

ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

IA PER LA MODA

Data Visualization & Orange

28/03/2022

Alessia Angeli

Studente di dottorato in Data Science and Computation
Dipartimento di Informatica – Scienza e Ingegneria



CV (in breve)

- Studente di dottorato in Data Science and Computation
- Laurea Magistrale in Matematica, curriculum generale/applicativo
- Laurea Triennale in Matematica



ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

VARLAB: VIRTUAL AND AUGMENTED REALITY LAB

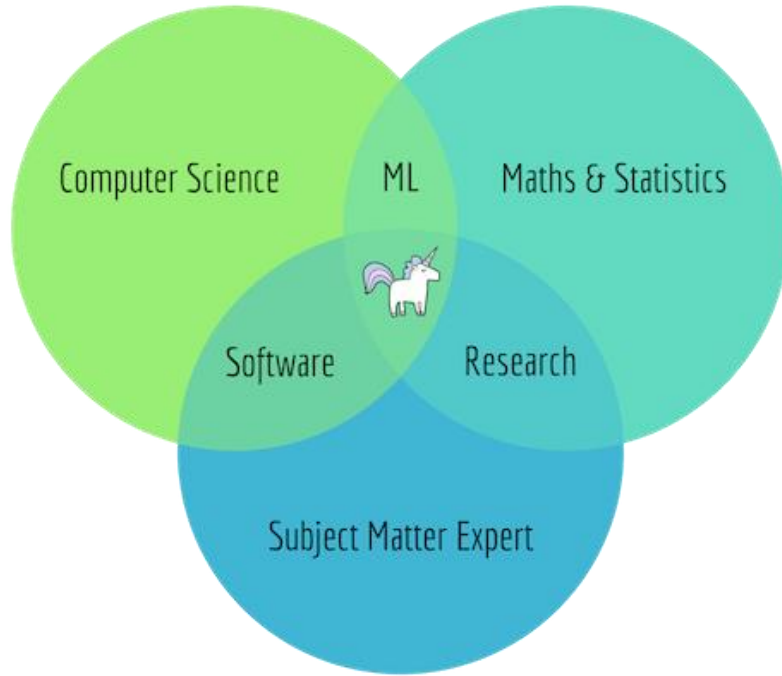


Contatti



ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

DATA SCIENCE



Maths



Machine and Deep Learning



Data Visualization

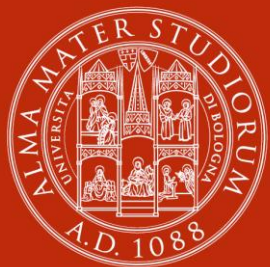


Augmented Reality

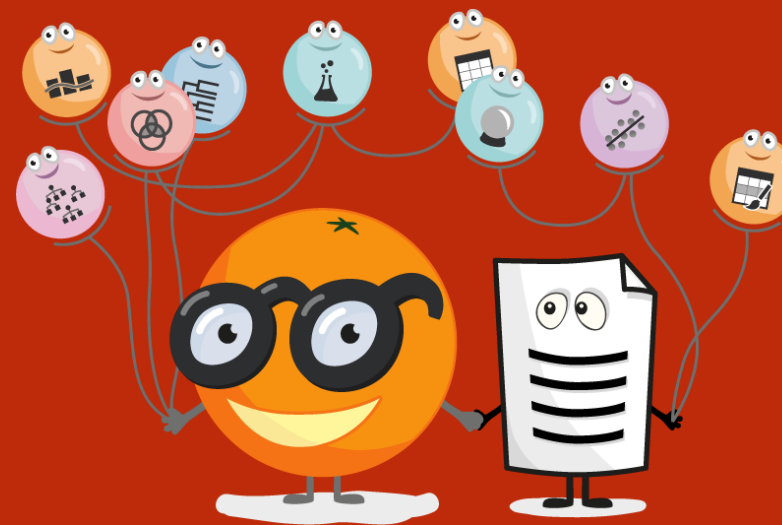


Fashion





ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA



Download Orange

Alessia Angeli

Studiante di dottorato in Data Science and Computation
Dipartimento di Informatica – Scienza e Ingegneria

Download Orange

[Screenshots](#)[Workflows](#)[Download](#)[Blog](#)[Docs](#)[Workshops](#)[Donate](#)

Data Mining Fruitful and Fun

Open source machine learning and data visualization.

Build data analysis workflows visually, with a large, diverse toolbox.

[Download Orange](#)

<https://orangedatamining.com/>



ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

Download Orange

[Screenshots](#)[Workflows](#)[Download](#)[Blog](#)[Docs](#)[Workshops](#)[Donate](#)

Windows



macOS



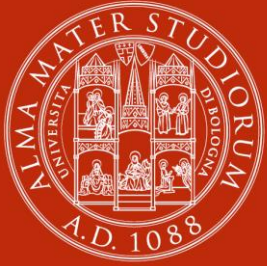
Linux / Source

Download the latest version for Windows

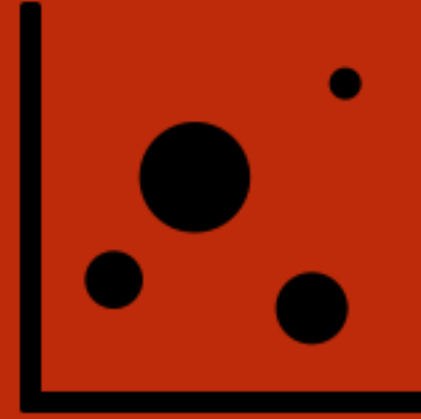
[Download Orange 3.31.1](#)

ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

Riprendiamo... (in parte)



ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA



Data Visualization – Plots/Graphs

Alessia Angeli

Studente di dottorato in Data Science and Computation
Dipartimento di Informatica – Scienza e Ingegneria

Quantitative Attributes

Scatter plot
Histogram
Scatter plot matrix
Box plot
Violin plot
Radar chart
...



Scatter plot

What?

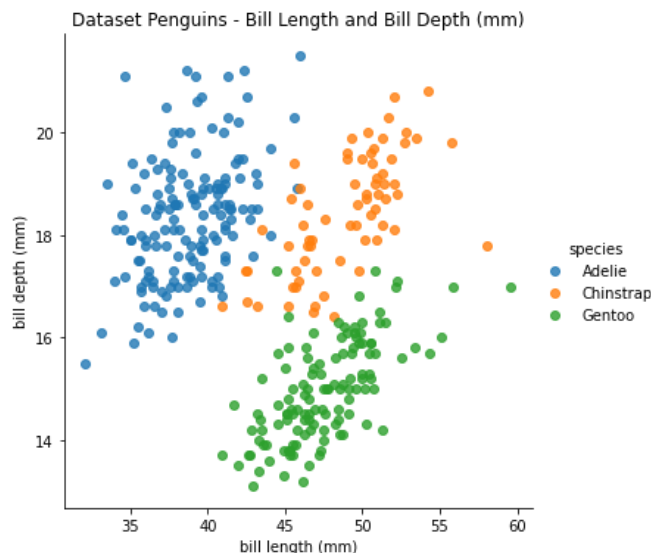
- 2 quantitative attributes;

Why?

