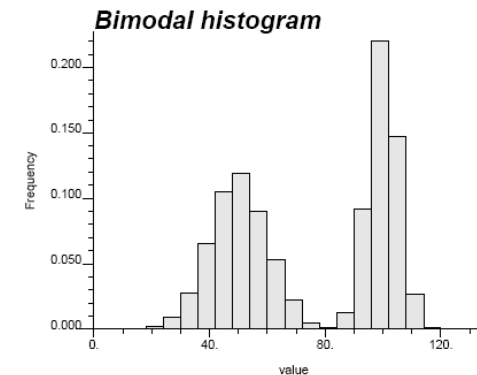
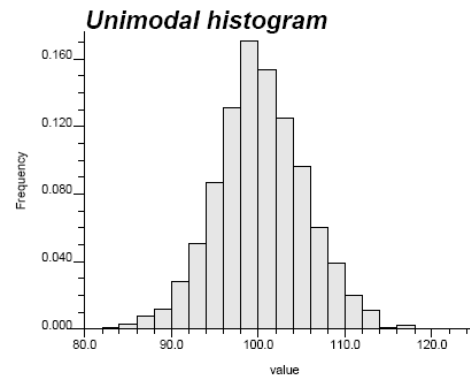


Histogram Shape Characteristics

- Peaked or uniform:
 - Is it peaked or not?



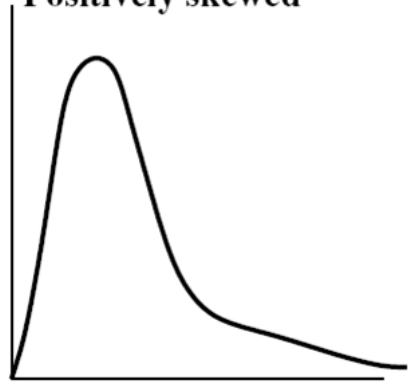
- Number of peaks:
 - How many peaks are there?



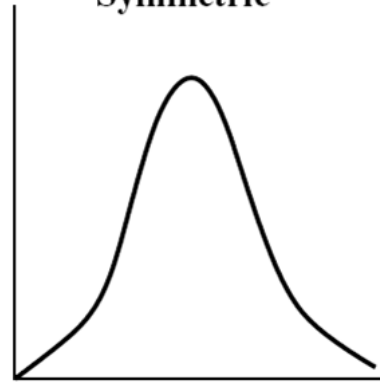
Histogram Shape Characteristics

- Symmetric or skewed:

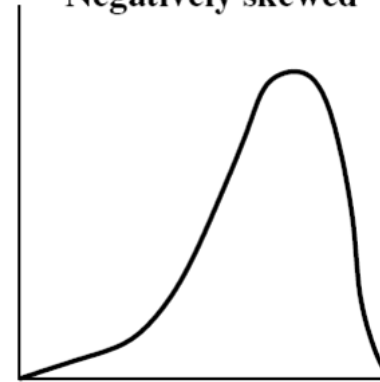
Positively skewed



Symmetric



Negatively skewed



$$\text{coefficient of skewness} = \frac{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^3}{s^3}$$

Univariate vs. Bivariate Analysis

- **Univariate statistics**
 - Describe the distributions of **individual** variables (or one variable at a time in a set of multivariate data)
 - Is not sufficient to describe spatial patterns, because the spatial arrangement of attribute values matters
- So ...
 - Relationships and dependencies between variables are very important in most (earth) science data sets
 - For **comparing the distributions** of paired data (we have two measurements per observation, e.g., porosity and permeability)
 - Histograms + summary statistics → reveals only gross differences
 - Two very similar distributions → not helpful
 - ⇒ suitable visual comparisons + **bivariate analysis**

Scatterplots

- Offer qualitative information on how **two variables are related**
- Useful also for error checking:
make aberrant data obvious

