VSB TECHNICAL UNIVERSITY OF OSTRAVA | FACULTY OF ELECTRICAL ENGINEERING AND COMPUTER SCIENCE | DEPARTMENT OF COMPUTER SCIENCE

Algorithms for Big Data

Vector data processing, Exploratory analysis

Jan Platoš

November 21, 2020

Department of Computer Science
Faculty of Electrical Engineering and Computer Science
VŠB - Technical University of Ostrava

Table of contents

Explorative Analysis

Explorative Analysis

- What do you imagine?

Explorative Analysis

- What do you imagine?
- What is the shape of the data?

Explorative Analysis

- What do you imagine?
- What is the shape of the data?
- What is the distribution of the data?

Explorative Analysis

- What do you imagine?
- What is the shape of the data?
- What is the distribution of the data?
- What is the distribution among the features?

Explorative Analysis

- What do you imagine?
- What is the shape of the data?
- What is the distribution of the data?
- What is the distribution among the features?
- Is there any relation between features?

Explorative Analysis

Exploratory Data Analysis refers to the critical process of performing initial investigations on data so as to **discover patterns**, to spot **anomalies**, to **test hypothesis** and to **check assumptions** with the help of **summary statistics** and **graphical representations**.[1]

[1] https://towardsdatascience.com/exploratory-data-analysis-8fc1cb20fd15

Processing of the Data

- Dataset investigations
  - How many features it has?
  - How many records it has?
  - What is the content of the features?
  - What type of the features it has - Categorial, Numeric, Text?

Feature types - Categorial features

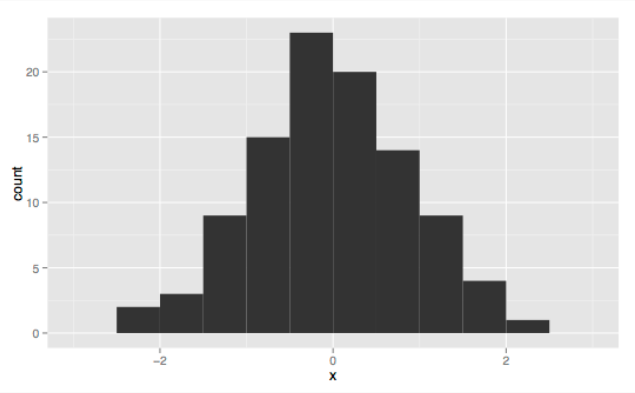
- How many distinct features it has?
- Is it a class label or regular feature?
- What is the distribution of the values?

Feature types - Numeric features

- What is the minimum and maximum?
- What are the quartiles?
- What is the distribution/histogram of the data?
- What are the properties of the data distribution?
  - Mean $\overline{x} = \frac{\sum_{i=1}^{n} x_i}{n}$
  - Median - a middle value of sorted values.
  - Variance $s^2 = \frac{\sum_{i=1}^{n} (x_i - \overline{x})^2}{n-1}$
  - Std. deviation $\sigma = \sqrt{s^2}$
  - Inter-quartile range $IQR = Q_3 - Q_1$

Feature types - Graphical analysis

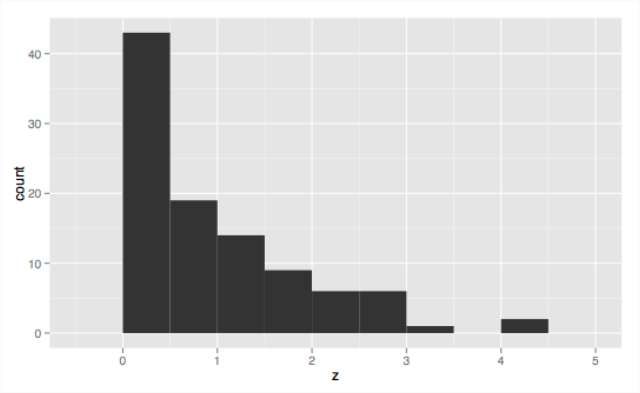
- Histograms[2]



[2] https://en.wikipedia.org/wiki/Histogram

Feature types - Graphical analysis

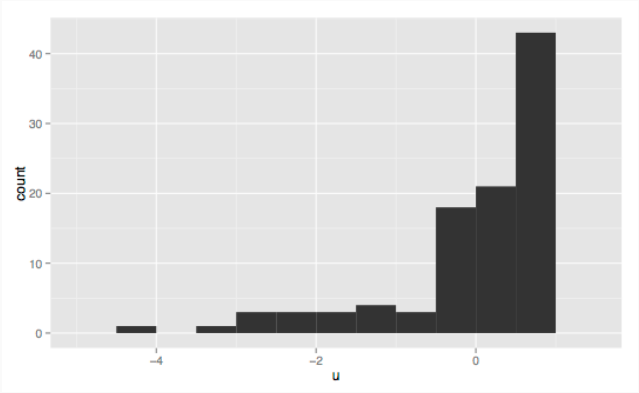
- Histograms[2]



