DATA SCIENCE MEI/1

University of Beira Interior
Department of Informatics

Hugo Pedro Proença,
hugomcp@di.ubi.pt, **2021/22**

Data Visualization and Summarization

- Vision is the Human dominant sense: **80–85% of information** we perceive, learn or process is mediated through vision.
- It is particularly important **to understand and interpret data,** when trying to find (illustrate) relationships among hundreds or thousands of variables (features).
- Data visualization can be regarded as a way to perform **exploratory analysis**
  - The goal is to "get to know" the dataset.
- There are two major advantages resulting of this kind of preprocessing
  - We gain valuable hints for Data Cleaning, which often represents the key feature between successful/failed models.
  - We get potential ideas for Feature Engineering, which can also play a significant role in the final effectiveness of our models.
- A third advantage is to "feel" the dataset, which can help us to communicate results and deliver greater impact.

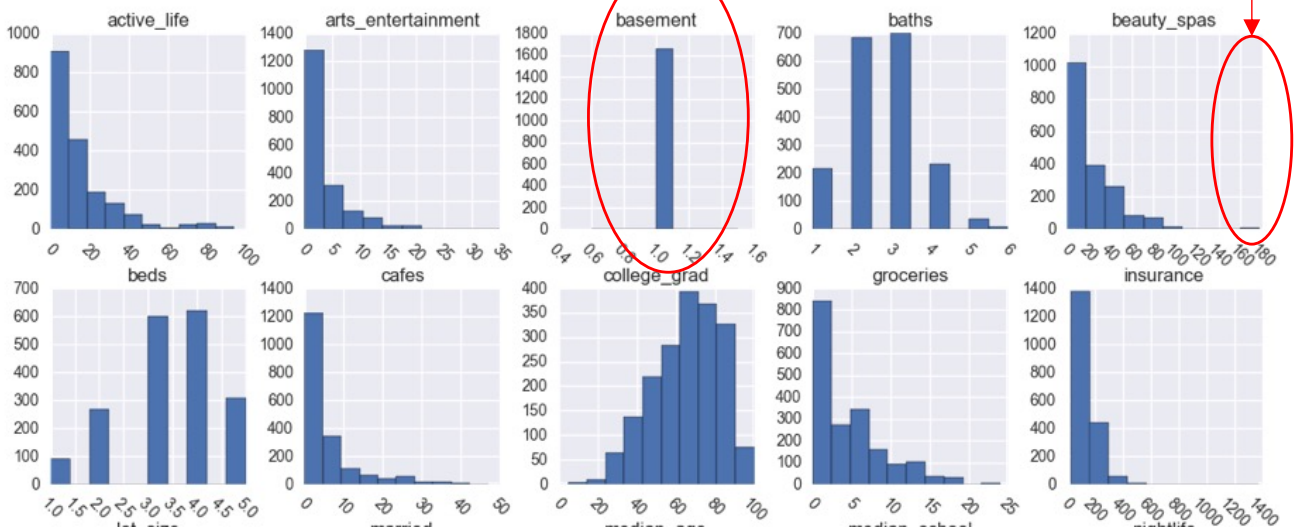
Data Visualization and Summarization

- The first step is to perceive the basic properties of the dataset:
  - How many observations do we have? How many features?
  - What is the data type of each feature? Numeric? Categorical?
  - What is the scale of each feature? Do we need to perform any kind of normalization?
  - Do we have target (label) variables?
  - Are there any missing values?
- It might be important to display some (random?) instances of the dataset, to "feel" the values of each feature, and perceive if everything makes sense.
  - Example: (real-estate dataset)

	tx_price	beds	baths	sqft	year_built	lot_size	property_type	exterior_walls	roof	basement	restaurants	groceries	nightlife
0	295850	1	1	584	2013	0	Apartment / Condo / Townhouse	Wood Siding	NaN	NaN	107	9	30
1	216500	1	1	612	1965	0	Apartment / Condo / Townhouse	Brick	Composition Shingle	1.0	105	15	6
2	279900	1	1	615	1963	0	Apartment / Condo / Townhouse	Wood Siding	NaN	NaN	183	13	31
3	379900	1	1	618	2000	33541	Apartment / Condo / Townhouse	Wood Siding	NaN	NaN	198	9	38
4	340000	1	1	634	1992	0	Apartment / Condo / Townhouse	Brick	NaN	NaN	149	7	22

