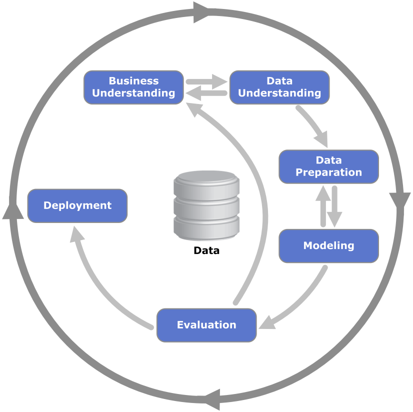
Data Understanding

Rita P. Ribeiro

Machine Learning - 2022/2023

From previous class . . .



Shearer C.: The CRISP-DM model: the new blueprint for data mining, J Data Warehousing (2000)

Today

References

- Aggarwal, Charu C. 2015.Data Mining, the Texbook. Ch 1.1, 1.2, 1.3.
- Moreira, João, et al. 2018. Data Analytics: A General Introduction. Ch 2, Ch 3
- Gama, João, et al. 2015. Data Mining -3rd Ed. Ch 2.
- Wilke, Claus O. 2022. Fundamentals of Data Visualization.

Today

- Data Understanding
  - Data
  - Summarization
  - Visualization

Data

Data

Collection of data objects (cases) described by attributes (features)

- Attribute: a property or characteristic of an object
    - date, country, temperature, precipitation
- Object: described by a collection of attributes

- It can be structured (e.g. data table) or non-structured (e.g. text)
- It can have non-dependency or dependency between objects (e.g. time, space)

Data

Examples of data sets

- Data tables
  - tabular data, document data, transactional data
- Ordered data
  - time series, data streams, genetic sequences
- Graphs and networks
  - social networks, transportation networks, molecular structures
- Multimedia
  - images, audio, maps, video

Types of data sets

- Nondependency-oriented data
  - the cases do not have any dependencies between them
  - examples: simple data tables, transactions
- Dependency-oriented data
  - implicit or explicit relationships between cases
  - examples: time series, discrete sequences, spatialtemporal data, network and graph data.

Data

- A tidy data table with 15 cases described by 4 attributes.

country	year	sex	age	cases
AD	2000	m	0-14	0
AD	2000	m	15-24	0
AD	2000	m	25-34	1
AD	2000	m	35-44	0
AD	2000	m	45-54	0
AD	2000	m	55-64	0
AD	2000	m	65+	0
AE	2000	m	0-14	2
AE	2000	m	15-24	4
AE	2000	m	25-34	4
AE	2000	m	35-44	6
AE	2000	m	45-54	5
AE	2000	m	55-64	12
AE	2000	m	65+	10
AE	2000	f	0-14	3

Data: Attributes

- Type of Attributes
  - Categorical
  - Numeric
- Scale of Attributes
  - Nominal
  - Ordinal
  - Interval
  - Ratio

Data: Type of Attributes

Categorical Attributes

- finite number of symbols or names
- if represented by numbers, they don't represent quantities
- no arithmetic operation can be performed on them
- e.g.eye color, t-shirt size

Data: Type of Attributes

Numeric Attributes

- Discrete
  - finite or countably infinite set of values
  - it can take only distinct or separate values
  - e.g. number of students in a class
- Continuous
  - infinite set of values, real numbers
  - measurable data
  - e.g. distance, income

Scale of Categorical Attributes

- Nominal
  - there is no relationship between the values
  - only equality is meaningful
  - e.g. eye color
- Ordinal
  - there is an order between the values
  - both equality and inequality is meaningful
  - e.g. size $\in \{small, medium, large\}$

Scale of Numeric Attributes

- Interval
  - values vary within an interval
  - equality, inequality and differences are meaningful
  - the value 0 or scale origin, is defined arbitrarily
  - there is no absolute zero
  - e.g. calendar year, temperature ($^\circ C$)

Data: Scale of Attributes

Scale of Numeric Attributes

- Ratio
  - numbers have an absolute meaning
  - equality, inequality, differences and ratios are meaningful
  - there is an absolute zero
  - e.g. number of visits to a hospital, distance, income

Data: Type and Scale of Attributes

In summary

Amount of Information

Attributes		Operations			
Type	Scale	$=, \neq$	$<, \leq, >, \geq$	$+, -$	$\times, \div$
Numeric	Ratio	✓	✓	✓	✓
	Interval	✓	✓	✓	
Categorical	Ordinal	✓	✓		
	Nominal	✓			