[2] https://en.wikipedia.org/wiki/Histogram

Feature types - Graphical analysis

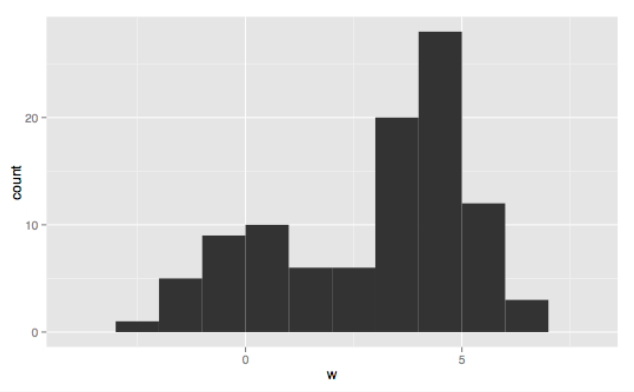
- Histograms[2]



[2] https://en.wikipedia.org/wiki/Histogram

Feature types - Graphical analysis

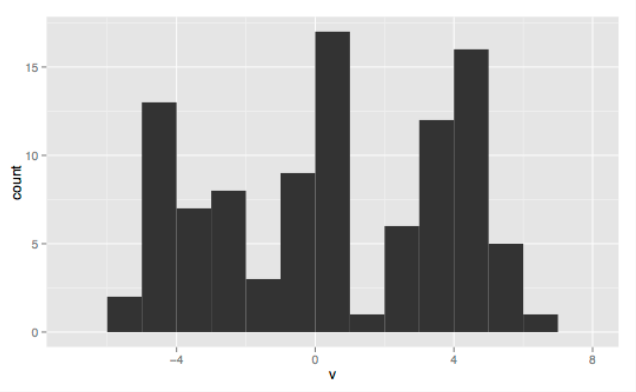
- Histograms[2]



[2] https://en.wikipedia.org/wiki/Histogram

Feature types - Graphical analysis

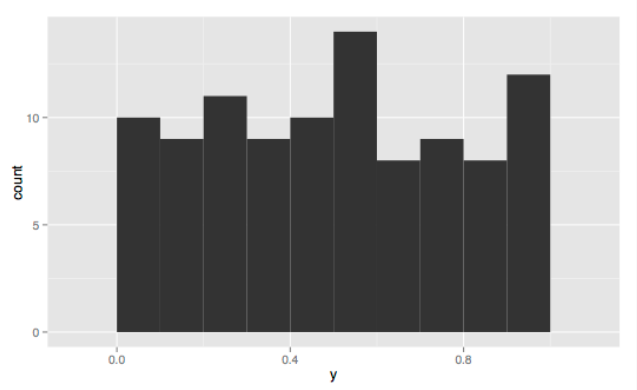
- Histograms[2]



[2] https://en.wikipedia.org/wiki/Histogram

Feature types - Graphical analysis

- Histograms[2]



[2] https://en.wikipedia.org/wiki/Histogram

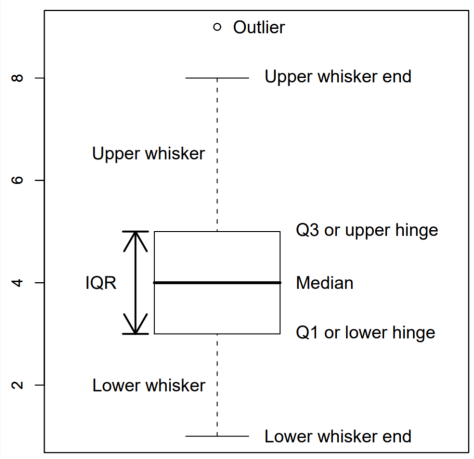
Feature types - Graphical analysis

- Histograms
- Ideal for categorial data
- Numeric values requires another parameter - **Bin size**