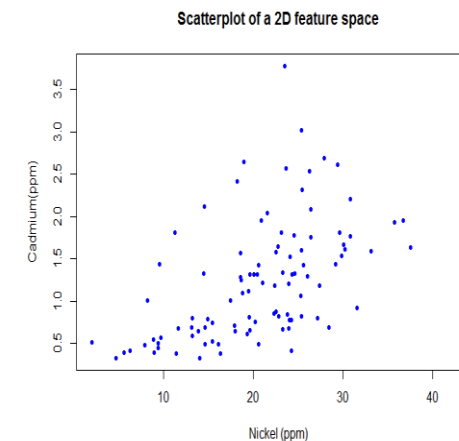
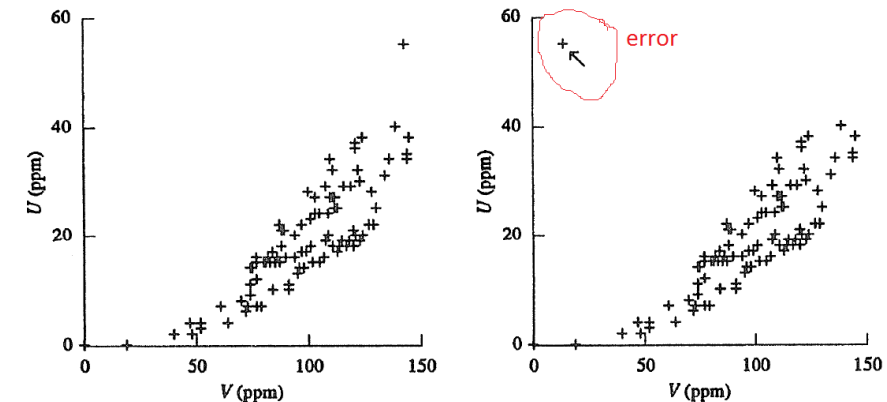
- Setting:

- Data pairs of two variables X & Y,
measured at N sampling units
- There are N pairs of attribute values

$$\{(x_n, y_n), n = 1, \dots, N\}$$

- **Scatterplot:**

- graph of x-values versus y-values in the **bivariate attribute space**:
 - x-values as coordinates in horizontal axis
 - y-values as coordinates in vertical axis
 - n-th point in scatterplot has coordinates (x_n, y_n)



- Key feature in a scatterplot → correlation (association or trend) between variables U and V
- There are three patterns that can be observed on a scatterplot:
 - **positive correlation**
 - Higher (lower) X values are associated with higher (lower) Y values
 - E.g.: porous rocks → porosity and permeability
 - **negative correlation**
 - Higher (lower) X values are associated with lower (higher) Y values
 - E.g.: geological data sets → concentration of major elements is often negatively correlated (dolomitic limestone – increase in amount of calcium results in a decrease of magnesium)
 - **no correlation (uncorrelated)**
 - An increase in one variable has no apparent effect on the other

Correlation Coefficient

- Most frequently used to summarize the relationship between two attributes
- Quantifies numerically the trend in a bivariate scatterplot
- Provides a measure of the linear relationship of two variables
 - If there is no linear relationship \Rightarrow poor summary statistic

$$\rho = \frac{\frac{1}{n} \sum_{i=1}^n (x_i - m_x)(y_i - m_y)}{\sigma_x \sigma_y}$$