Transformation of attributes

... changing the scale type

- more informative $\rightarrow$ less informative
  - loss of information from the original scale
  - e.g. age $\rightarrow$ age group
- less informative $\rightarrow$ more informative
  - information limited by the original scale
  - e.g. birth date $\rightarrow$ age at current date

Transformation of attributes

... maintaining the scale type

- the scale type defines
- summarization and visualization operations
- admissible transformations that yield to equally legitimate representations
- so that genuine patterns from data are discovered

Data: Transformation of Attributes

Examples of transformations maintaining the scale:

- nominal: any permutation
  - eyecolor: $\{green, blue, brown\} \equiv \{blue, brown, green\}$
- ordinal: monotonic function that preserves the order
  - size: $\{small, medium, large\} \equiv \{36, 38, 40\}$
- interval: change the origin and the unit
  - temperature: $\{0^\circ C, 5^\circ C, 10^\circ C\} \equiv \{32^\circ F, 41^\circ F, 50^\circ F\}$
- ratio: change the unit
  - distance: $\{0\ km, 5\ km, 10\ km\} \simeq \{0\ mi, 3\ mi, 6\ mi\}$

Data: Important Characteristics

- Dimensionality (i.e. number of attributes)
  - high dimensional data brings several challenges
- Sparsity
  - only presence counts
- Resolution
  - patterns depend on the scale
- Size
  - type of analysis may depend on size of data

Data: Exploratory Analysis

"First things, first"

- For any data mining task to succeed,
  - analyzing and exploring data is essential!
- Summarization and visualization techniques
  - play a crucial role in data understanding and data preparation.

Data Summarization

Data Summarization

Motivation

- With big data sets it is hard to have an idea of what is going on in the data
- Data summaries provide overviews of key properties of the data
- Help selecting the most suitable tool for the analysis
- Describe important properties of the distribution of the values

Data Summarization

Common questions in data analysis

- What is the most common value?
- What is the variability in the values?
- Are there strange values?

Choosing the appropriate data analysis dependends on

- number of variables: univariate or multivariate
- type of variables: categorical or numeric

Descriptive Statistics

- Frequency
- Location or central tendency
- Dispersion
- Distribution

Frequency

- Absolute (or relative) occurrence of each value
- e.g. nr. of water samples by season

autumn	spring	summer	winter
40	53	45	62
20%	26.5%	22.5%	31%

- e.g. exam grades

8	10	11	13	15	17	18
1	2	3	4	8	5	2
4%	8%	12%	16%	32%	20%	8%

*For both categorical and numeric variables

Univariate analysis of location

- Minimum: the lowest value
- Maximum: the highest value
- Mode*: the most frequent value
- Mean: the average value (sensitive to extremes)

$$\mu_x = \frac{1}{n}\sum_{i=1}^{n} x_i$$

*For both categorical and numeric variables

Data Summarization: Univariate Data

Univariate analysis of location

- 1st Quartile (Q_1):
  - the value that is larger than 25% of the values
- Median / 2nd Quartile (Q_2):
  - the value above (below) which there are 50% of the values
- 3rd Quartile (Q_3):
  - the value that is larger than 25% of the values

Univariate analysis of variability or dispersion

- Range: $max_x - min_x$
- Standard Deviation - sensitive to extreme values

$$\sigma_x = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \mu_x)^2}$$

- Variance σ_x^2 - sensitive to extreme values
- Inter-quartile Range (*IQR*)
  - It is the difference between the 3rd (Q_3) and 1st (Q_1) quartiles

Frequency

- Contingency tables: cross-frequency of values for two variables
  - season and size

	autumn	spring	summer	winter
large	11	12	10	12
medium	16	21	21	26
small	13	20	14	24

Multivariate analysis of variability or dispersion

- Covariance Matrix: variance between every pair of numeric variables, .i.e. how they vary together;

$$cov(x, y) = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \mu_x)(y_i - \mu_y)$$

  the value depends on the magnitude of the variable.