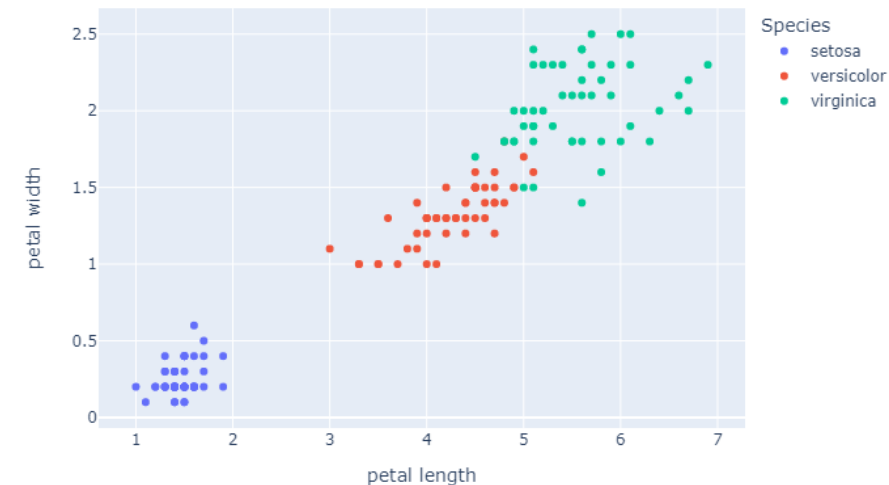
- Visualizzare correlazioni e distribuzioni;
- Identificare outliers, patterns e clusters;

Remarks

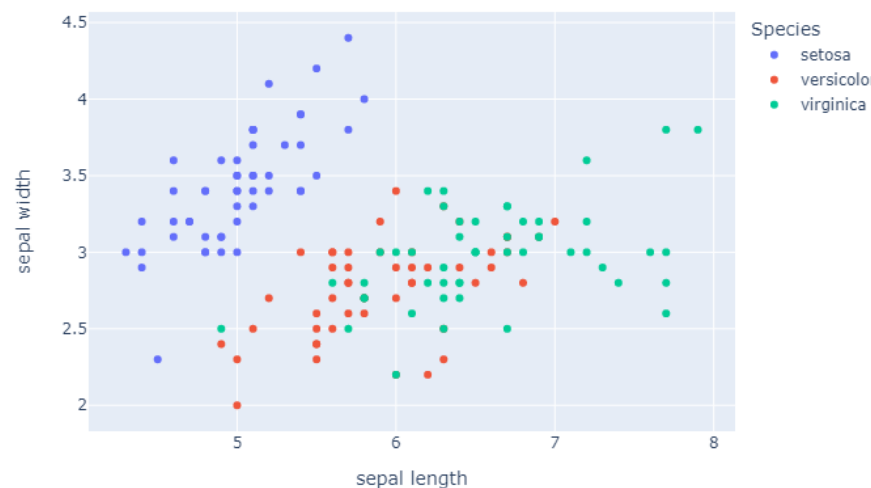
- Fino a ~100 items;
- Colore e dimensione possono essere usati per codificare categorical attributes aggiuntivi (bubble plot).



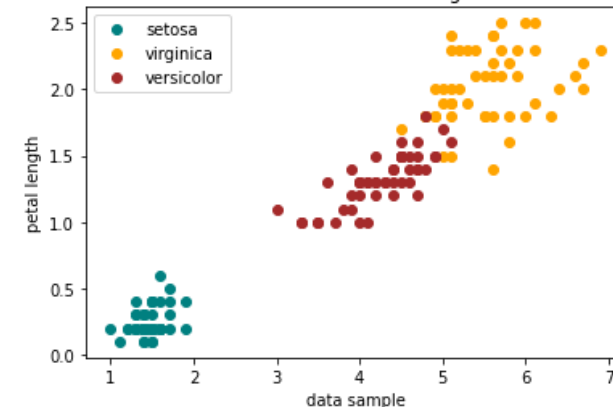
Petal Length and Petal Width of different Iris Species



Sepal Length and Sepal Width of different Iris Species



Dataset Iris - Petal Length

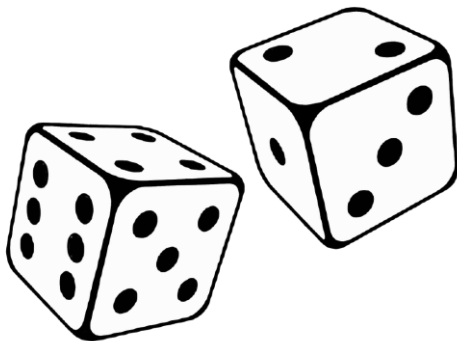


Definizione

DISTRIBUZIONE DI PROBABILITA': Una distribuzione di probabilità è un modello matematico che associa ai valori (possibili) di una variabile aleatoria (continua o discreta) le probabilità che tali valori possano essere assunti da tale variabile. Formalmente le distribuzioni vengono espresse da funzioni matematiche, **funzione densità di probabilità** e **funzione di probabilità**, rispettivamente per variabili aleatorie continue e discrete.

ESEMPIO

Si lanciano 2 dadi e si considera come variabile aleatoria la somma risultante.



Somma	# Combinazioni	Probabilità
2	1	0.03
3	2	0.06
4	3	0.08
5	4	0.11
6	5	0.14
7	6	0.17
8	5	0.14
9	4	0.11
10	3	0.08
11	2	0.06
12	1	0.03

$\Sigma 36$

$\Sigma 1$



Histogram

What?

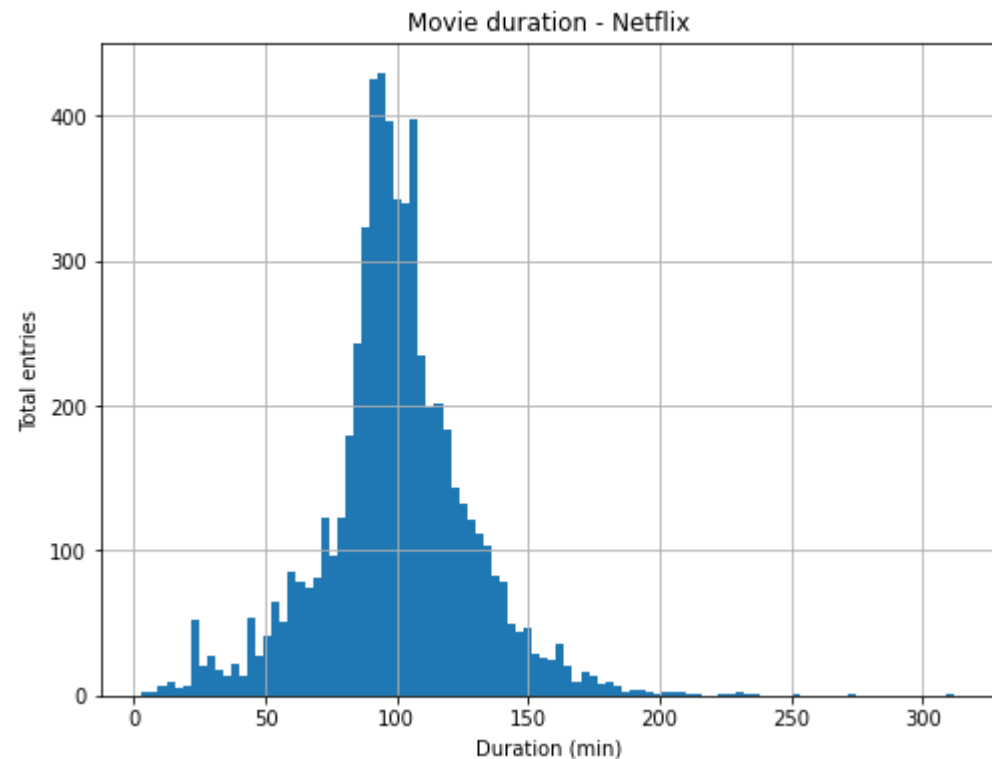
- 1 quantitative attribute;

Why?

- Visualizzare distribuzioni;
- Identificare patterns e range;

Remarks

- Una linea (o un'area) può essere visualizzata per mostrare la funzione di densità calcolata;
- Gli items possono essere visualizzati con dei punti.



Definizione

MATRICE: una matrice è una tabella ordinata. Le righe orizzontali vengono chiamate *righe* della matrice e le righe verticali *colonne* della matrice.

Generalmente una matrice si indica con una lettera maiuscola e viene scritta nel modo seguente:

$$A = \begin{pmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \dots & \dots & \dots & \dots \\ a_{m1} & a_{m2} & \dots & a_{mn} \end{pmatrix}$$

dove i pedici di ogni elemento della matrice indicano, rispettivamente, la riga e la colonna in cui l'elemento è posizionato.

Quindi a_{ij} è l'elemento della matrice A che si trova nella riga i -esima e nella colonna j -esima.



Scatter plot matrix

What?