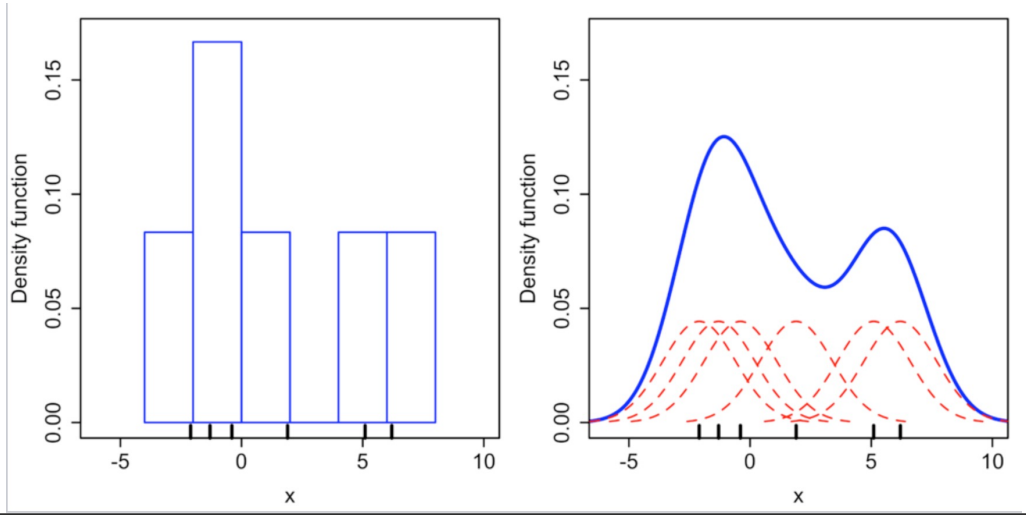
Data Visualization and Summarization

- After converting all features into a numerical scale, it is typically important to plot the (independent) distributions of the features.
- This can be done using two major techniques: 1) Histogram analysis; or 2)Density estimation
- A grid of 1D histograms/densities can help us to perceive:
  - Unexpected distributions
  - Potential outliers
  - Features that should be binary
  - Potential measurement errors



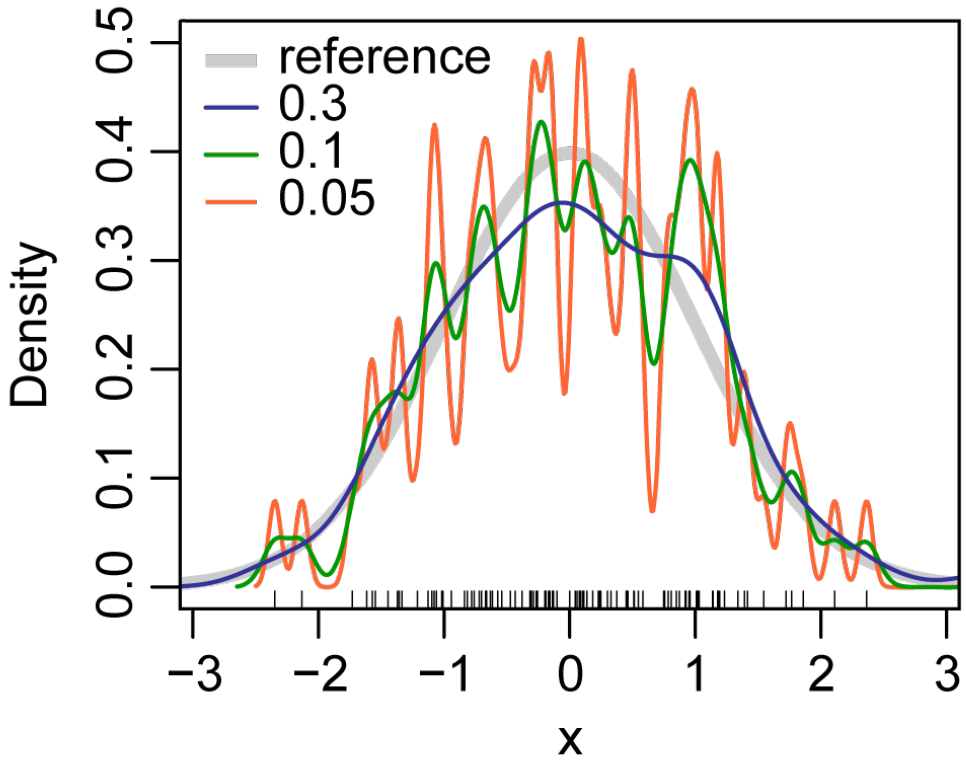
Kernel Density Estimates

- Kernel density estimation (KDE) is a non-parametric way to estimate the probability density function of a random variable.
- It is an important data smoothing problem where inferences about the population are made.
  - This technique is also termed as **Parzen–Rosenblatt** window method, after Emanuel Parzen and Murray Rosenblatt, who are credited with independently creating it in its current form.
- KDEs are closely related to histograms, typically adding properties such as smoothness or continuity by using a suitable **kernel**.
  - The algorithm evolves to find the position of a set of kernels (e.g., Gaussian), that when summed up, approximate the observed sample:



Kernel Density Estimates

- The bandwidth of the kernel is a free parameter which exhibits a strong influence on the resulting estimate.
  - Changing the bandwidth changes the shape of the kernel:
  - Lower bandwidths mean that only points very close to the current position are given any weight, which leads to the estimate looking squiggly;
  - In opposition, higher bandwidth provide shallow kernels, where distant points can still contribute.