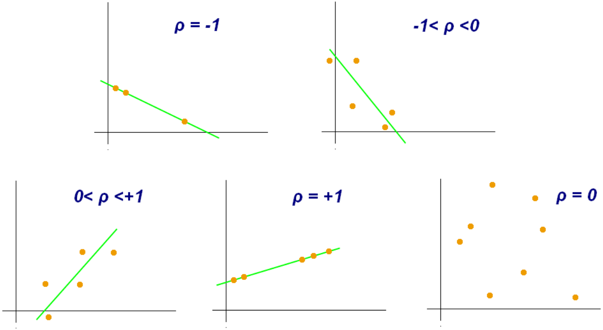
- Correlation Matrix: correlation between every pair of numeric variables, i.e. how a change in one variable will impact the other;

$$cor(x, y) = \frac{cov(x, y)}{\sigma_x \sigma_y}$$

  the influence of the magnitude is removed

Multivariate analysis of variability or dispersion

- Pearson Correlation Coefficient (ρ):
  - measures the linear correlation between two variables;
  - it has a value between +1 and -1.



Multivariate analysis of variability or dispersion

- Pearson Correlation Coefficient - cont.

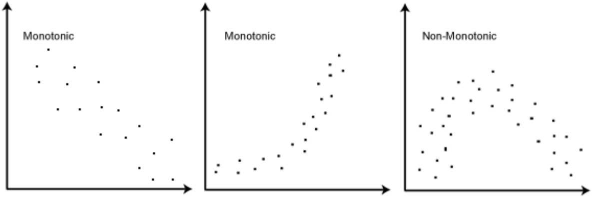
For a given sample of two variables x and y, $\{(x_1, y_1), ..., (x_n, y_n)\}$, the correlation coefficient is defined as

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

where n is the sample size, x_i and y_i are the individual sample points and $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ is the sample mean, the same for $\bar{y}$

Multivariate analysis of variability or dispersion

- Spearman Rank-Order Correlation Coefficient:
  - measures the strength and direction of monotonic association between two variables;
  - two variables can be related according to a type of non-linear but still monotonic relationship.



Multivariate analysis of variability or dispersion

- Spearman Rank-Order Correlation Coefficient: cont.
  - a rank-based, and non-parametric, version of *Pearson* correlation coefficient;
  - it has a value between +1 and -1;

$$rs_{xy} = r_{rank_x rank_y}$$

- if all n ranks are distinct integers, it can be computed using the popular formula

$$rs_{xy} = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n(n^2-1)}$$

where $d_i = rank_{x_i} - rank_{y_i}$ is the difference between the two ranks of each observation.

Data Summarization: Outliers

"An outlier is a point that deviates so much from the other data points as to arouse suspicions that it was generated by a different mechanism" (Hawkins, 1980)

- Outliers can be univariate or multivariate
- Statistical Parametric Techniques:
  - univariate case: boxplot definition (Tukey, 1977) is the most used one; any value outside the interval $[Q_1 - 1.5 \times IQR, Q_3 + 1.5 \times IQR]$
  - multivariate case: Mahalanobis distance (Mahalanobis, 1936).
- Statistical Non-parametric Techniques
  - Kernel functions
  - ...

Data Visualization

Data Visualization

Motivation

- Humans are outstanding at detecting patterns and structures
- Data visualization methods try to explore these capabilities
- Help detecting patterns and unusual patterns

Main Types of Visualization

- amounts
- distributions
- proportions
- associations
- trends
- time series
- geospatial data
- uncertainty

Some Graphs

- Barplots
- Piecharts
- Histograms
- Density Plots
- QQ Plots
- Boxplots
- Scatterplots
- Heatmaps
- Correlograms
- etc.

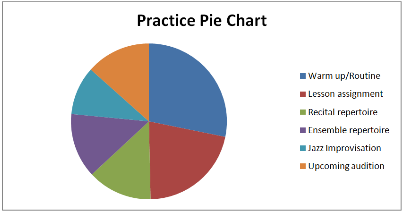
Data Visualization

Consider the people in this room.

- What graph would you choose for plotting
  - the distribution of ages?
  - the number of individuals by gender?
  - the proportion of individuals by gender?
  - the height and weight of each individual?
  - the height and weight of each individual by gender?

Piecharts

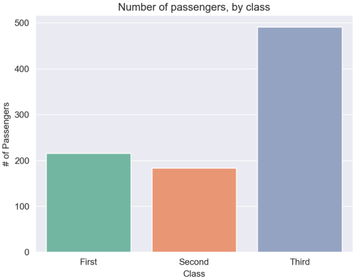
- Display the relative frequency of different values of a **categorical variable** in the form of a pie.