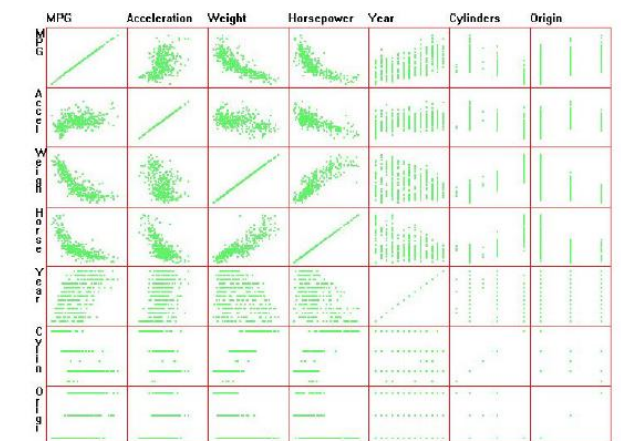
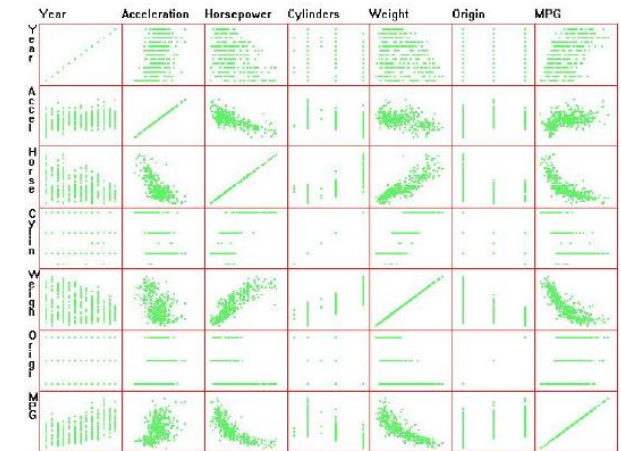
Diagram annotations:

- Red arrows point from the word "mean" to m_x and m_y in the numerator.
- The entire numerator is highlighted in a light blue box, with the word "covariance" written in blue to its right.
- A green arrow points from the words "standard deviations" to $\sigma_x \sigma_y$ in the denominator.

- Average of the data deviations from their means
- Depends on the magnitude of the data values
- Division by standard deviation \rightarrow ρ values are between -1 and 1
- Strongly influenced by a few aberrant pairs

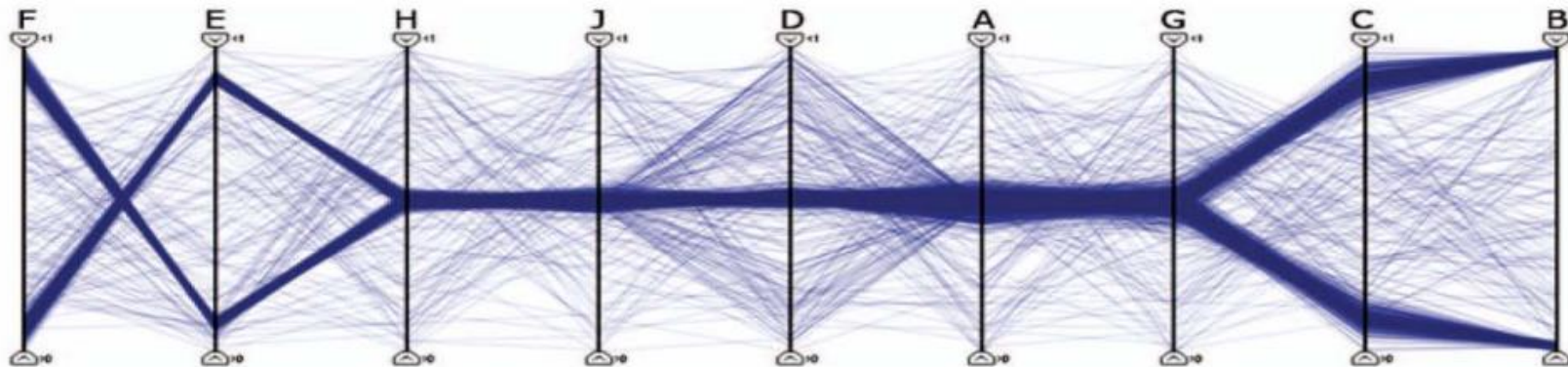
Scatterplot Matrix

- Matrix of scatterplots for all pairs of variables
- Ordering of dimensions (variables) is important
- Dimension re-ordering
 - The interestingness of different orderings can be evaluated with quality metrics
 - Reduces clutter
 - Better visualization and understanding of data