Feature types - Graphical analysis

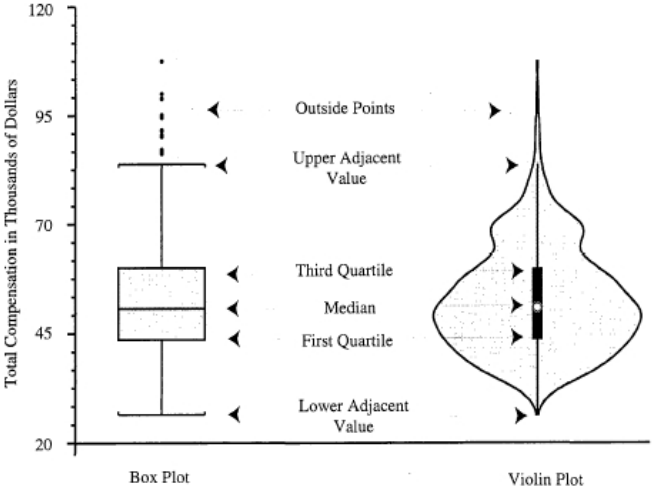
Box plot (Box and Whiskers plot)

- Minimum : the lowest data point excluding any outliers.
- Maximum : the largest data point excluding any outliers.
- Median : the middle value of the dataset.
- First quartile : the median of the lower half of the dataset.
- Third quartile : the median of the upper half of the dataset.



Feature types - Graphical analysis

- Violin plot[3]



Relationship between features

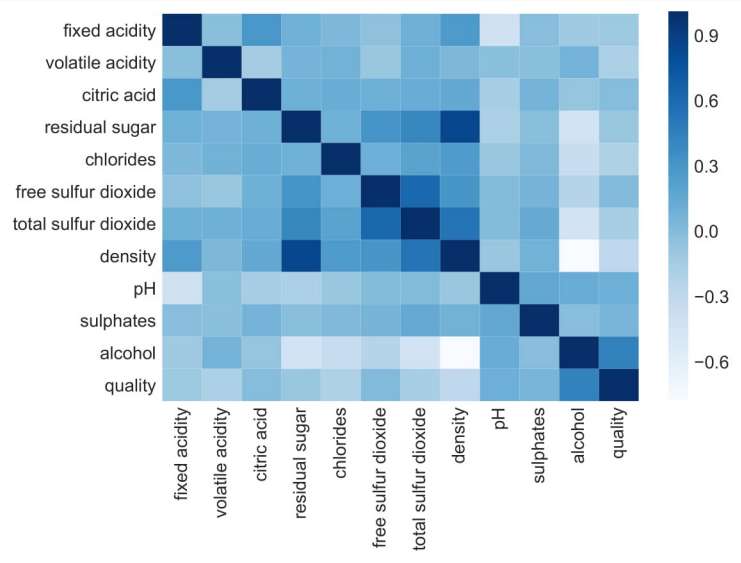
- There are many ways how to identify the relationships between features.
- **Covariance** - how much (and in what direction) should we expect one variable to change when the other changes.

$$Cov(X, Y) = \frac{\sum_{i-1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{n - 1}$$

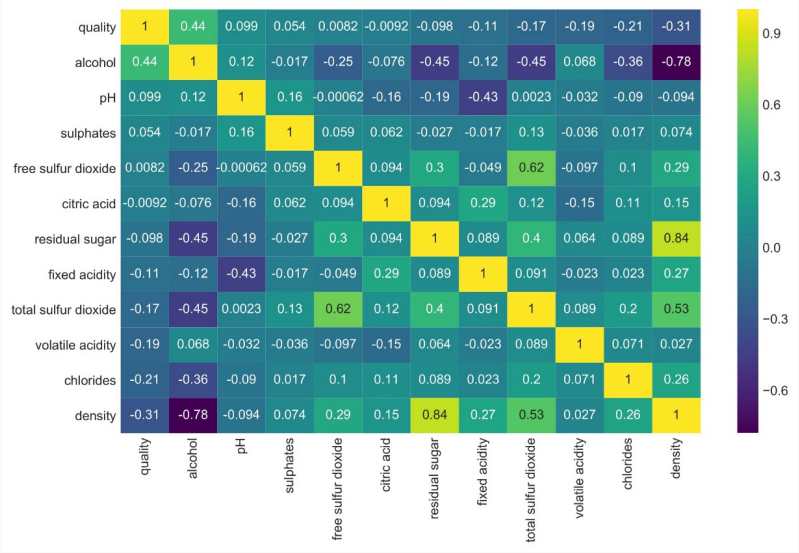
- **Correlation** - similarly to covariance, the power and direction of the relation.