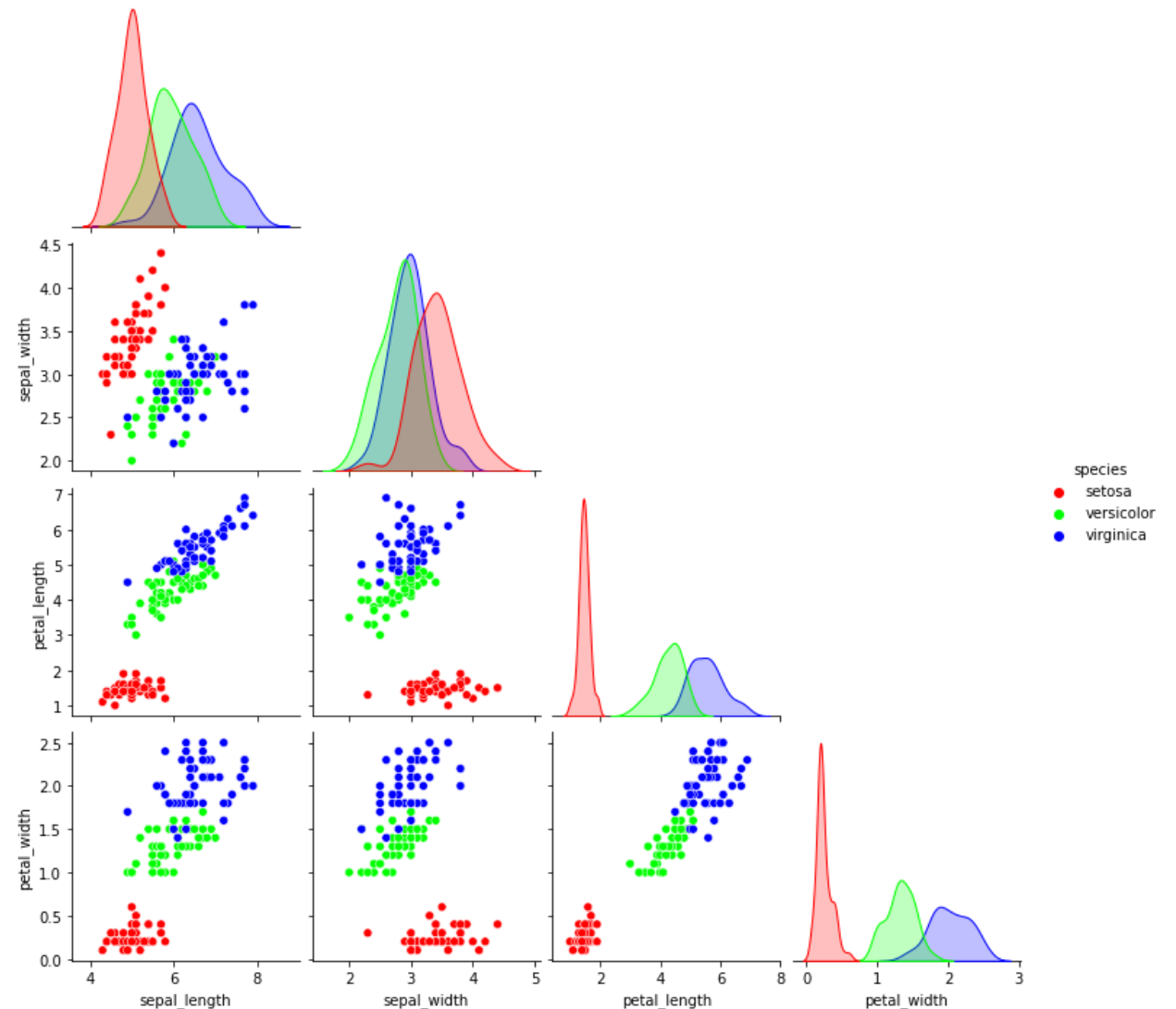
- N quantitative attributes;

Why?

- Visualizzare correlazioni e distribuzioni;
- Identificare outliers, patterns e clusters;

Remarks

- Fino a ~12 attributi e ~100 items;
- E' possibile visualizzare solo la parte triangolare inferiore della matrice.



Statistica descrittiva – alcune definizioni

Considerando un insieme di dati numerici si definiscono:

5 7 4 6 5

MEDIA (MEDIA ARITMETICA): rapporto tra la somma dei dati e il numero dei dati.

$$(5+7+4+6+5)/5=5.4$$

MODA: il valore del dato che si presenta con maggiore frequenza (possono essere presenti più valori di moda).

5 – dato con massima frequenza (2)

MEDIANA: è il valore centrale tra i dati ordinati in modo crescente o decrescente. Se l'insieme contiene un numero di dati dispari c'è un unico valore centrale e questo è la mediana. Se l'insieme contiene un numero di dati pari, invece, ci sono due valori centrali e di solito come mediana viene considerata la media aritmetica di questi.

5 – è il valore centrale in 4 5 5 6 7



Statistica descrittiva – alcune definizioni

Oltre alla mediana, che divide a metà un insieme di dati ordinati, vengono usati anche altri indici che dividono tale insieme in determinate percentuali detti **quantili**, **quartili** e **percentili**.

PERCENTILI: sono un caso particolare dei quantili e, come si intuisce dal nome, dividono l'insieme di dati ordinati in 100 parti.

- il 1° percentile lascia alla sua sinistra un centesimo (1%) degli elementi dell'insieme ordinato;
- il 10° percentile lascia alla sua sinistra il 10% degli elementi;
- il 50° percentile (che coincide con la mediana) lascia alla sua sinistra il 50% degli elementi;
- ...

QUARTILI: questi si ottengono dividendo l'insieme di dati ordinati in 4 parti uguali.

- il **primo quartile** (che coincide con il 25-esimo percentile) è il valore che lascia alla sua sinistra il 25% degli elementi;
- il **secondo quartile** (che coincide con la mediana e con il 50-esimo percentile) è il valore che lascia alla sua sinistra il 50% dei dati;
- il **terzo quartile** (che coincide con il 75-esimo percentile) è il valore che lascia il 75% degli elementi a sinistra e il 25% a destra.



Box plot

What?

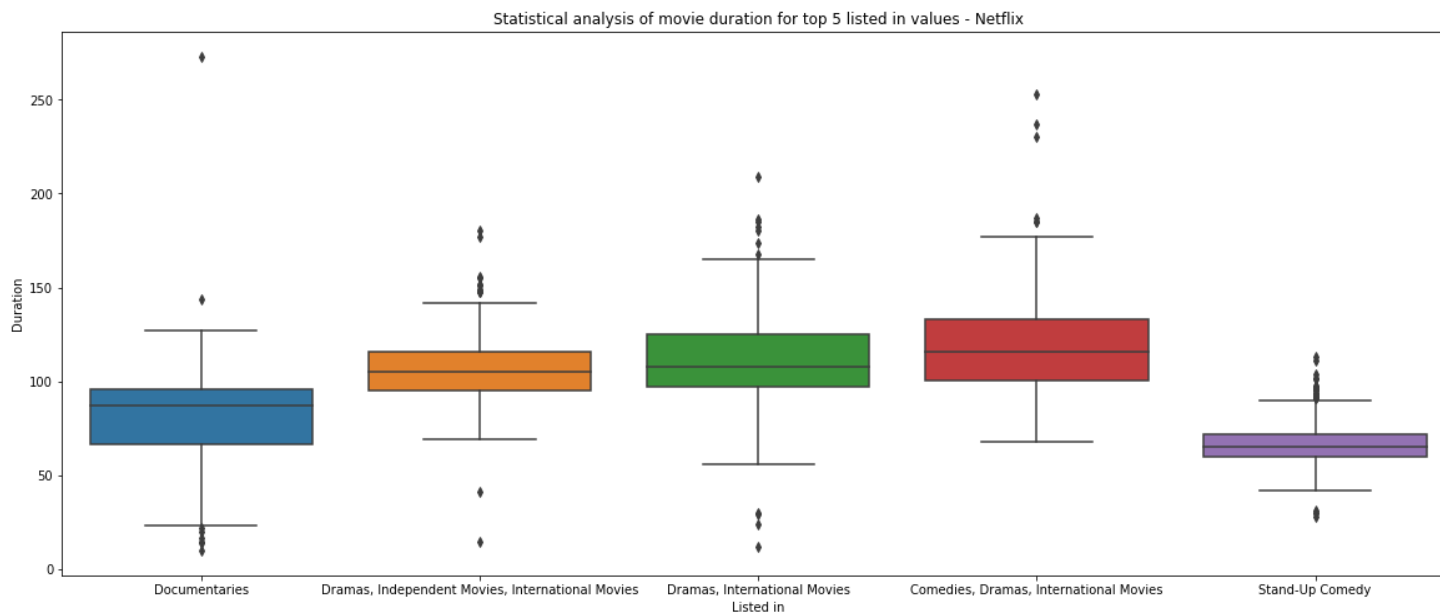
- N quantitative attributes (oppure 1 quantitative attribute ed 1 categorical key);

Why?