Categorical Features Conversion

- When converting categorical features into a numerical scale, it might be particularly sensitive to consider the notion of **Manifold**
- A manifold is a topological space that locally resembles a Euclidean space near each point.
- In practice, the categorical values that are **semantically adjacent** should be projected into neighbor regions of the manifold.
- Example: Consider a categorical feature that provides the current state of the weather.
  - Possible values are: "*Stormy*", "*Rainy*", "*Cloudy*" and "*Sunny*".

# Instance	Weather
1	Stormy
2	Sunny
3	Rainy
4	Cloudy
5	Rainy

# Instance	Weather
1	3
2	4
3	2
4	1
5	2

"Alphabetic order"

"Semantic order"

# Instance	Weather
1	1
2	4
3	3
4	2
5	3

One-hot Encoding

- There are cases where no ordinal relationship exists between the categorical variables
  - For such cases, the "integer encoding" scheme is not appropriate
- Using such encoding scheme will let the model assume a natural ordering between categories, which may result in poor performance or unexpected results (predictions halfway between categories).
- In this case, a **one-hot encoding** can be applied to the integer representation.
- In this scheme, a binary variable is added for each possible feature value.
- For example, consider a feature that provides the "brand" of a specific product (e.g., a car).
  - A possibility will be to encode the values based on alphabetical order
  - However, we do not want to assume that "**A**udi" is closer to "**B**MW" than to "**V**olvo".
- For that cases, we create as many columns as possible values in the feature and set a specific cell to "1" to specify "Yes" and "0" otherwise.

# Instance	Brand
1	Audi
2	BMW
3	Audi
4	Volvo
5	Renault

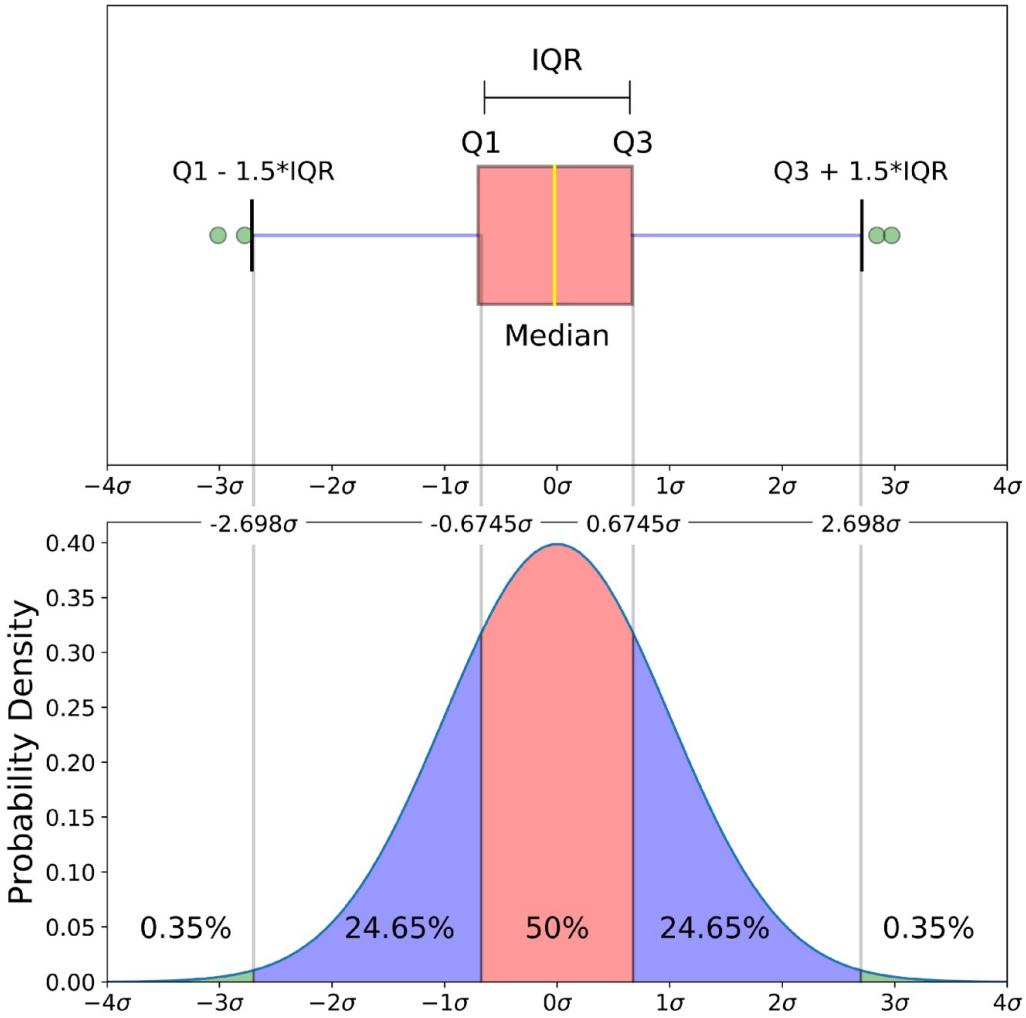
#Instance	Audi	BMW	Renault	Volvo
1	1	0	0	0
2	0	1	0	0
3	1	0	0	0
4	0	0	0	1
5	0	0	1	0