$$Cor(X, Y) = \frac{Cov(X, Y)}{\sigma_x \sigma_y}$$

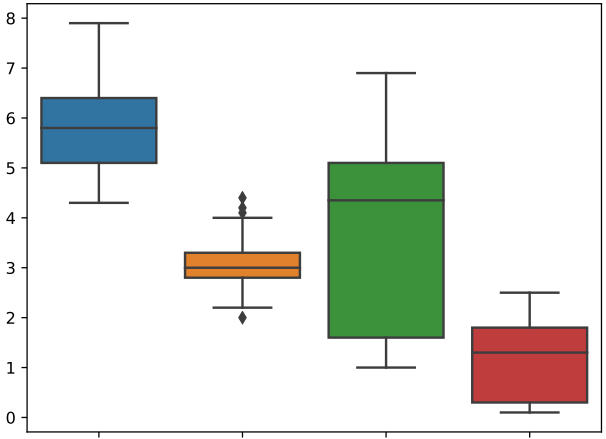
Relationship between features - Correlation Heatmap



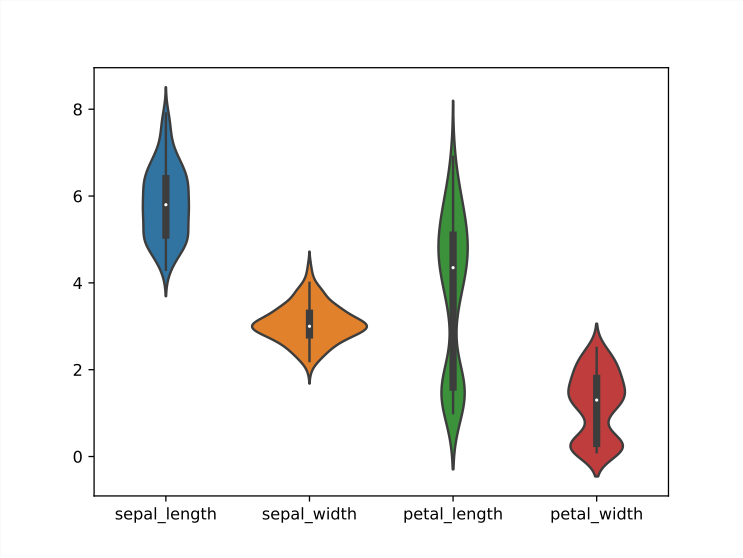
Relationship between features - Correlation Heatmap



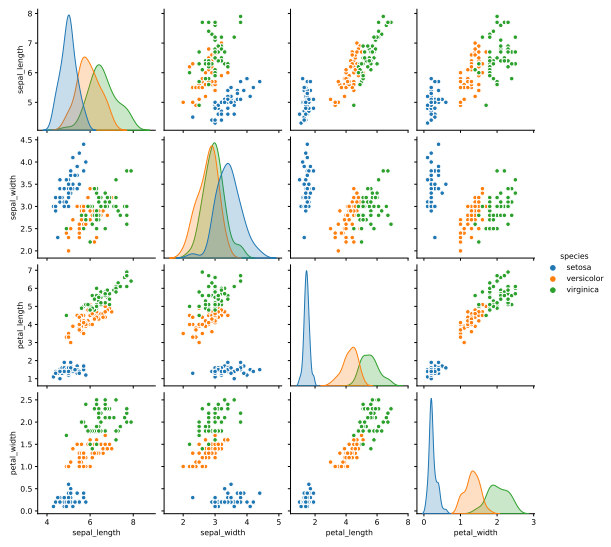
Relationship between features - Box plot



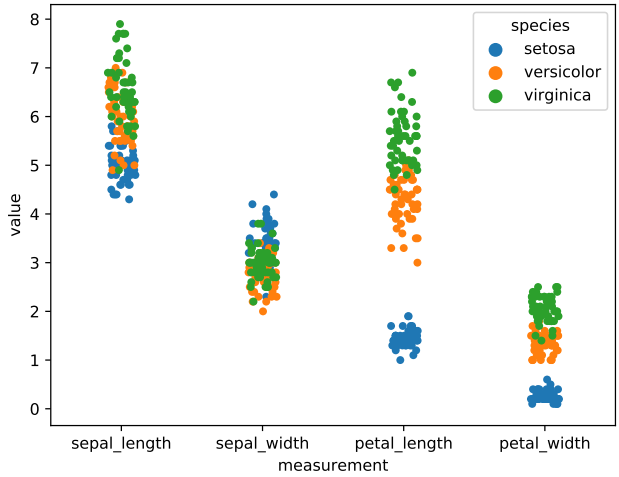
Relationship between features - Violin plot