Parallel Coordinates

- A d -dimensional data space is visualized by d parallel axes
- Each axis is scaled to the min-max range in the corresponding dimension
- A data point is visualized as a **polyline**, which intersects each of the axes at the point that corresponds to the value of the object in the respective dimension

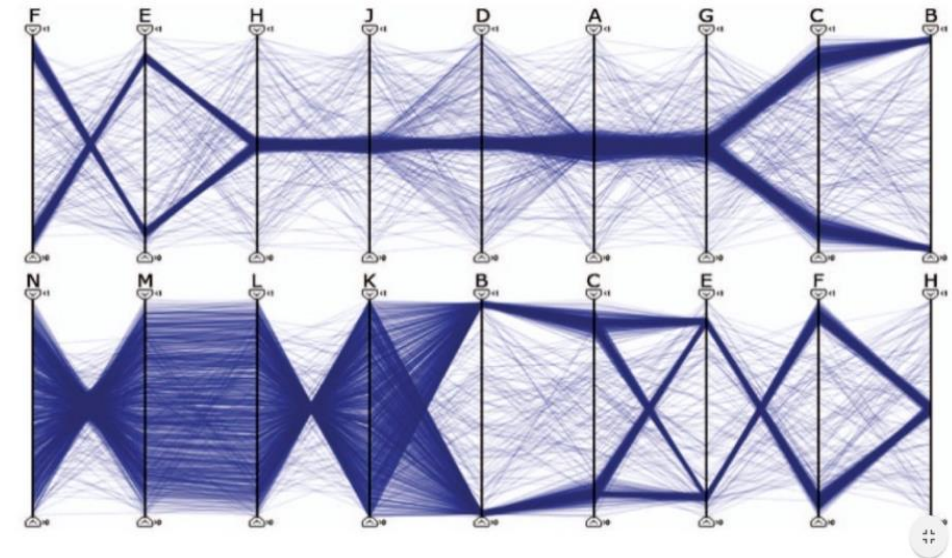


Parallel Coordinates

- The ordering of the dimensions matters!
- Interestingness of an ordering can be measured with a quality metric
- Quality or interestingness of orderings depends on what you want to visualize

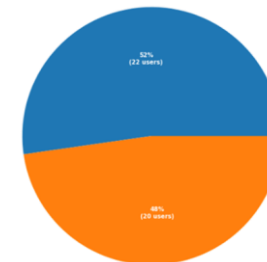
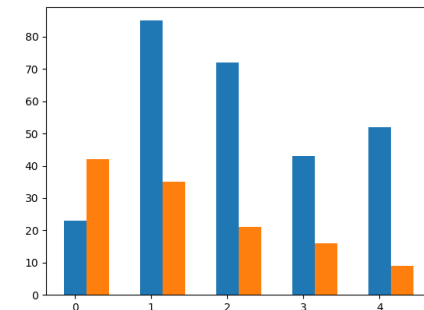
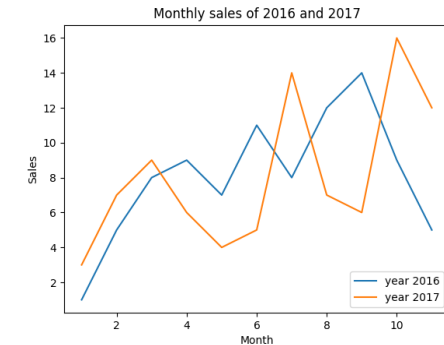
- Example:

- The first ordering is well-suited to visualize clusters in the data
- The second ordering is well-suited to visualize correlation between the dimensions



Other Plots

- **Line plots:**
 - Display information as a series of (ordered) data points connected by straight lines
 - Visualize trends in data over intervals (of time)
- **Bar charts:**
 - Categorical data with rectangular bars, where the height corresponds to the value it represents
- **Pie charts:**
 - Circular plots divided into slices to show numerical proportions



Thank you for your attention!