Boxplots

- Apart the well-known **Linear/Scatter plots** (to express the relationship between two variables), **Bar charts** (to compare the quantities of different groups or features) and **Pie charts** (to compare the parts of a whole), **Boxplots** offer a tool to summarize the distribution of a feature.
- They are binned plots with whiskers, displaying five statistics:
  - The minimum, lower quartile, median, upper quartile and maximum
  - The lower quartile (25th percentile) is represented by the lower edge of the box,
  - The upper quartile (75th percentile) is represented by the upper edge of the box.
  - The median (50th percentile) is represented by a central line that divides the box into sections.
  - Extreme values are represented by whiskers that extend out from the edges of the box.
- Box plots are often used to understand the outliers in the data.

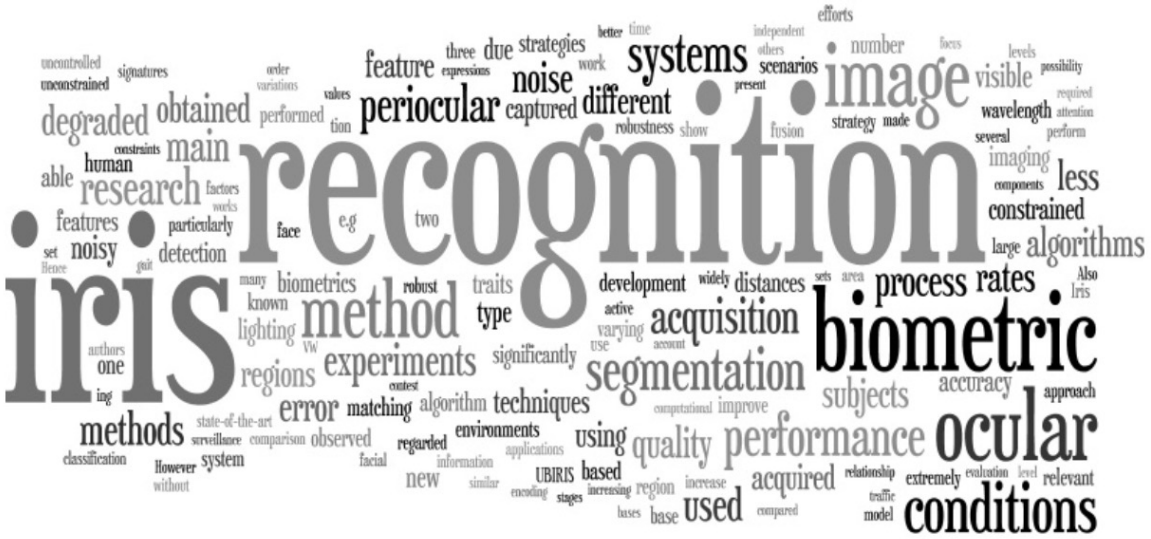
Boxplots

- The interquartile range (IQR) is the amplitude between the 25th and the 75th percentiles.
- The "*maximum*" is defined as the upper quartile value + 1.5*IQR
- Similarly, the "*minimum*" is defined as the lower quartile value - 1.5*IQR
- Values outside this interval are considered "outliers"



Semi-Structured or Unstructured Data

- For semi-structured or even unstructured data other (less conventional techniques) can be considered.
- For instance, **Word Clouds** can be a useful illustration tool to summarize the most important terms of a text source.
- A word cloud visual represents the frequency of a word within a body of text with its relative size in the cloud. This technique is used on unstructured data to display high- or low-frequency words.
- Example: “Word cloud” of all publications of a group in the SOCIA Lab.



Semi-Structured or Unstructured Data

- Another interesting visualization technique that can be used for semi-structured or unstructured data is the **Network Diagram**.
  - Network diagrams represent relationships as nodes (individual actors within the network) and ties (relationships between the individuals). They are used in many applications, for example for analysis of social networks or mapping product sales across geographic areas.
  - Example: Network diagram about the co-authorship of a set of scientific publications: