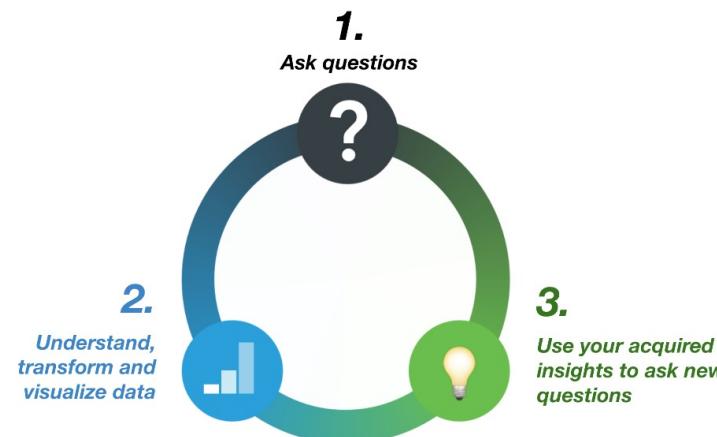

Tutorial

Exploratory Data Analysis

and Tools – Orange and Python

Exploratory Data Analysis

- EDA is an **iterative** cycle:
 - Generate questions about your data:
 - Search for answers by visualising, transforming, and modelling your data
 - Use what you learn to refine your questions and/or generate new questions



<https://duo.com/labs/research/gamifying-data-science-education>

We are considering the popular data set “iris”

-The Iris flower data set or Fisher's Iris data set is a multivariate data set introduced by the British statistician and biologist Ronald Fisher in his 1936 paper The use of multiple measurements in taxonomic problems as an example of linear discriminant analysis.^[1](Wikipedia)



□UCI Machine Learning Repository

Iris Data Set

Download: [Data Folder](#), [Data Set Description](#)

Abstract: Famous database; from Fisher, 1936

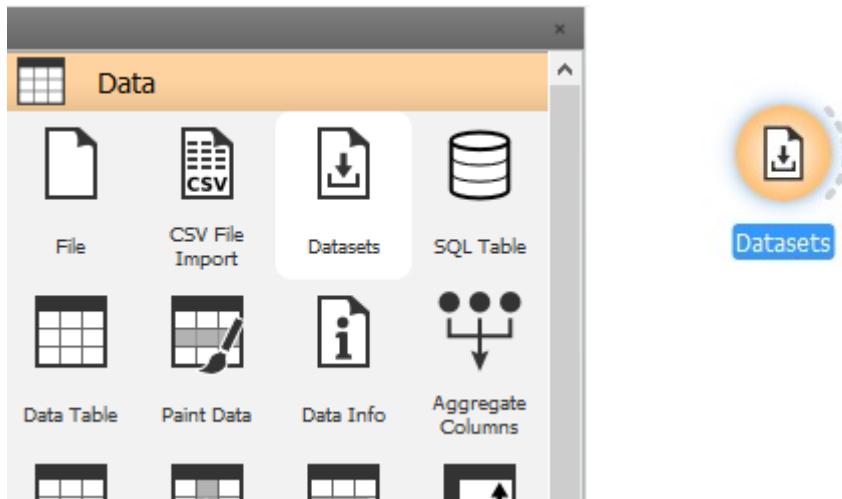


Data Set Characteristics:	Multivariate	Number of Instances:	150	Area:	Life
Attribute Characteristics:	Real	Number of Attributes:	4	Date Donated	1988-07-01
Associated Tasks:	Classification	Missing Values?	No	Number of Web Hits:	4128602

[1]. R. A. Fisher (1936). "The use of multiple measurements in taxonomic problems". *Annals of Eugenics*. 7 (2): 179–188. doi:[10.1111/j.1469-1809.1936.tb02137.x](https://doi.org/10.1111/j.1469-1809.1936.tb02137.x). hdl:[2440/15227](https://hdl.handle.net/2440/15227).

Let's start with Orange

- Load the data set



Title	Size	Instances	Variables	Target	Tags
Iris	4.5 KB	150	5	C categorical	biology
Breast Cancer and Docetaxel Treatment	1.8 MB	24	9486	C categorical	biology
Smoking effect on B lymphocytes	1.8 MB	79	3000	C categorical	genomics
Bone marrow mononuclear cells with AML	582.0 KB	96	1000	C categorical	genomics
HDI	65.1 KB	188	66	N numeric	economy, geo
TKI resistance	1.2 MB	280	467	C categorical	spectral
Abalone	187.5 KB	4177	8	N numeric	biology
Adult	4.1 MB	32561	15	C categorical	economy
Roman Amphorae	23.7 KB	164	16	C categorical	archaeology, image analytics
Attrition - Predict	838 bytes	3	18	C categorical	economy, synthetic, education
Attrition - Train	182.2 KB	1470	18	C categorical	economy, synthetic
Auto MPG	17.3 KB	398	9	N numeric	