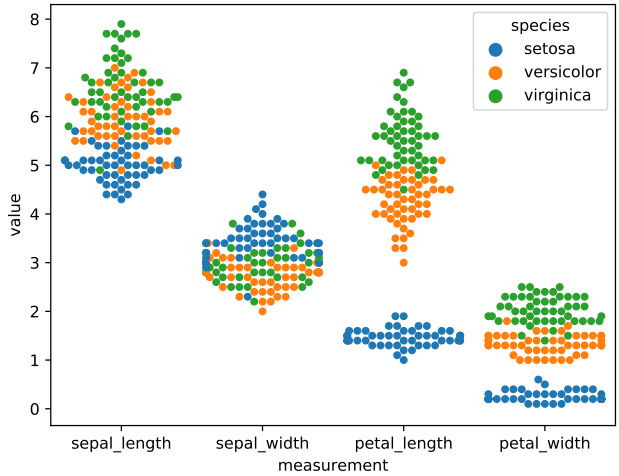
Relationship between features - Pair plot



Relationship between features - Strip plot



Relationship between features - Swarm plot



Feature rescaling

Feature rescaling

- Feature scaling is a method used to normalize the range of independent variables or features of data.
- The process is known also as *data normalization*.
- The scaling is necessary to compare features between each other and use a metric to compute non-biased distance.
- The scaling have to respect the nature of the data.

Min-Max Scaling

- The most common scaling principle.
- Unifies the *min* and *max* among the features.
- Normalize into range $[0, 1]$

$$x' = \frac{x - \min(x)}{\max(x) - \min(x)}$$

- Normalize into range $[a, b]$

$$x' = a + \frac{(x - \min(x))\,(b - a)}{\max(x) - \min(x)}$$

- Maximum Absolute Scaling where we normalize to $|\max(x)|$

Mean Scaling

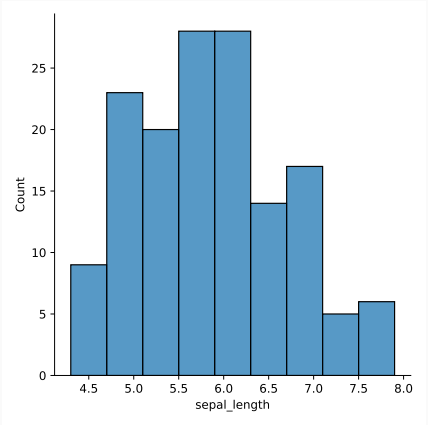
- Moves the mean of the data into 0.
- The final range is $[-1, +1]$.

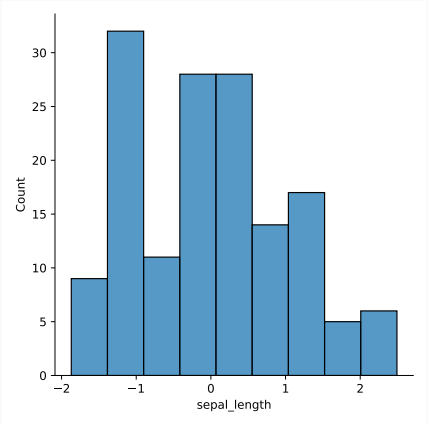
$$x' = \frac{x - \overline{x}}{\max(x) - \min(x)}$$

Standardizing

- Normalize the values to zero mean and unit variance.

$$x' = \frac{x - \overline{x}}{\sigma}$$





Non-linear Transform

- Remapping features distribution into form close to Normal distribution
- Box-Cox transform or Yeo-Johnson transform

$$x^{(\lambda)} = \begin{cases} \frac{x_i^{\lambda}-1}{\lambda} & if\ \lambda \neq 0 \\ \ln(x_i) & if\ \lambda = 0 \end{cases}$$

Categorial Data Encoding

Categorial Data Encoding

- Ordinal encoding.
- One-hot encoding.
- Embedding (mainly for words)
  - Word2Vec
  - Glove
  - ...

Classification

Classification

Basic questions:

- What it is?
- What it needs?
- What it produces?

Classification

Basic questions:

- What it is?
- What it needs?
- What it produces?

Definition

Given a set of training data points, each of which is associated with a class label, determine the class label of one or more previously unseen test instances.

Classification

Phases of classification:

- Training phase - construction of models from the training instances.
- Testing phase - determining class labels of one or more training instances.

Output of classification:

- Label prediction - one fixed label is predicted.
- Numerical score - numerical evaluation of each label assignment to the instance.

Questions?