- They are **not a good option for comparison purposes**

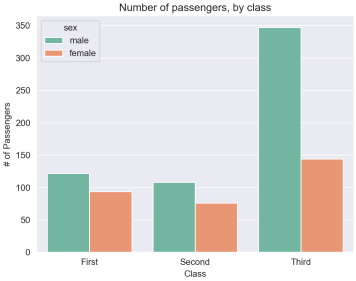
Barplots

- The main purpose is to display a set of values as heights of bars
- It can be used to display the frequency of occurrence of different values of a **categorical variable**



Barplot with two variables

- dodge
- stacked
- stacked (percent)

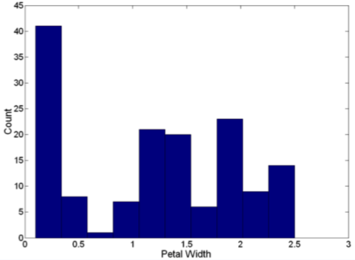


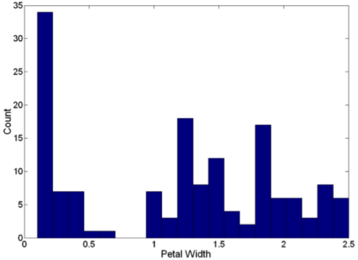
Histograms

- The main purpose is to display how the values of a **continuous variable** are distributed
- It is obtained as follows:
  - divide the range of the variable into a set of bins (intervals of values)
  - count the number of occurrences of values on each bin
  - display this number as a bar

Problems with Histograms

- Histograms may be misleading in small data sets
- The shape of the histogram depends on the number of bins
- How are the limits of the bins chosen? There are several algorithms for this.





Data Visualization: Distributions

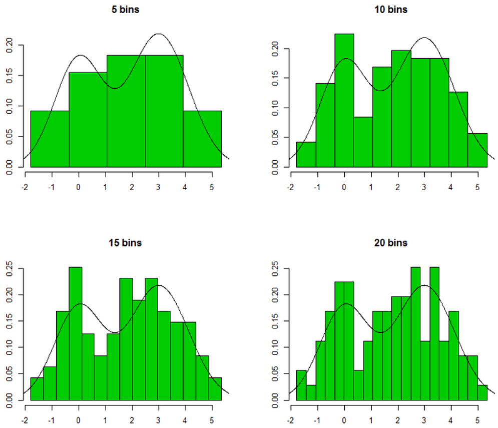
- Some of the problems of histograms can be tackled by smoothing the estimates of the distribution of the values. That is the purpose of kernel density estimates
- Kernel estimates calculate the estimate of the distribution at a certain point by smoothly averaging over the neighboring points
- Namely, the density is estimated by

$$\hat{f}_h(x) = \frac{1}{n}\sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right)$$

- where $K(.)$ is the kernel — a non-negative function — and $h > 0$ is a smoothing parameter called the bandwidth.

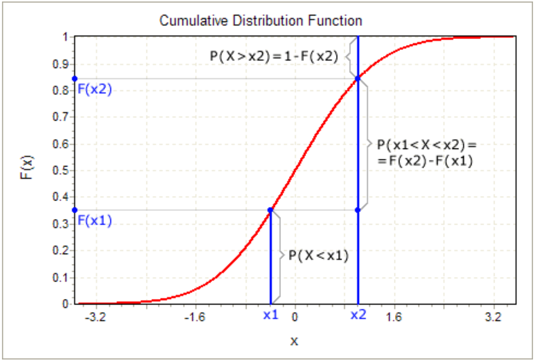
Data Visualization: Distributions

- Histograms with density estimate



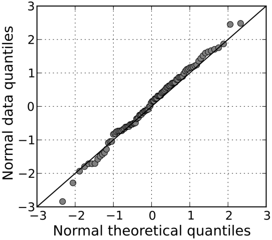
Cumulative Distribution Function (CDF)