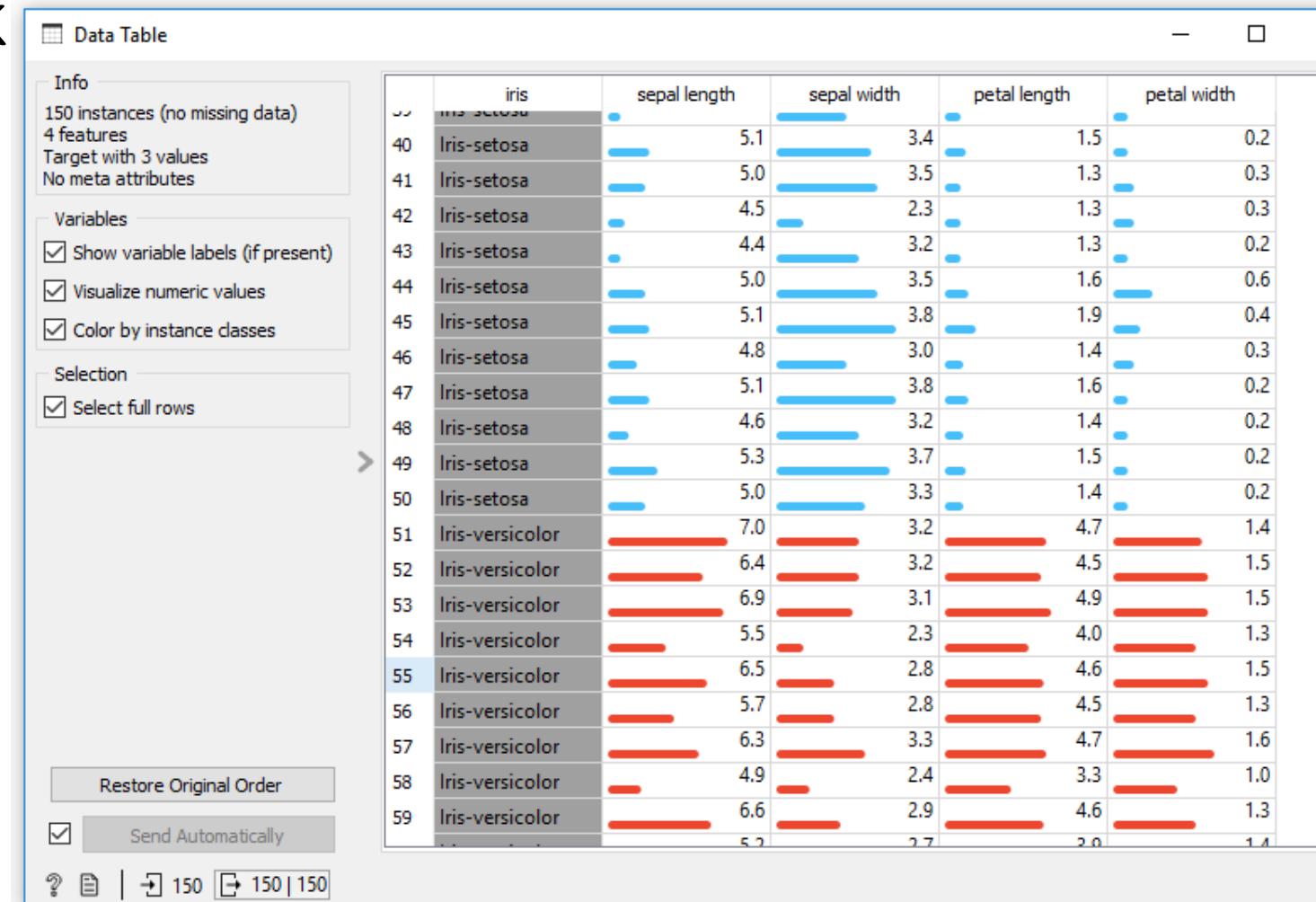
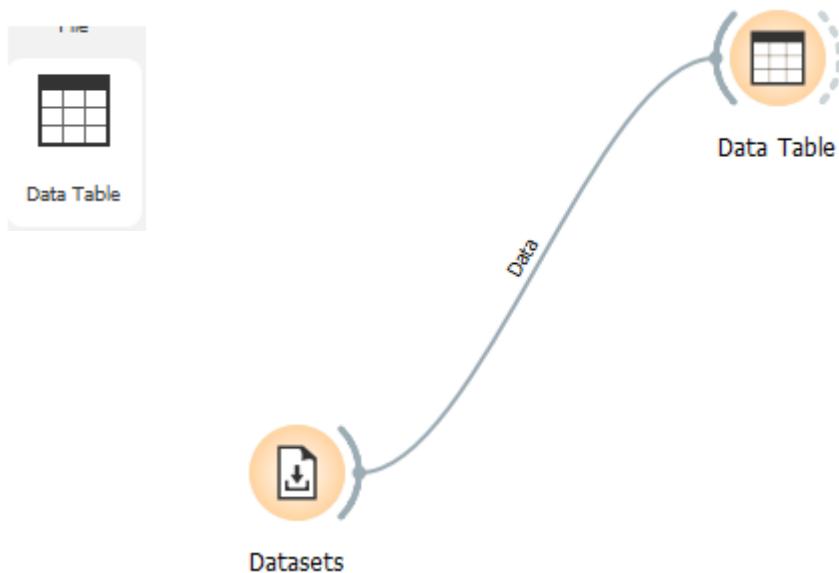
Description
Iris (1936), from [UCI ML Repository](#)
The Iris flower data set or Fisher's Iris data set was introduced by the British statistician and biologist Ronald Fisher in his 1936 paper as an example of linear discriminant analysis. The data on length and width of petal and sepal leaves was actually collected by American botanist Edgar Anderson to quantify the morphologic variation of Iris flowers of three related species.