- Visualizzare distribuzioni;
- Identificare outliers, valori estremi, range etc.;

Remarks

- Il colore può codificare un categorical attribute aggiuntivo;
- Possibile effettuare raggruppamenti.



Qualitative Attributes

Bar plot
Multi-set bar plot
Pie chart
Word Cloud
...



Bar plot

What?

- 1 quantitative attribute;
- 1 categorical key;

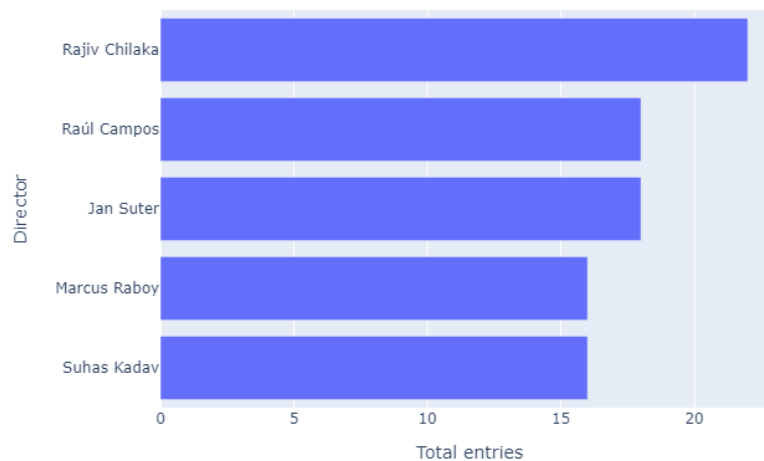
Why?

- Confrontare/evidenziare valori;
- Identificare valori estremi;

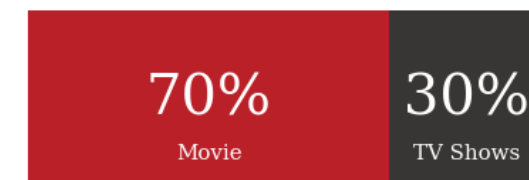
Remarks

- Fino a ~100 barre;
- Keys vs valori ordinati;
- Non adatto per visualizzare trends.

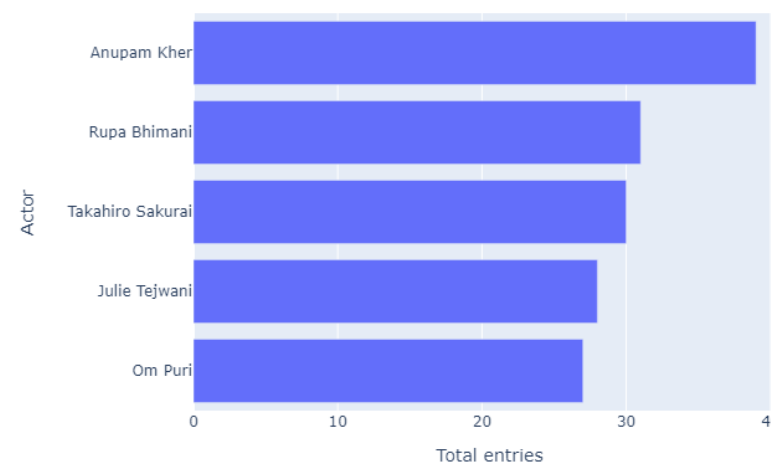
Top 5 Directors - Netflix



Movies & TV Shows distribution



Top 5 Actors - Netflix



Multi-set bar plot

What?

- 1 quantitative attribute;
- 2 categorical keys;

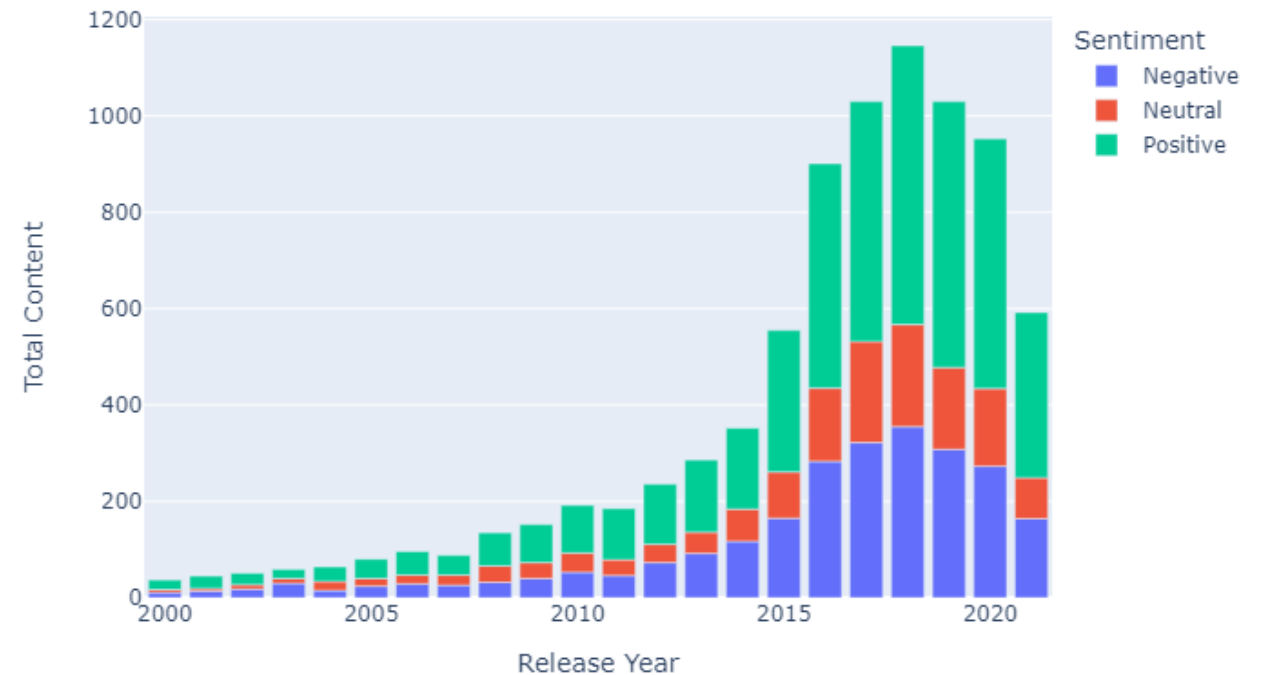
Why?

- Confrontare valori;
- Identificare patterns;

Remarks

- Visualizzare fino a ~100 barre;
- Riuscire a raggruppare/confrontare items, patterns.

Sentiment of contents - Netflix



2 Categorical Keys

Heatmap

...



Definizione

MATRICE DI CONFUSIONE: è un metodo per visualizzare le performance di un algoritmo rispetto ad un problema di classificazione dove gli outputs possono essere due o più classi.

Nel caso di problema di classificazione binario (due classi in outputs) la matrice di confusione sarà composta da quattro elementi: **True Positive (TP)**, **False Positive (FP)**, **False Negative (FN)**, **True Negative (TN)**.

		Actual Values	
		Positive (1)	Negative (0)
Predicted Values	Positive (1)	TP	FP
	Negative (0)	FN	TN

Inoltre, la matrice di confusione è estremamente comoda per calcolare *Precision*, *Recall*, *Accuratezza*, ... (se ne parlerà nelle prossime lezioni).



Heatmap

What?