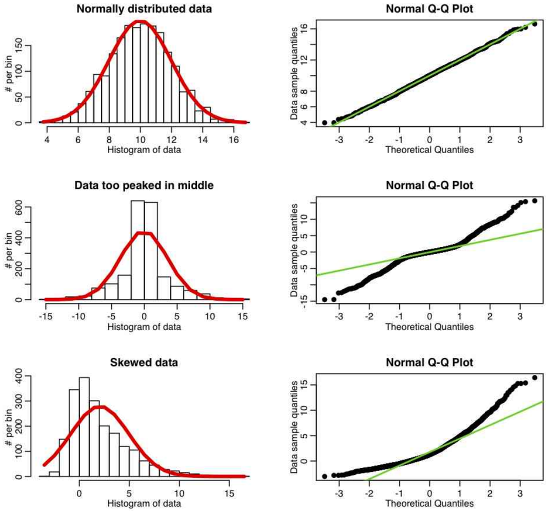
- CDF of a random variable X: $F_X(x) = P(X \leq x)$



QQ Plots

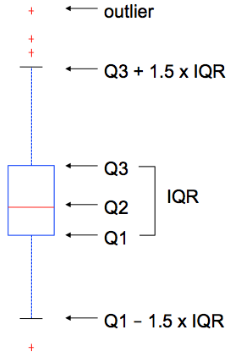
- Graphs that can be used to compare the observed distribution against the Normal distribution
- Can be used to visually check the hypothesis that the variable under study follows a normal distribution
- Obviously, more formal tests also exist

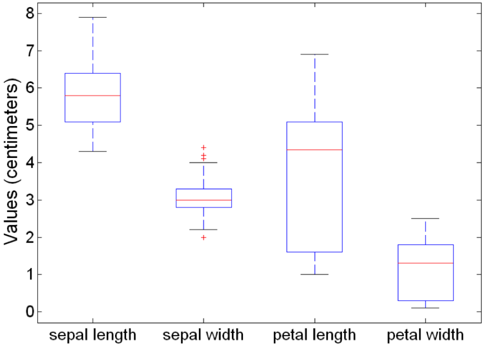

Normal data quantiles
3
2
1
0
−1
−2
−3
−3
−2
−1
0
1
2
3
Normal theoretical quantiles


Normally distributed data
per bin
0
50
100
150
4
6
8
10
12
14
16
Histogram of data
Normal Q-Q Plot
Data sample quantiles
4
6
8
12
16
-3
-2
-1
0
1
2
3
Theoretical Quantiles
Data too peaked in middle
per bin
0
200
400
600
-15
-10
-5
0
5
10
15
Histogram of data
Normal Q-Q Plot
Data sample quantiles
-15
-5
0
5
10
-3
-2
-1
0
1
2
3
Theoretical Quantiles
Skewed data
per bin
0
100
200
300
400
0
5
10
15
Histogram of data
Normal Q-Q Plot
Data sample quantiles
0
5
10
15
-3
-2
-1
0
1
2
3
Theoretical Quantiles

Boxplots

- An interesting summary of a variable distribution
- It inform us of the interquartile range and of the outliers (if any)

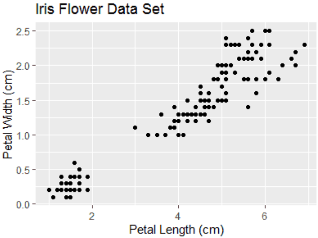




Data Visualization: Associations

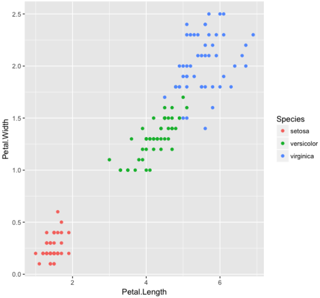
Scatterplots

- The natural graph for showing the relationship between two numeric variables



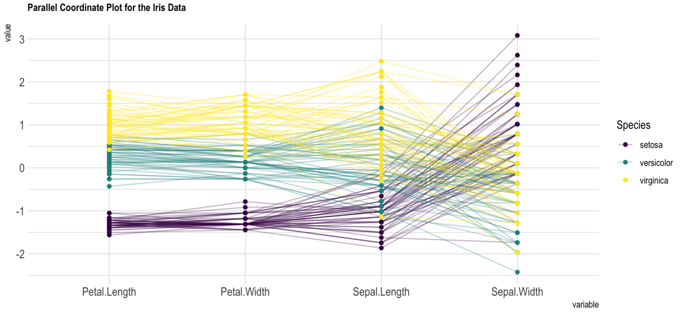
Data Visualization: Associations

- The scatterplot can plot the relationship between two numeric variables and with respect to a categorical variable



Parallel Sets

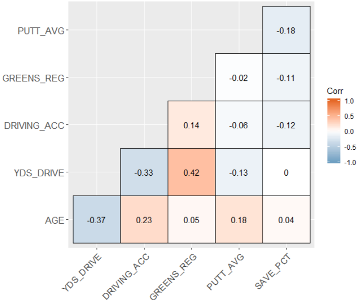
- Plots attributes values for each case (represented as a line)



- The order might be important to help identifying groups

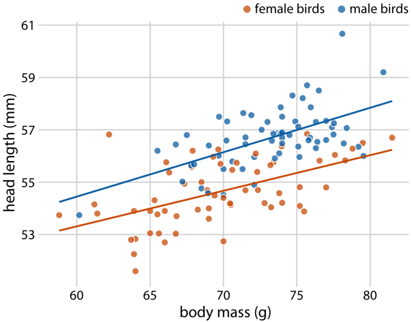
Correlograms

- visualization of correlation coefficients by a heatmap



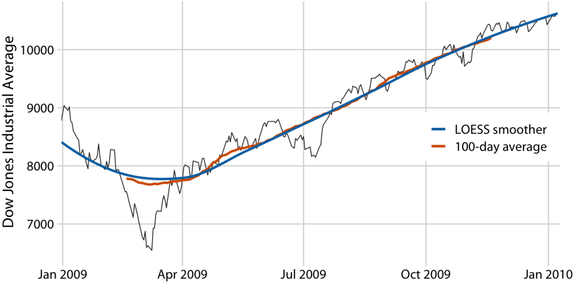
Scatterplots

- numerous functions exist to approximate the relationship between two numeric variables; scatter plot helps to perceive the trends



Time Series Plots

- moving average and other smoothing functions can be drawn on top of the original time series to perceive trends



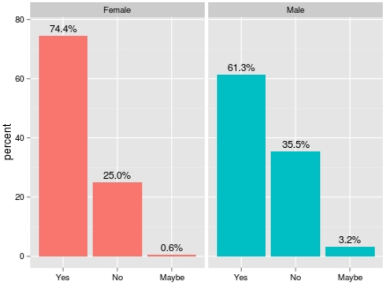
Data Visualization: Grouped Data

Graphs with grouped data

- Data sets frequently have categorical variables, which values can be used to create sub-groups of the data.
  - e.g. the sub-group of male/female clients of a company
- Conditioned plots allow the simultaneous presentation of these sub-group graphs to better allow finding eventual differences between the sub-groups

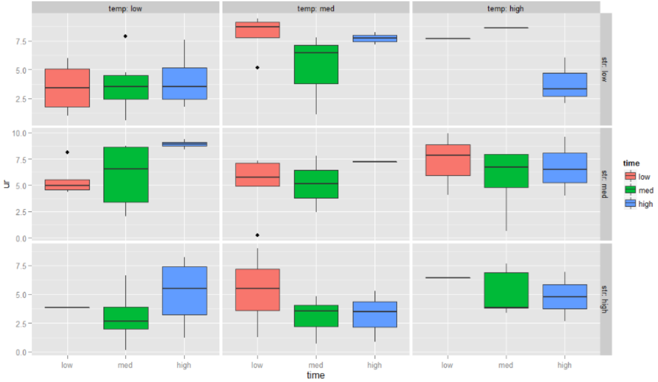
Graphs with grouped data

- groups on one categorical variable



Graphs with grouped data

- groups formed by cross-referencing of two categorical variables



Important Notes

- The purpose of data visualization is to convey meaningful information
- Is is very important to give it the right context providing appropriate
  - title
  - axis labels
  - legends
  - legend titles
  - other annotations

References

References

Aggarwal, Charu C. 2015. *Data Mining, the Texbook*. Springer.

Gama, João, André Carlos Ponce de Leon Ferreira de Carvalho, Katti Faceli, Ana Carolina Lorena, and Márcia Oliveira. 2015. *Extração de Conhecimento de Dados: Data Mining -3rd Edition*. Edições Sílabo.

Han, Jiawei, Micheline Kamber, and Jian Pei. 2011. *Data Mining: Concepts and Techniques*. 3rd ed. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc.

Moreira, João, Andre Carvalho, and Tomás Horvath. 2018. *Data Analytics: A General Introduction*. Wiley.

Tan, Pang-Ning, Michael Steinbach, Anuj Karpatne, and Vipin Kumar. 2018. *Introduction to Data Mining*. 2nd ed. Pearson.