See Also
[Scatter Plots: the Tour](#),
[All | See is Silhouette](#),
References
R. A. Fisher (1936) The use of multiple measurements in taxonomic problems. Annals of Eugenics 7(2):179–188

❑ What to notice?

- Type of target is categorical → classification
- Data size is small → might need cross validation

Orange EDA: A first look



What to notice?

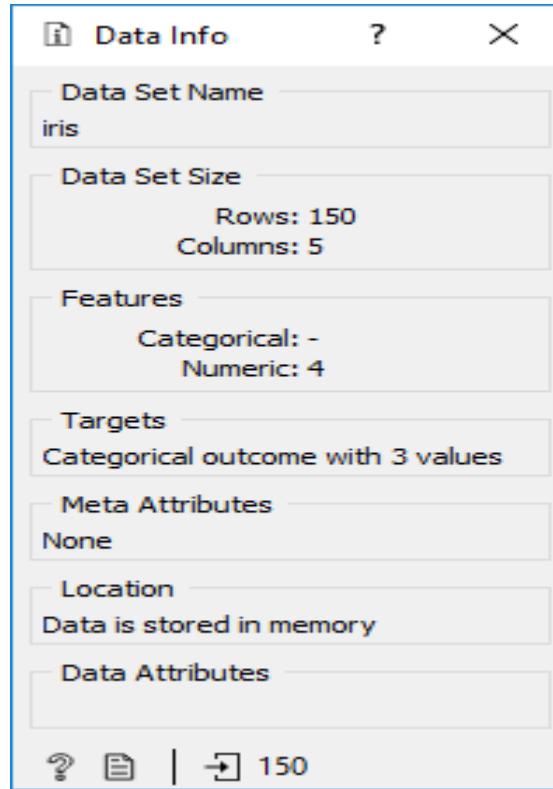
- Variable scale is different (e.g. sepal length is the widest and petal width is the least)
→ might need normalization

Selected Data: iris: 150 instances, 5 variables
Features: 4 numeric (no missing values)
Target: categorical
Data: iris: 150 instances, 6 variables
Features: 4 numeric (no missing values)
Target: categorical
Metas: categorical

Orange EDA: A first look



Data Info

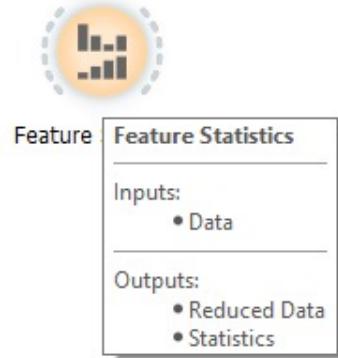


◻ What to notice?

- Type of target is categorical → classification
- Data size is small → might need cross validation

Orange EDA:

What are the stats of the variables?



❑ What to notice?

- Centre, spread, no missing values of variables
- Distributions of variables over classes
- Classes balance

Orange EDA:

What are the most important features?



Rank

Scoring Methods

- Information Gain
- Information Gain Ratio
- Gini Decrease
- ANOVA
- χ^2
- ReliefF
- FCBF

Select Attributes

- None
- All
- Manual
- Best ranked:

Send Automatically

Info. gain Gain ratio Gini ANOVA χ^2 ReliefF FCBF

	#	Info. gain	Gain ratio	Gini	ANOVA	χ^2	ReliefF	FCBF
N petal length		1.086	0.544	0.423	1179.034	98.946	0.368	1.542
N petal width		1.059	0.532	0.407	959.324	94.162	0.370	1.451
N sepal length		0.624	0.313	0.247	119.265	79.243	0.155	0.000
N sepal width		0.361	0.183	0.154	47.364	50.082	0.118	0.255



Rank

Data Table (1)