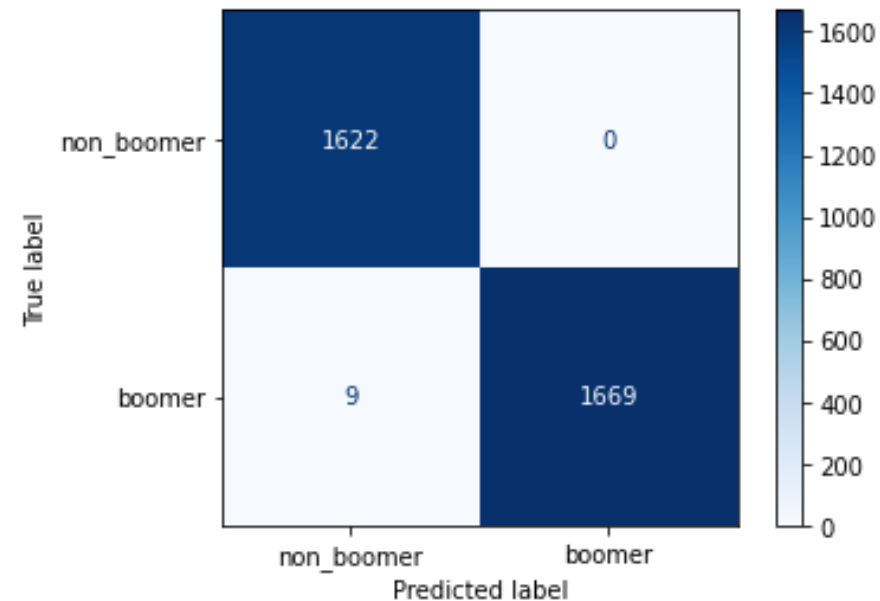
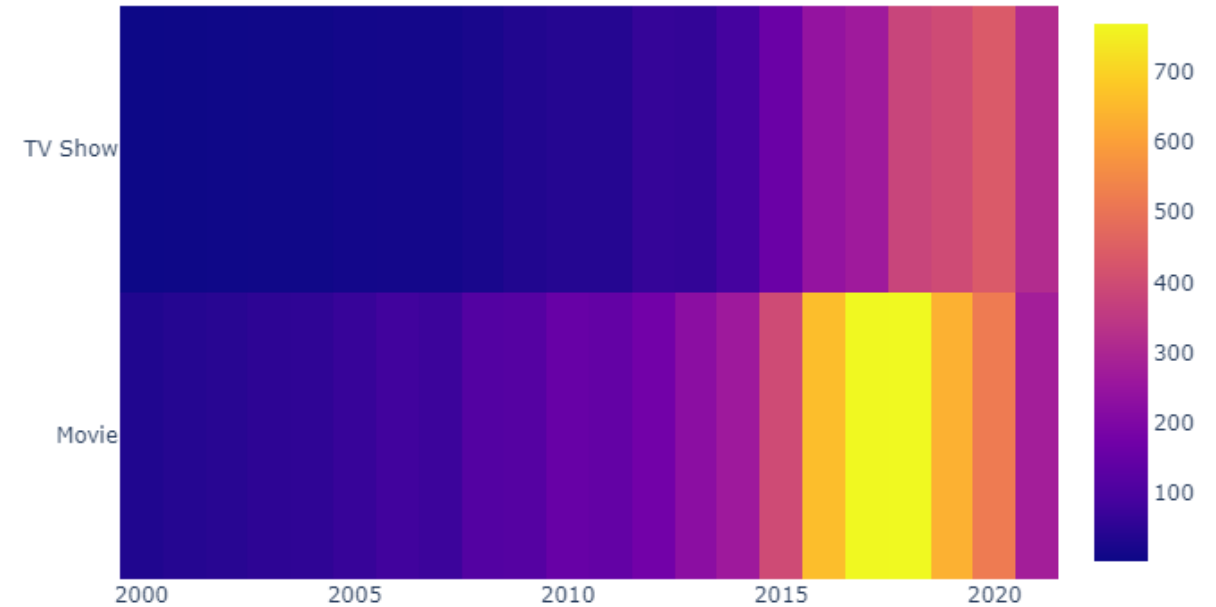
- 2 categorical key;
- 1 quantitative attribute;

Why?

- Visualizzare correlazioni;
- Identificare patterns, outliers;
- Confusion matrix for classification result visualization;

Remarks

- Fino a ~1M di items;
- L'ordine delle keys influisce la visibilità dei patterns.



For dealing with time

Line graph
Stacked area graph
...



Line graph

What?

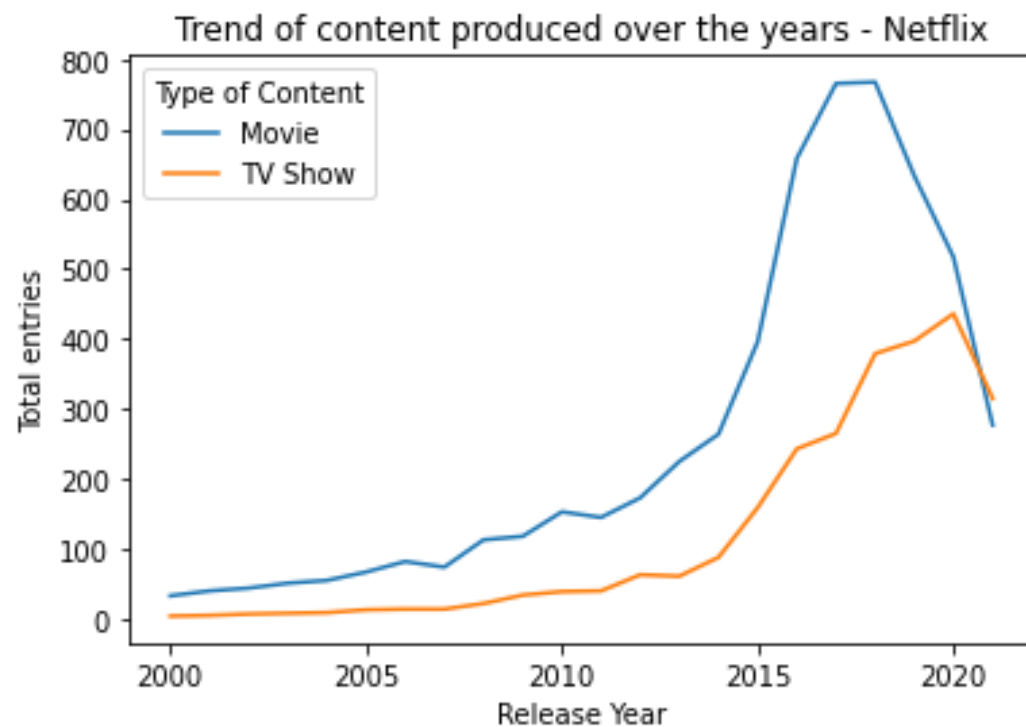
- 1 ordered key -> time;
- 1 quantitative attribute;

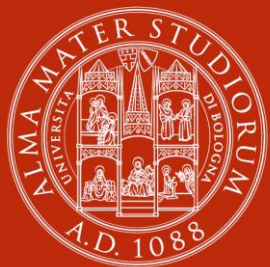
Why?

- Identificare e confrontare trends;

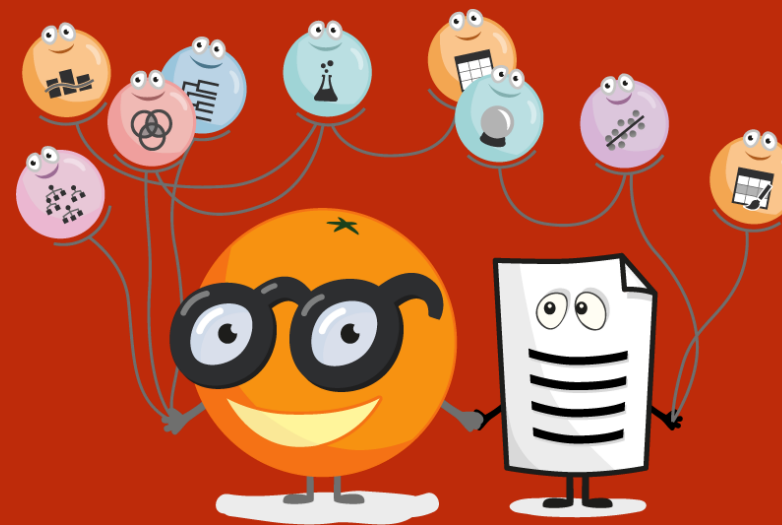
Remarks

- Fino a 10-20 linee;
- Il colore può codificare un categorical attribute additivo.





ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

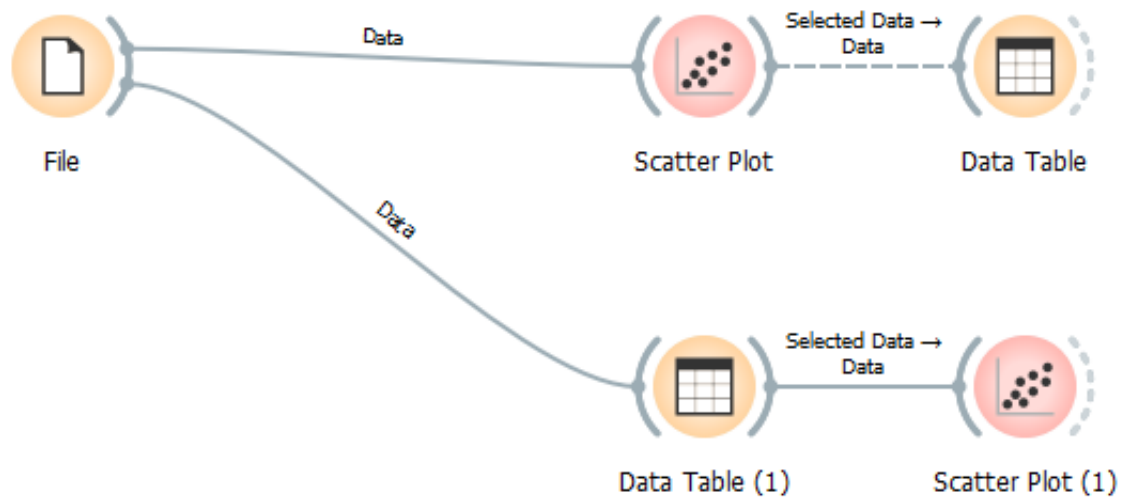


Data Visualization & Orange

Alessia Angeli

Studente di dottorato in Data Science and Computation
Dipartimento di Informatica – Scienza e Ingegneria

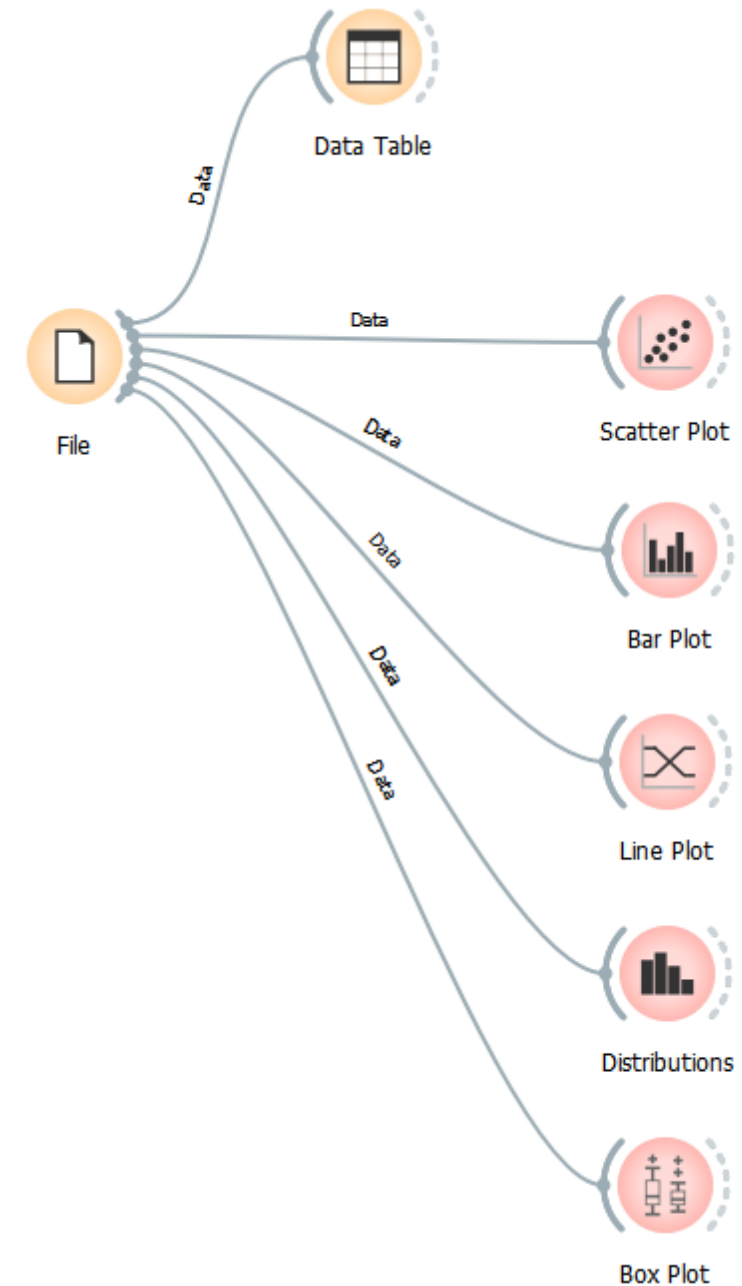
Orange – Visualization 1



1. DATA → VISUALIZE DATA and SELECT → TABLE OF SELECTED DATA
2. DATA → TABLE OF DATA and SELECT → VISUALIZE SELECTED DATA – **scatter plot**