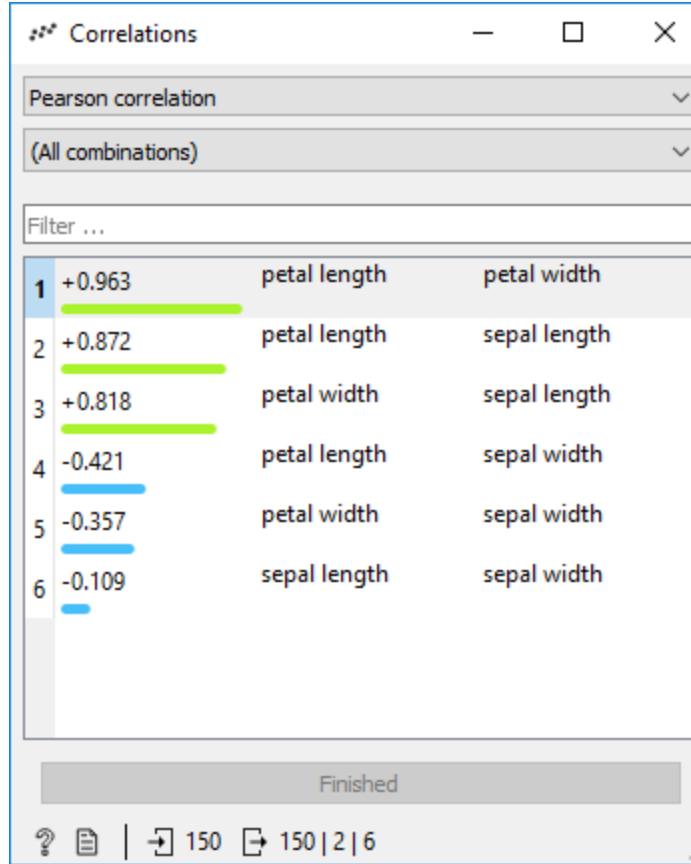
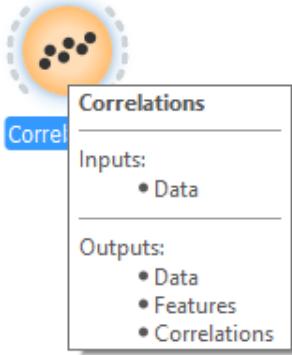
□ What to notice?

- Petal length seems the most important and sepal width is the least

→ Feature selection

Orange EDA:

What are the most important features?



❑ What to notice?

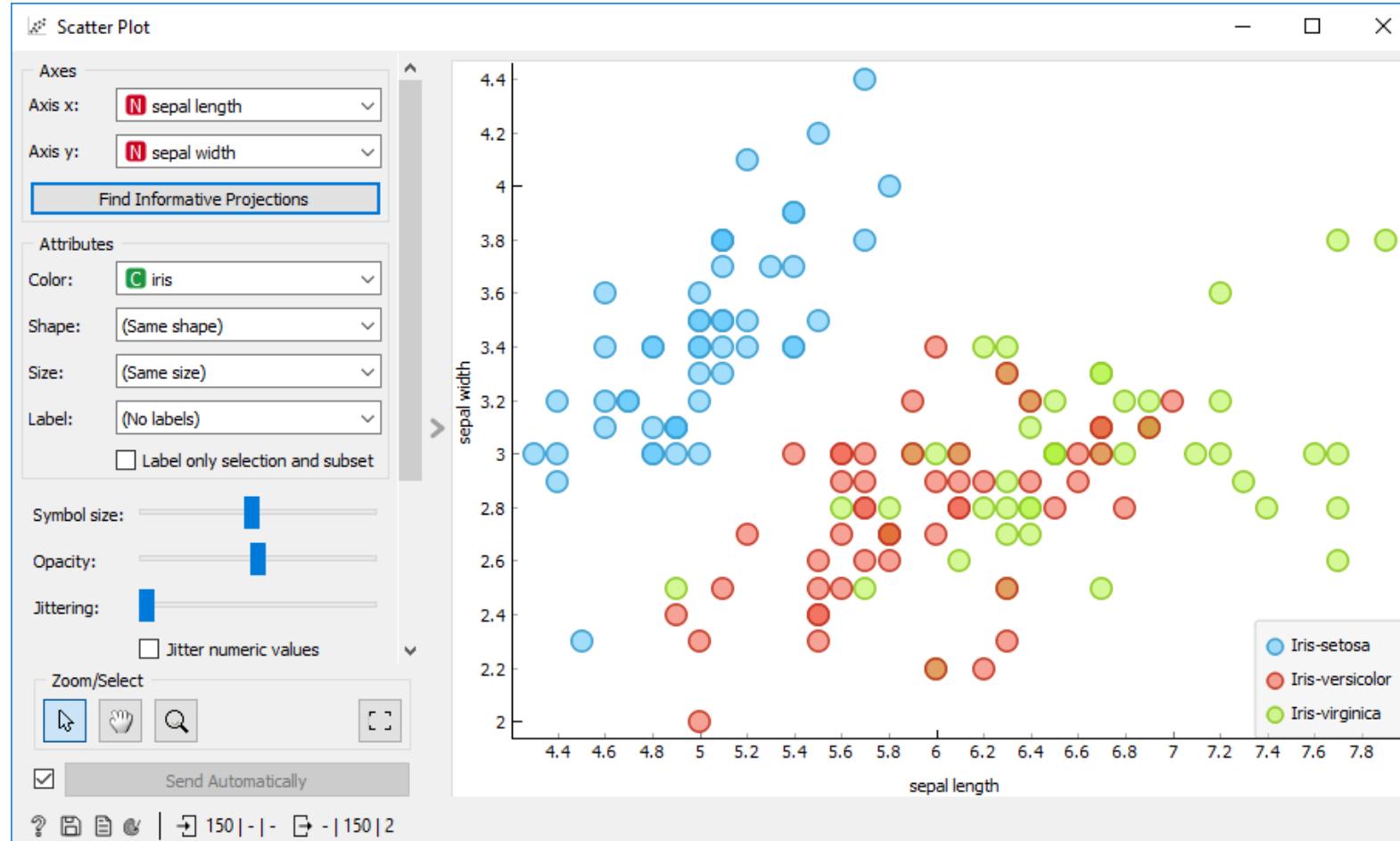
- [-1, 1], negative/positive, strong/weak...
 - Petal length and petal width look strongly correlated
- ➔ Feature selection

Orange EDA:

What is the relationship between two variables (e.g. the sepal length and width) per/regardless class?



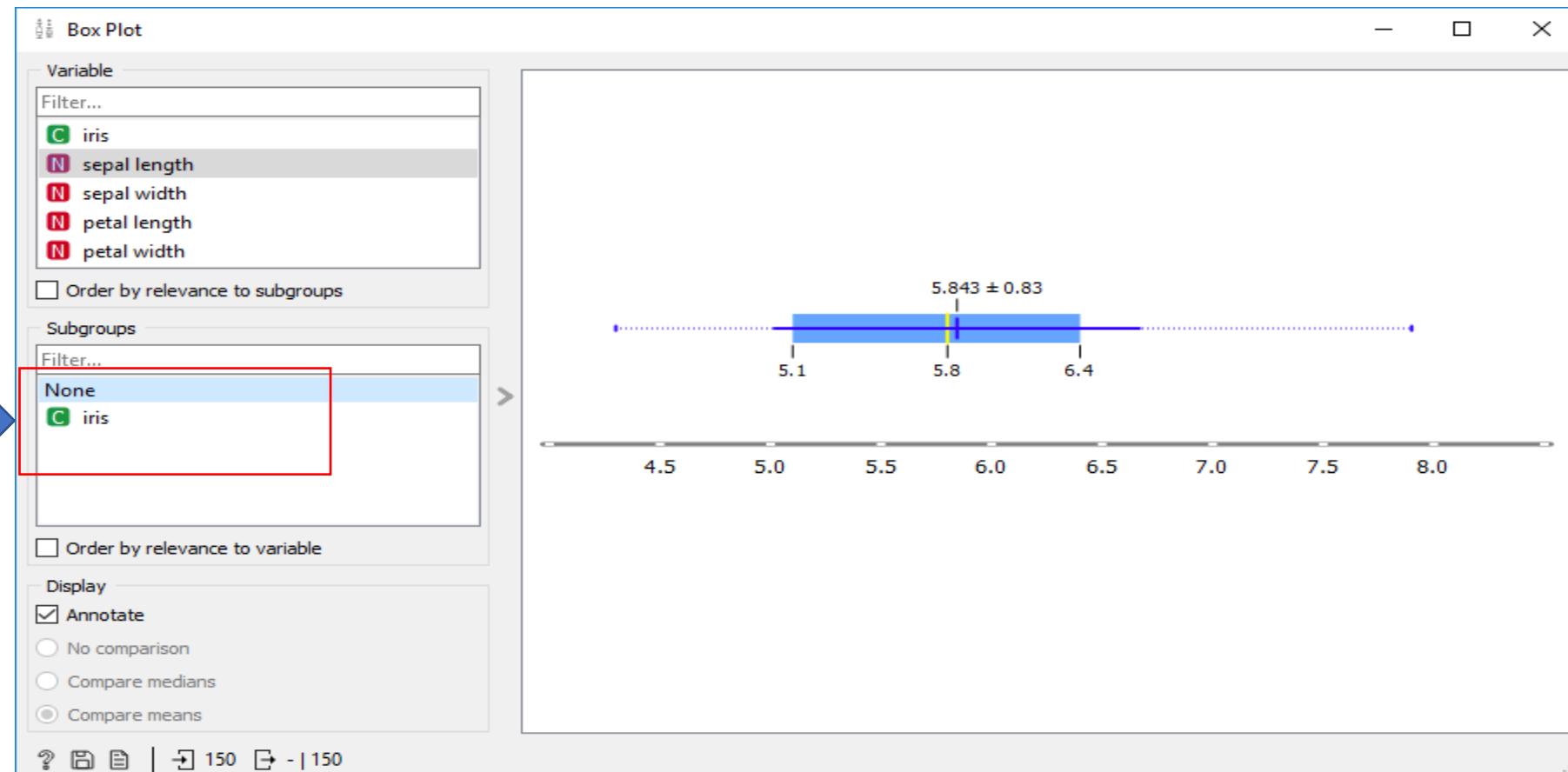
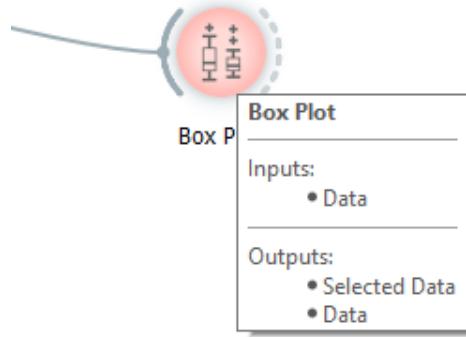
Scatter Plot



- Change variables
- What to notice?
- Compare with the correlation shown previously

Orange EDA:

How the values of a certain variable (e.g. sepal length) are distributed?