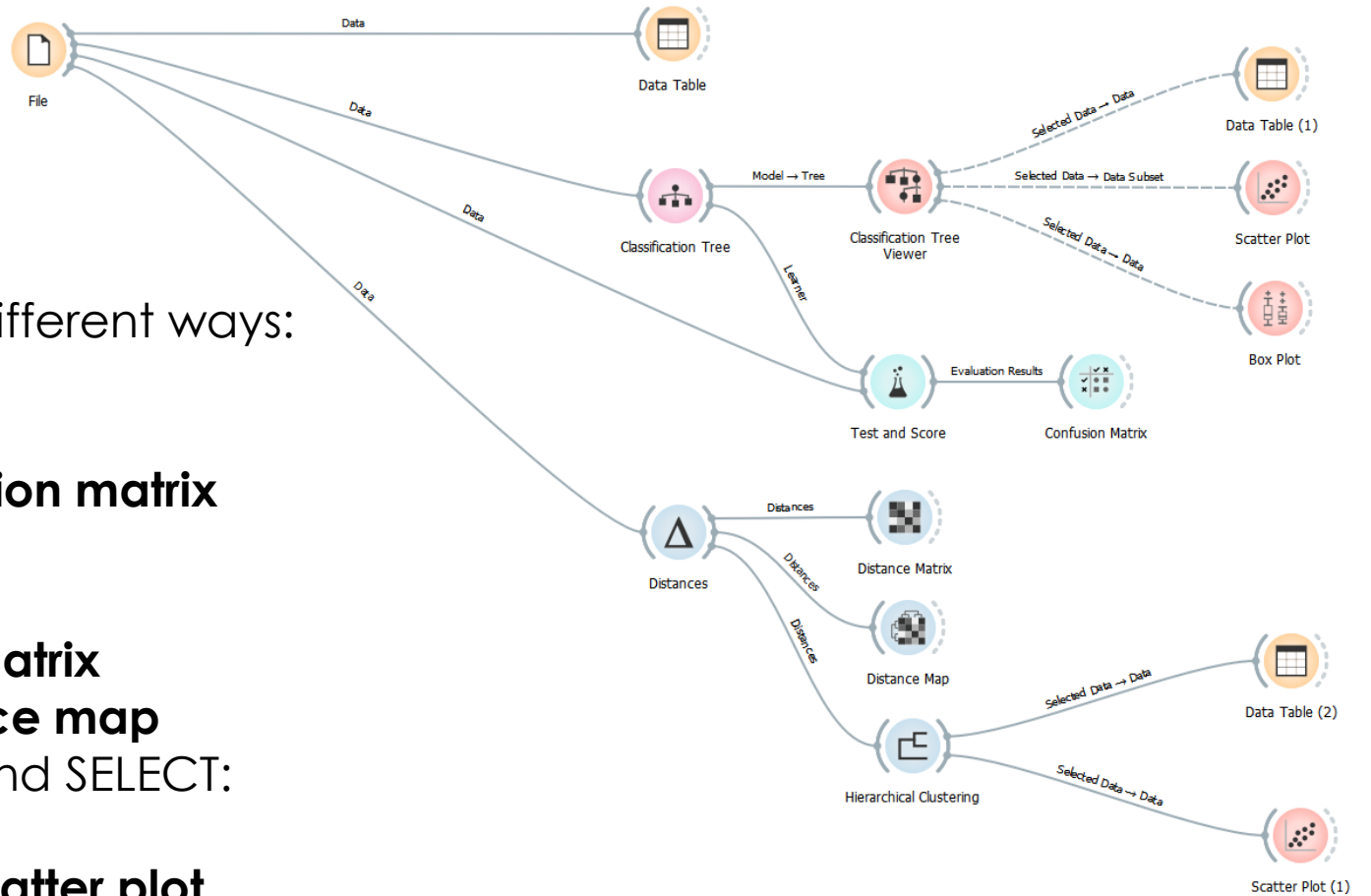
Orange – Visualization 2

1. DATA → TABLE OF DATA
2. DATA → VISUALIZE DATA in different ways:
 - **scatter plot**
 - **bar plot**
 - **line plot**
 - **distribution (histogram)**
 - **box plot**

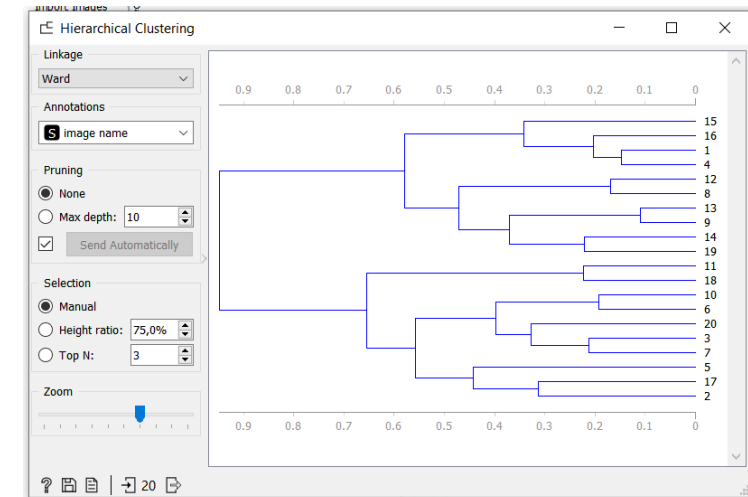
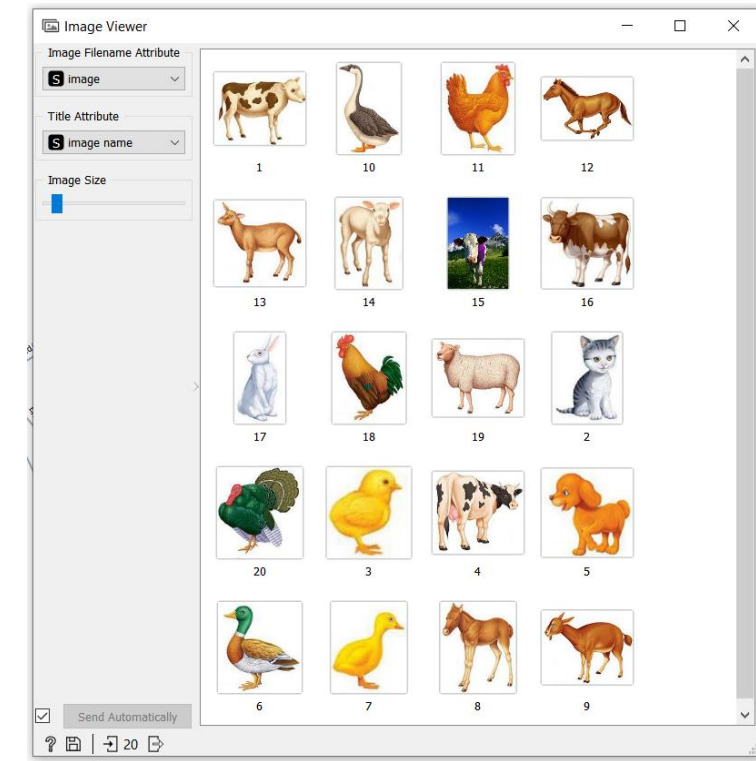
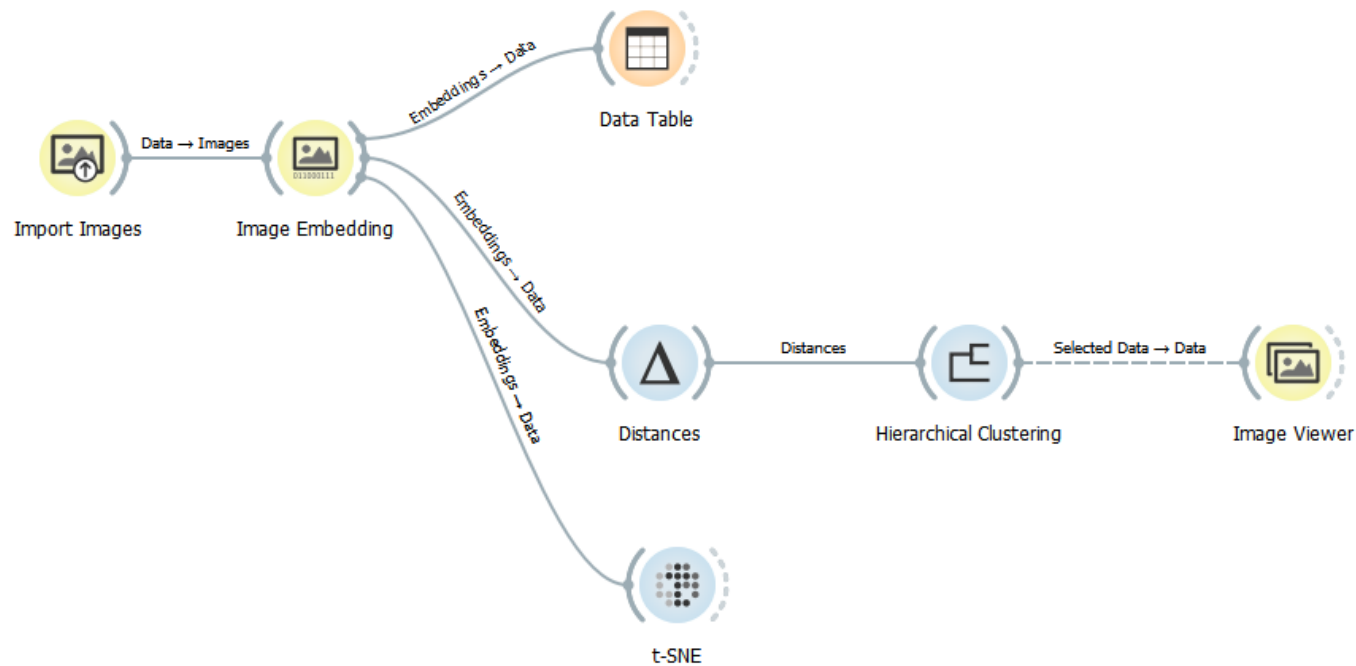
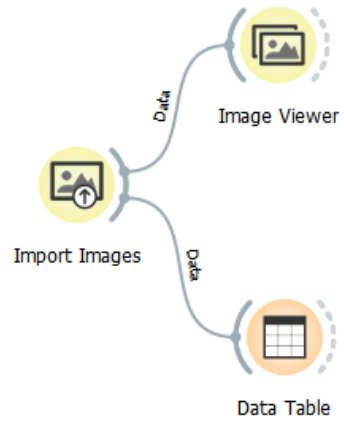


Orange – Visualization 3

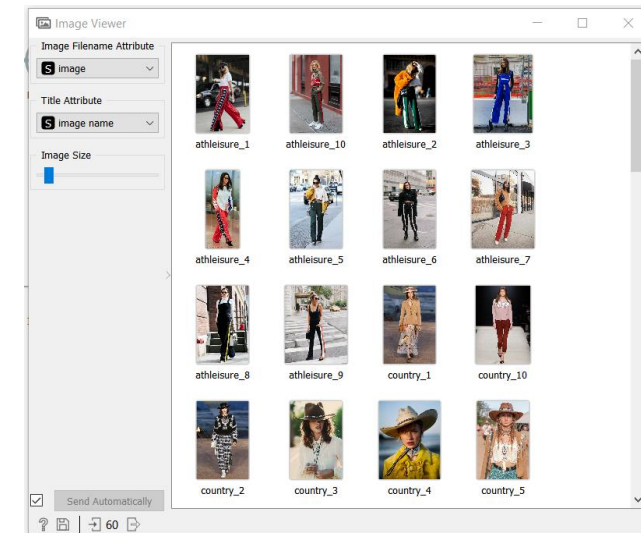
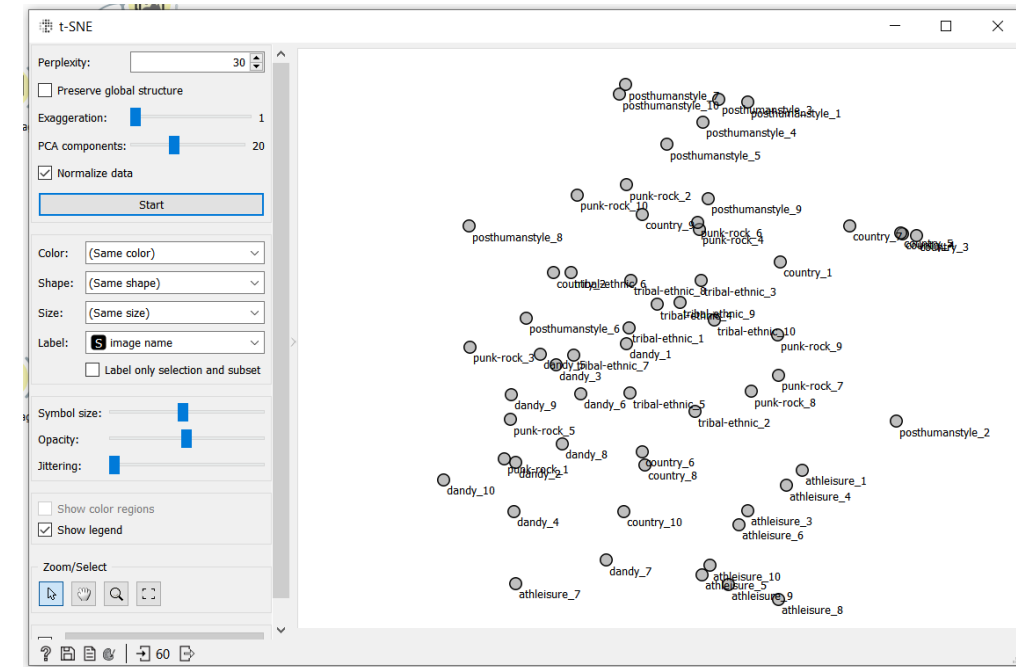
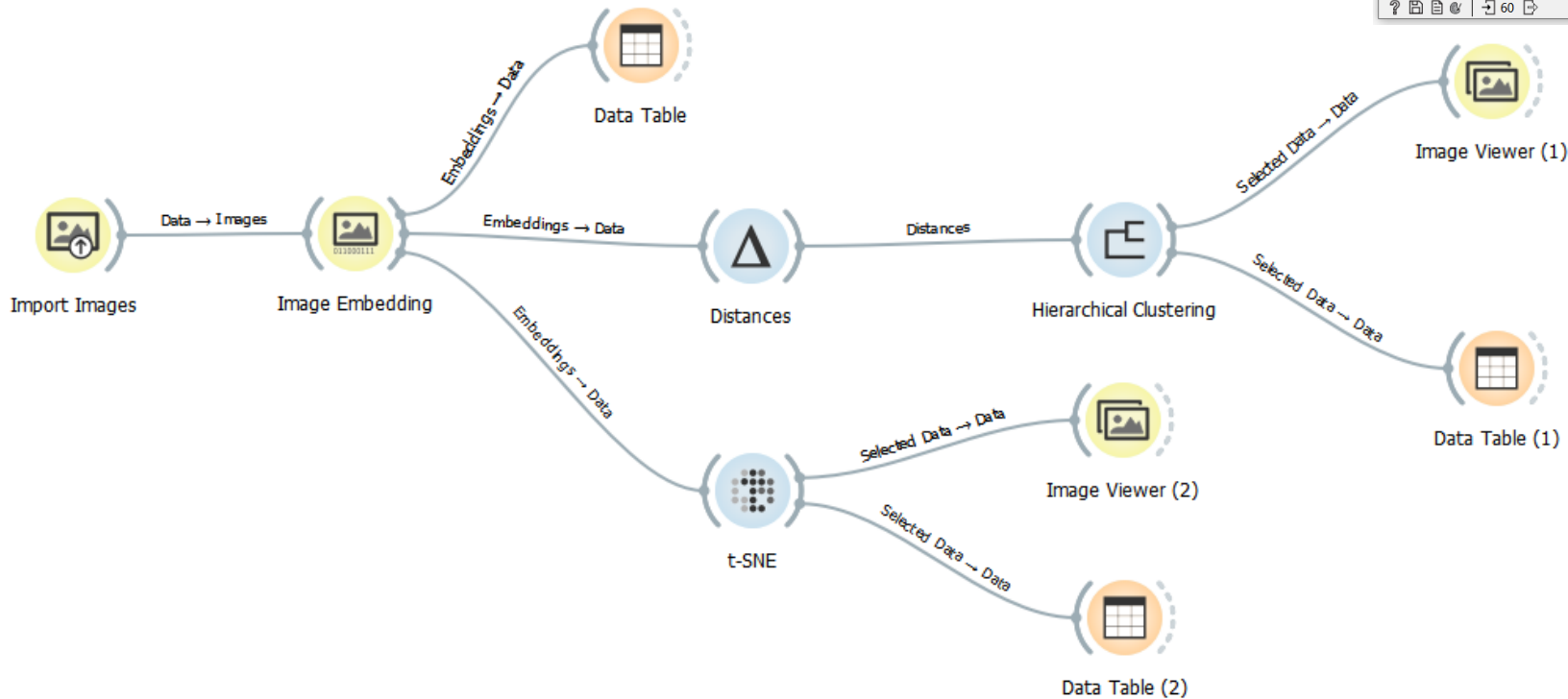
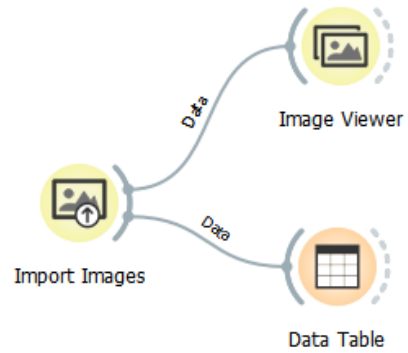
1. DATA → TABLE OF DATA
2. DATA → **MODEL (classification tree)**:
 - VISUALIZE MODEL and SELECT:
 - TABLE OF SELECTED DATA
 - VISUALIZE SELECTED DATA in different ways:
 - **scatter plot**
 - **box plot**
 - MODEL TEST AND SCORE → **confusion matrix**
3. DATA → **COMPUTE DISTANCES**:
 - VISUALIZE DISTANCES – **distance matrix**
 - VISUALIZE DISTANCE MAP – **distance map**
 - **MODEL (hierarchical clustering)** and SELECT:
 - TABLE OF SELECTED DATA
 - VISUALIZE SELECTED DATA – **scatter plot**



Orange – Domestic Animals Example



Orange – Fashion Style Example





ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

Alessia Angeli

Dipartimento di Informatica – Scienza e Ingegneria

alessia.angeli2@unibo.it

www.unibo.it