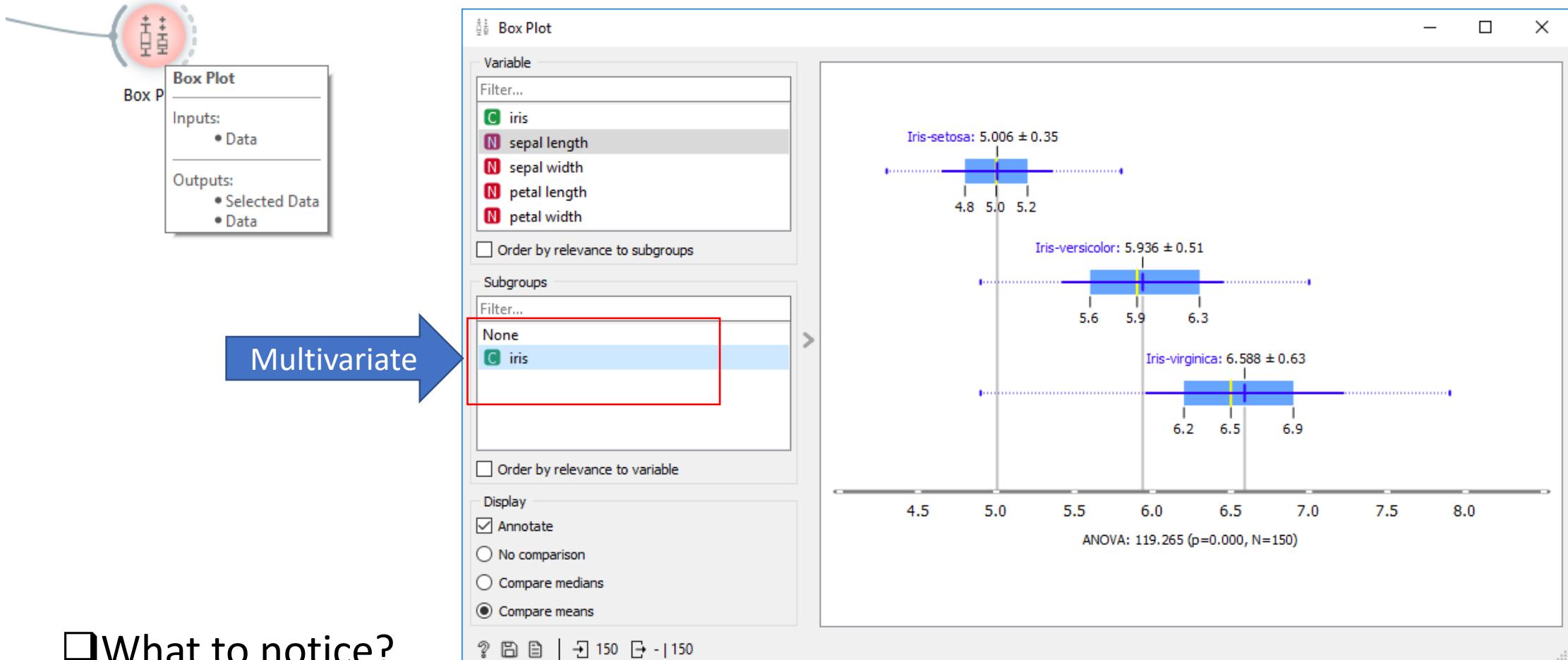


❑ What to notice?

- Graphical presentation for the stats

Orange EDA:

How the values of a certain variable (e.g. sepal length) are distributed per target class (iris species)?

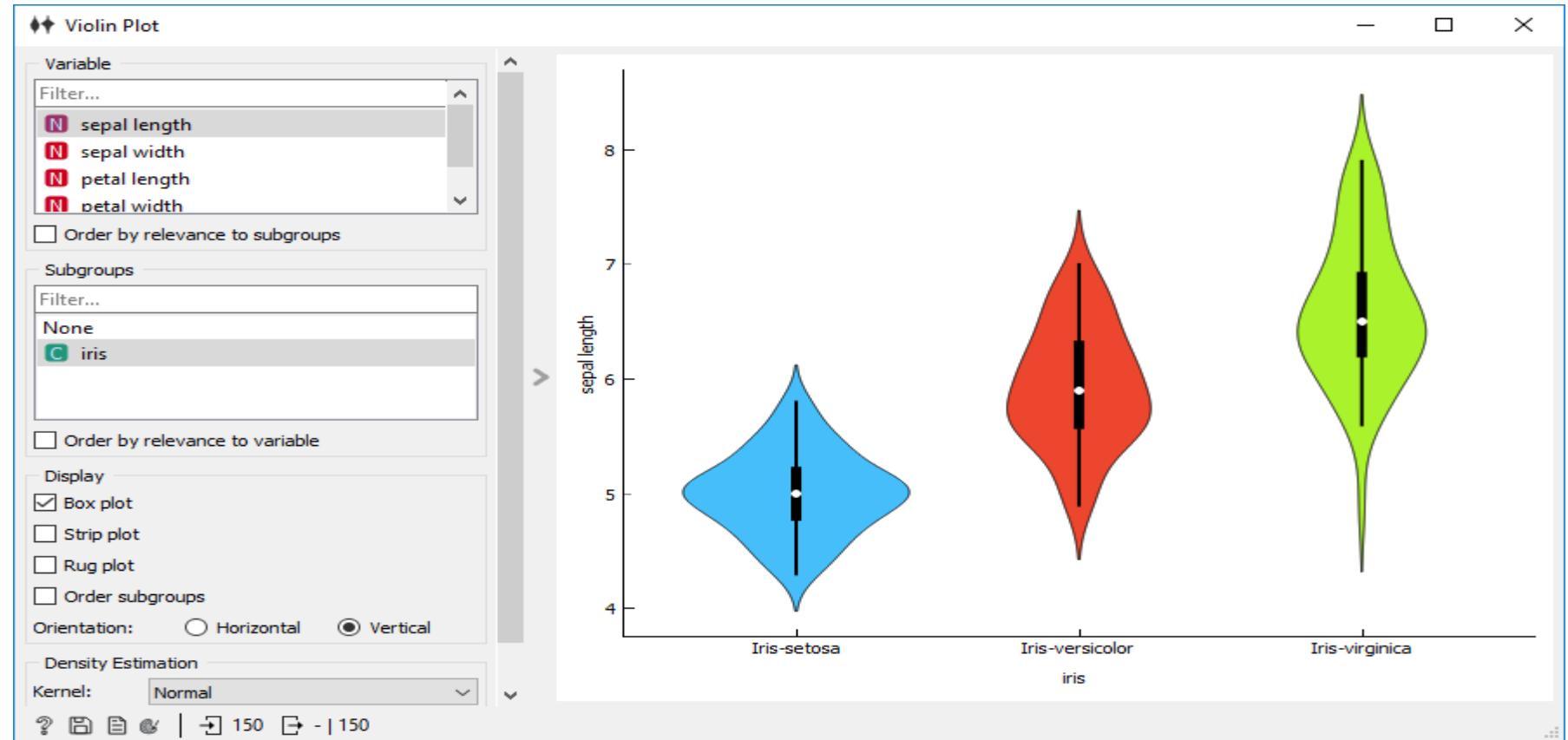
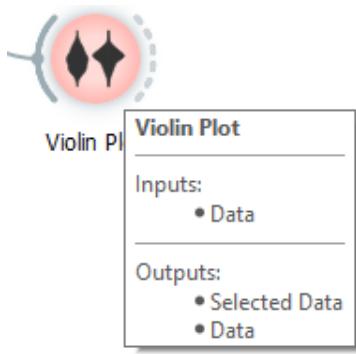


❑ What to notice?

- Graphical presentation for the stats per class
- Small sepal length → iris-setosa class

Orange EDA:

How the values of a certain variable (e.g. sepal length) are distributed per target class (iris species)?

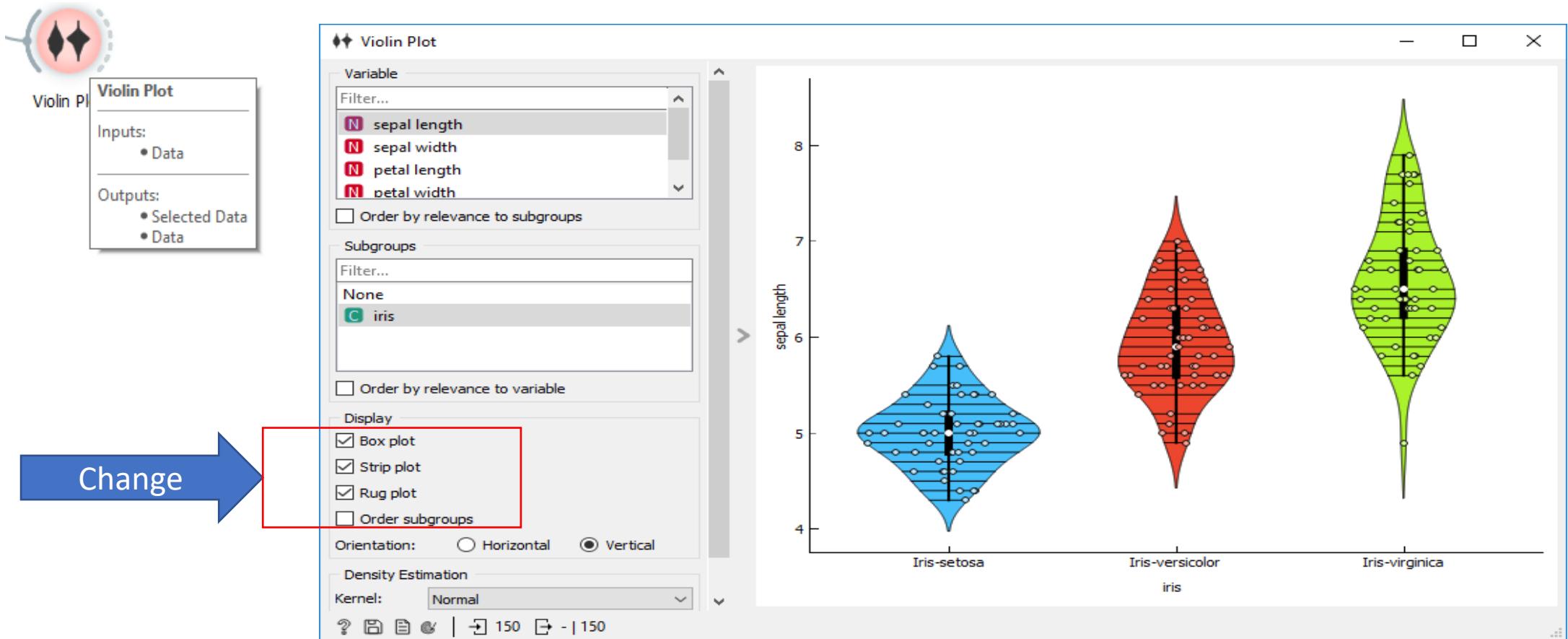


❑ What to notice?

- Similar to box plot but the density/frequency of the samples for variable values is visualized

Orange EDA:

How the values of a certain variable (e.g. sepal length) are distributed per target class (iris species)?

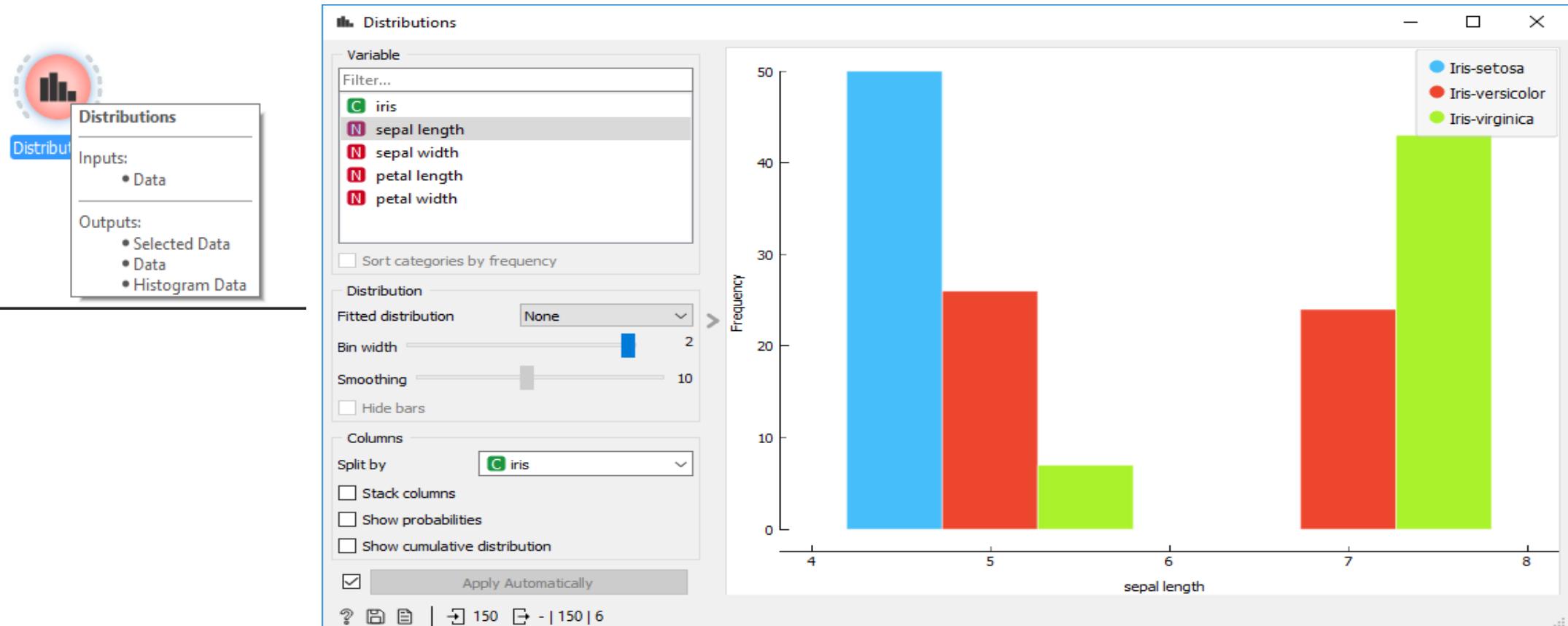


❑ What to notice?

- Show the points for clearer visualization

Orange EDA:

How the values of a certain variable (e.g. sepal length) are distributed per target class (iris species)?



❑ What to notice?

- Shorter sepal → Iris-setosa
- Longer sepal → more likely Iris-virginica

Orange EDA:

How the values of input variables are distributed w.r.t. another variable?



Bar Plot

