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

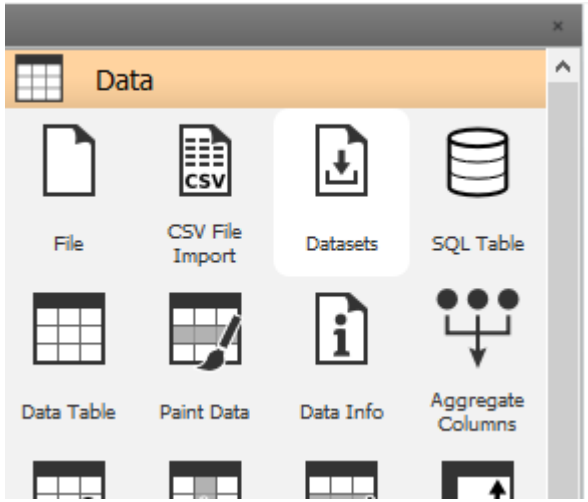
Abstract: Famous database; from Fisher, 1936

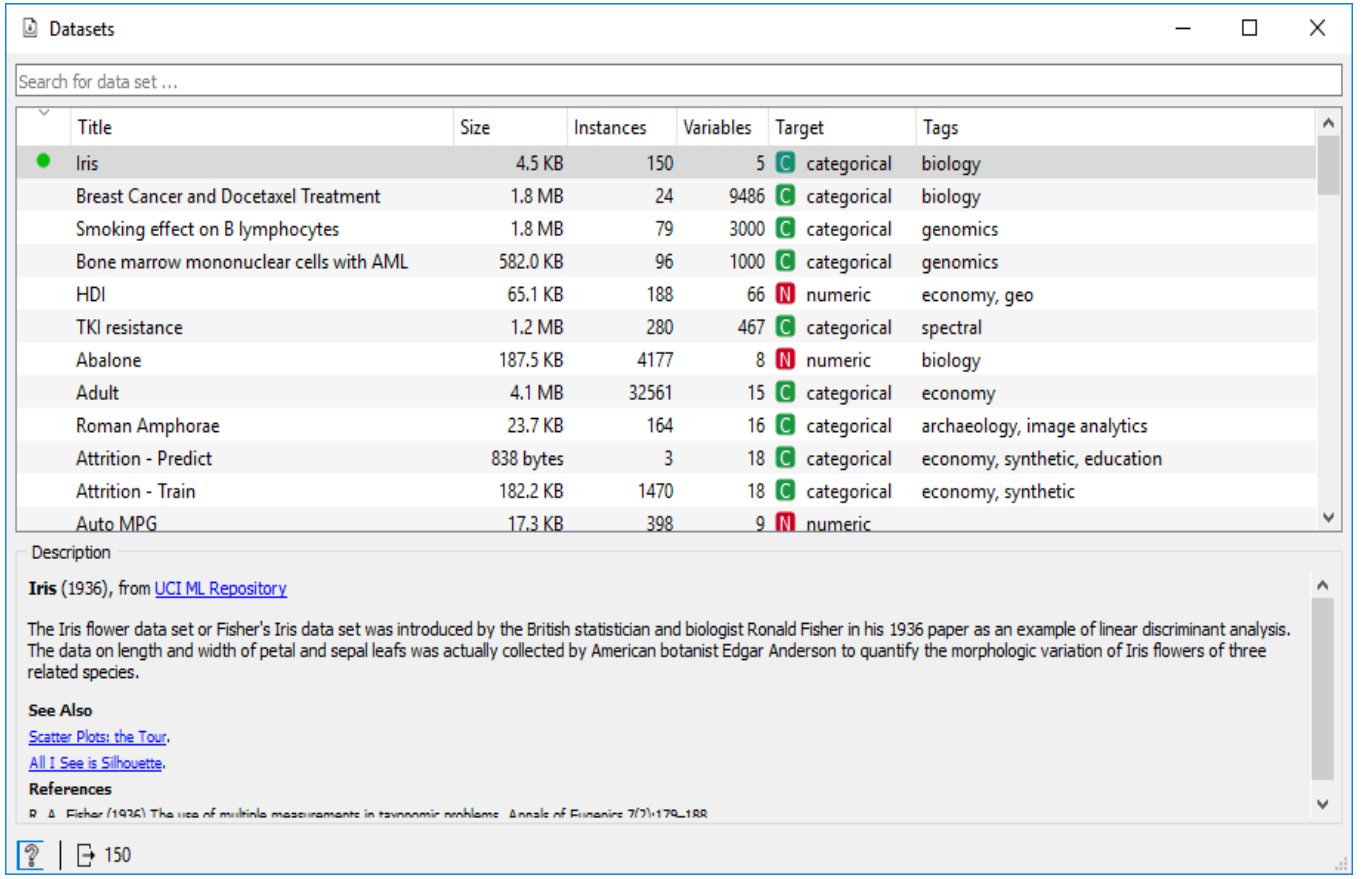
Data Set Characteristics:	Multivariate	**Number of Instances:**	150	**Area:**	Life
Attribute Characteristics:	Real	**Number of Attributes:**	4	**Date Donated**	1988-07-01
Associated Tasks:	Classification	**Missing Values?**	No	**Number of Web Hits:**	4128602

*[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. hdl:2440/15227.*

Let's start with Orange

- Load the data set

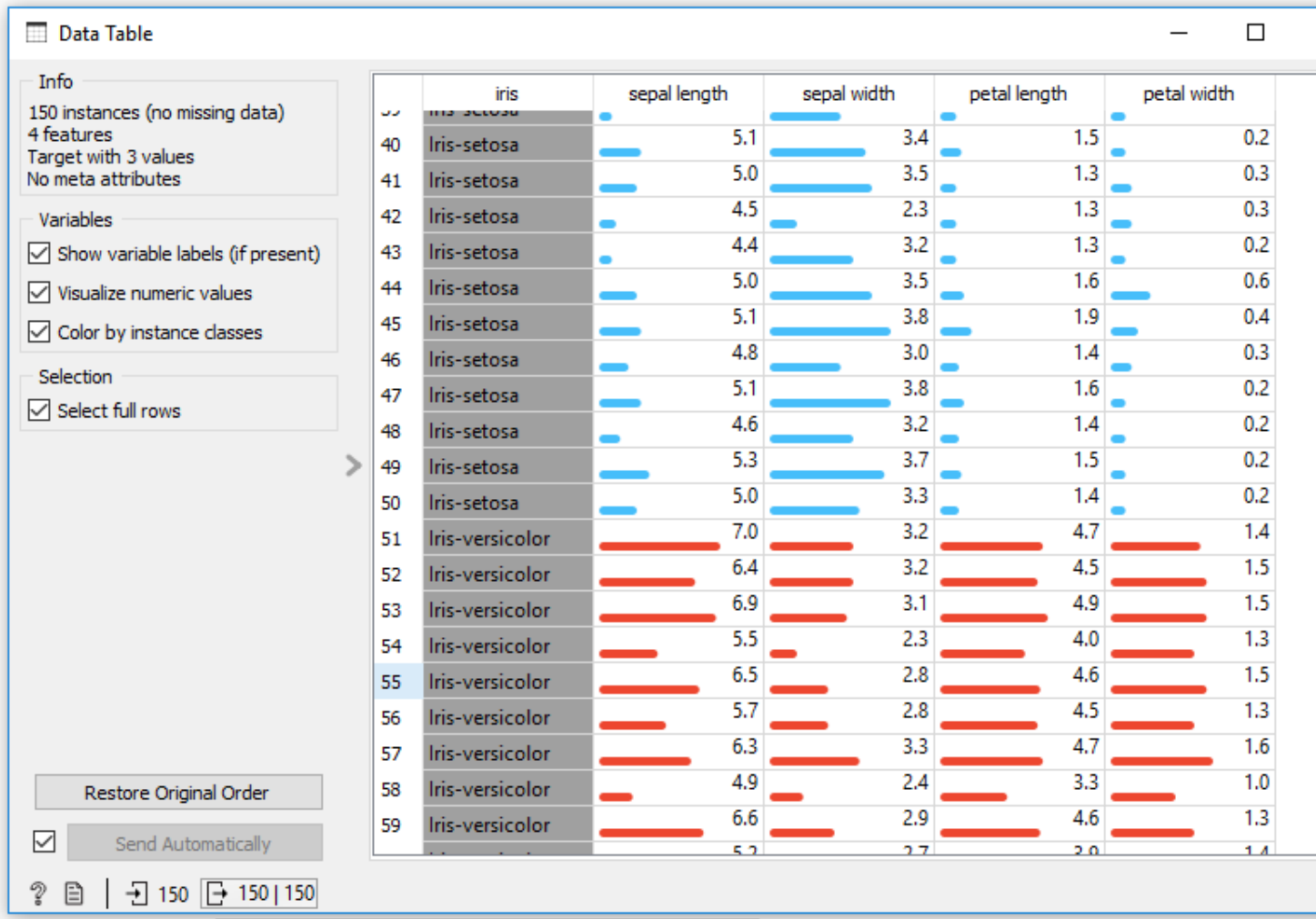




❑What to notice?

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

Orange EDA: A first look



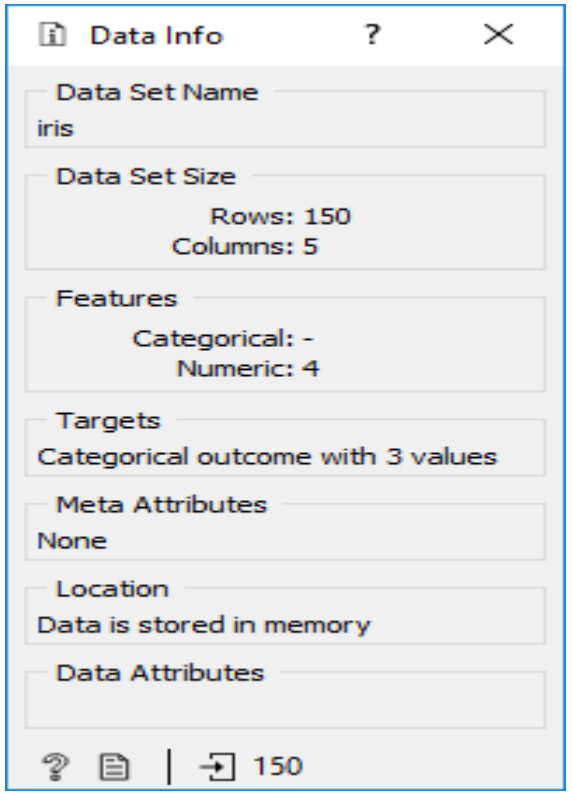
	iris	sepal length	sepal width	petal length	petal width
40	Iris-setosa	5.1	3.4	1.5	0.2
41	Iris-setosa	5.0	3.5	1.3	0.3
42	Iris-setosa	4.5	2.3	1.3	0.3
43	Iris-setosa	4.4	3.2	1.3	0.2
44	Iris-setosa	5.0	3.5	1.6	0.6
45	Iris-setosa	5.1	3.8	1.9	0.4
46	Iris-setosa	4.8	3.0	1.4	0.3
47	Iris-setosa	5.1	3.8	1.6	0.2
48	Iris-setosa	4.6	3.2	1.4	0.2
49	Iris-setosa	5.3	3.7	1.5	0.2
50	Iris-setosa	5.0	3.3	1.4	0.2
51	Iris-versicolor	7.0	3.2	4.7	1.4
52	Iris-versicolor	6.4	3.2	4.5	1.5
53	Iris-versicolor	6.9	3.1	4.9	1.5
54	Iris-versicolor	5.5	2.3	4.0	1.3
55	Iris-versicolor	6.5	2.8	4.6	1.5
56	Iris-versicolor	5.7	2.8	4.5	1.3
57	Iris-versicolor	6.3	3.3	4.7	1.6
58	Iris-versicolor	4.9	2.4	3.3	1.0
59	Iris-versicolor	6.6	2.9	4.6	1.3

❑What to notice?

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

➔ might need normalization

Orange EDA: A first look

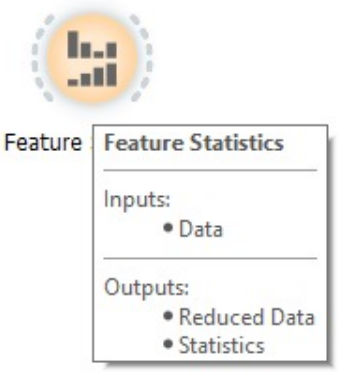


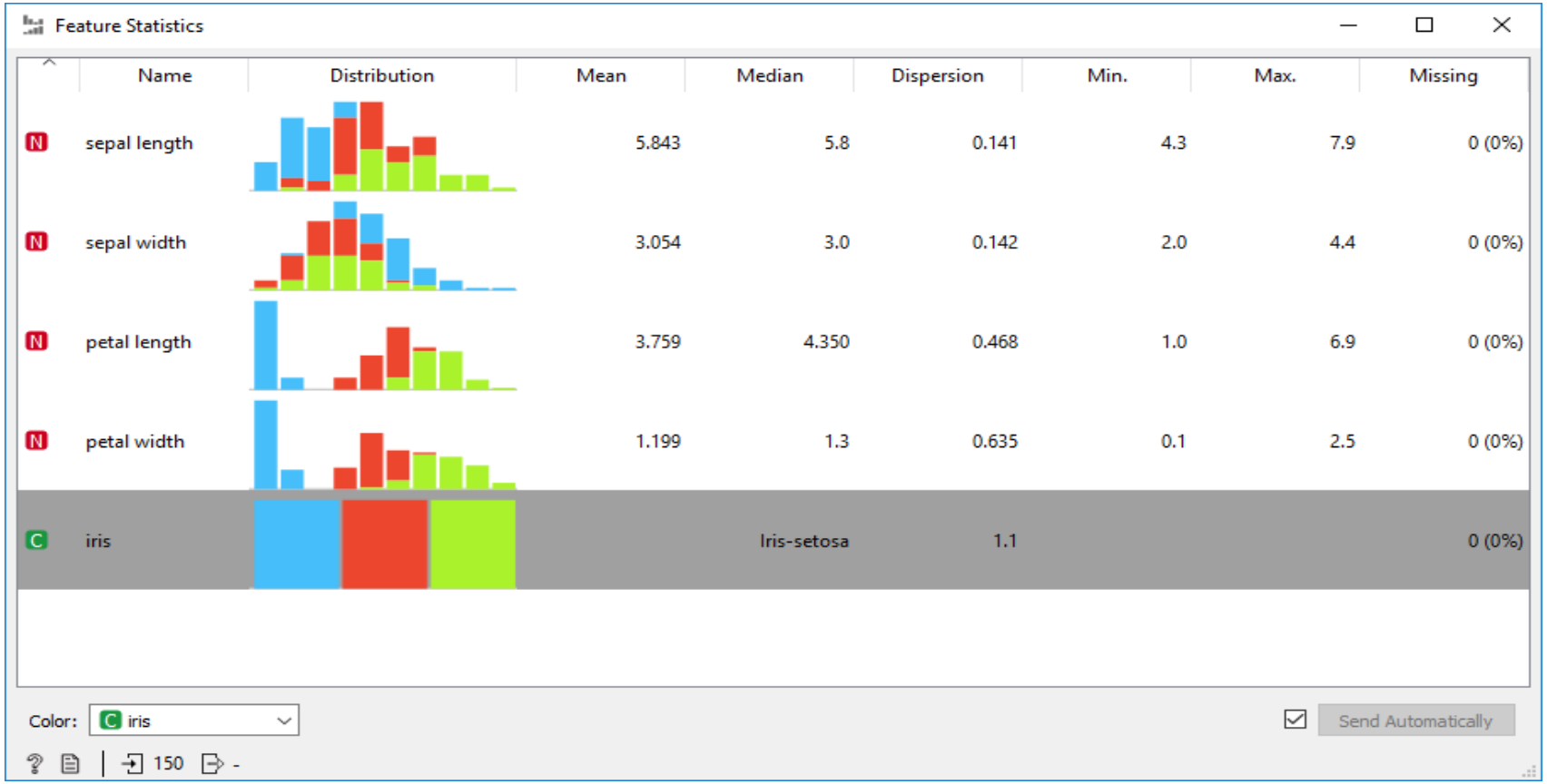
❑What to notice?

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

Orange EDA:

What are the stats of the variables?



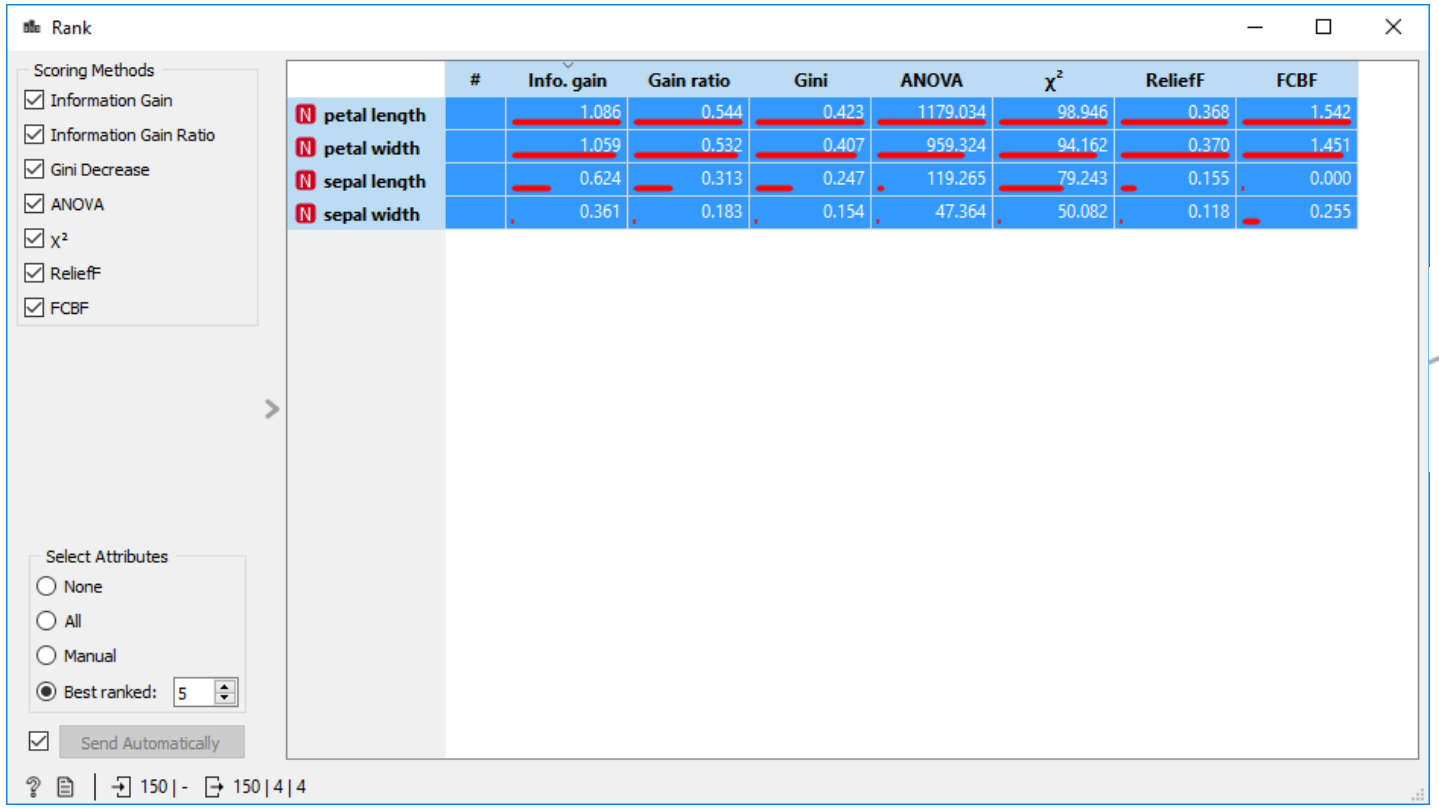


❑What to notice?

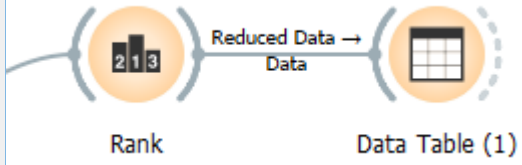
- Centre, spread, no missing values of variables
- Distributions of variables over classes
- Classes balance

Orange EDA:

What are the most important features?



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



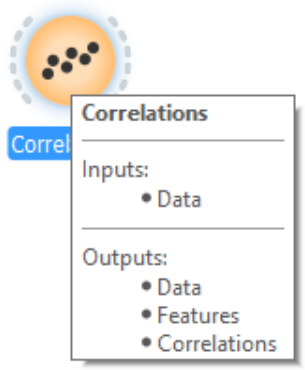
❑What to notice?

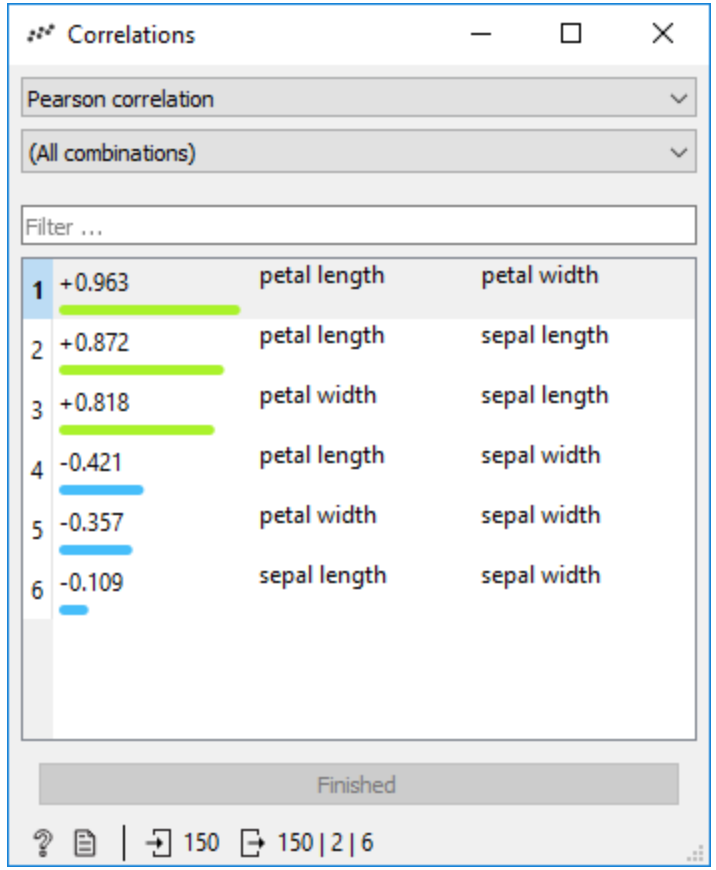
- Petal length seems the most important and sepal width is the least

➔ Feature selection

Orange EDA:

What are the most important features?





❑What to notice?

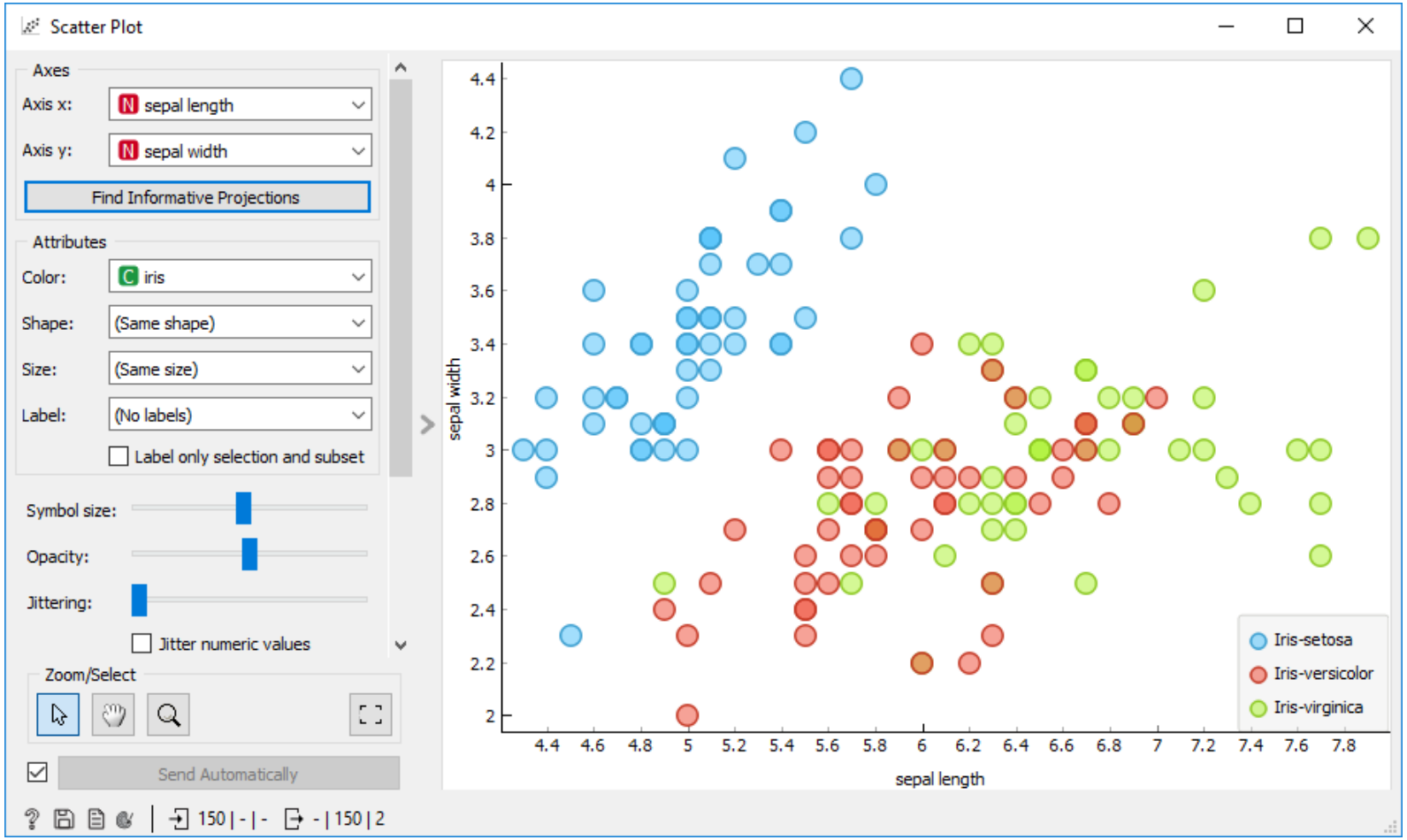
- [-1, 1], negative/positive, strong/weak…
- Petal length and petal width look strongly corelated

➔ Feature selection

Orange EDA:

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

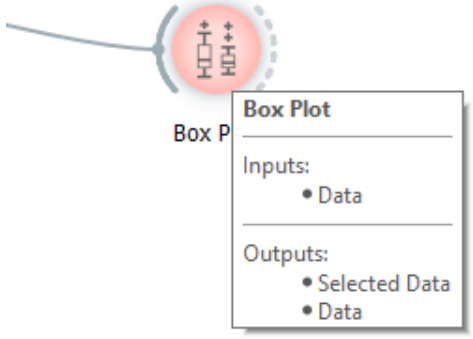
Scatter Plot



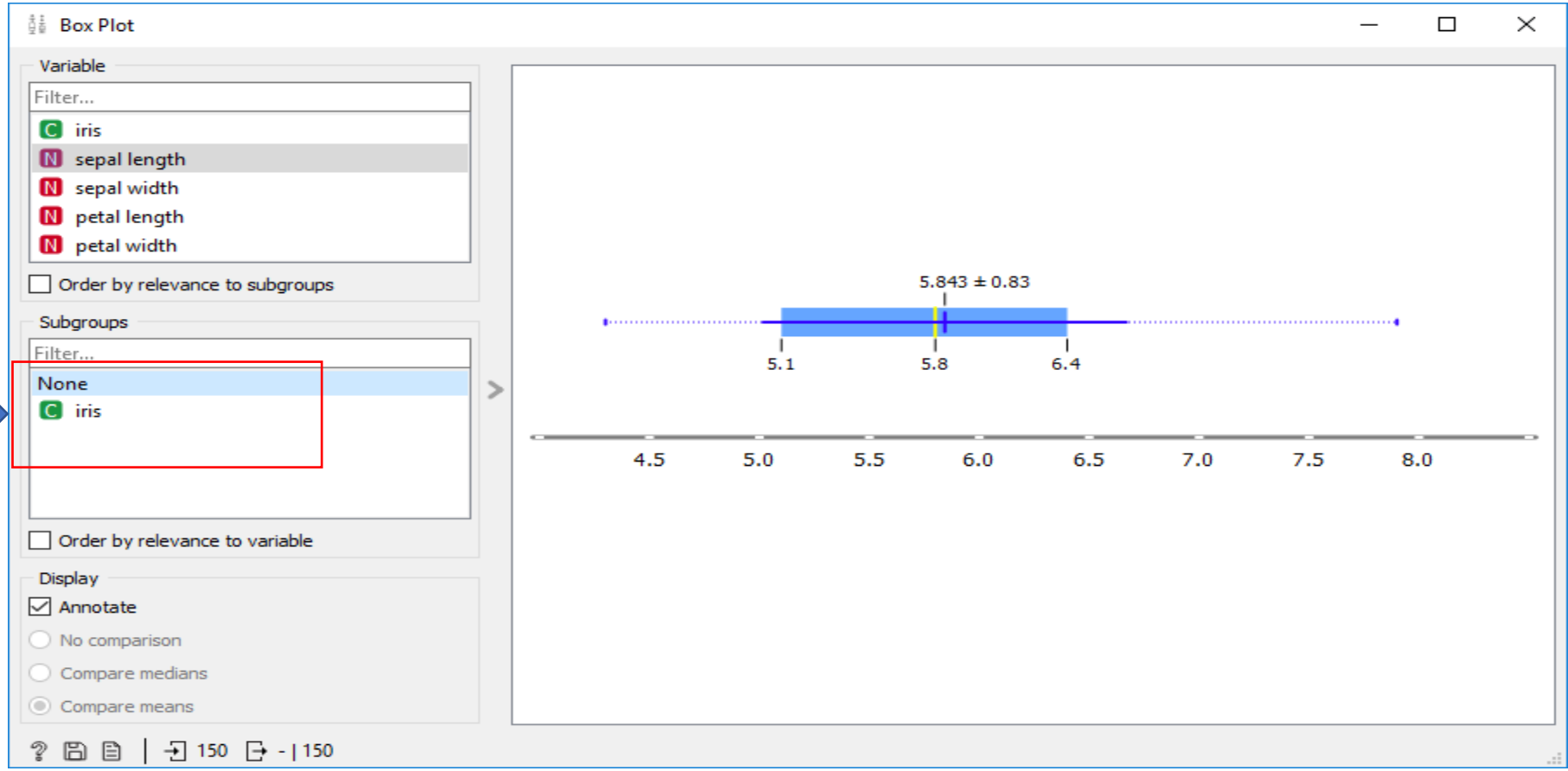
- Change variables
- What to notice?
- Compare with the correlation shown previously

Orange EDA:

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



Univariate

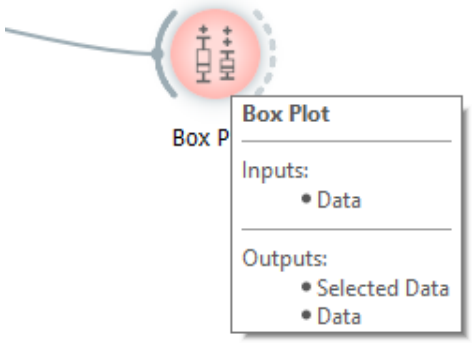


❑What to notice?

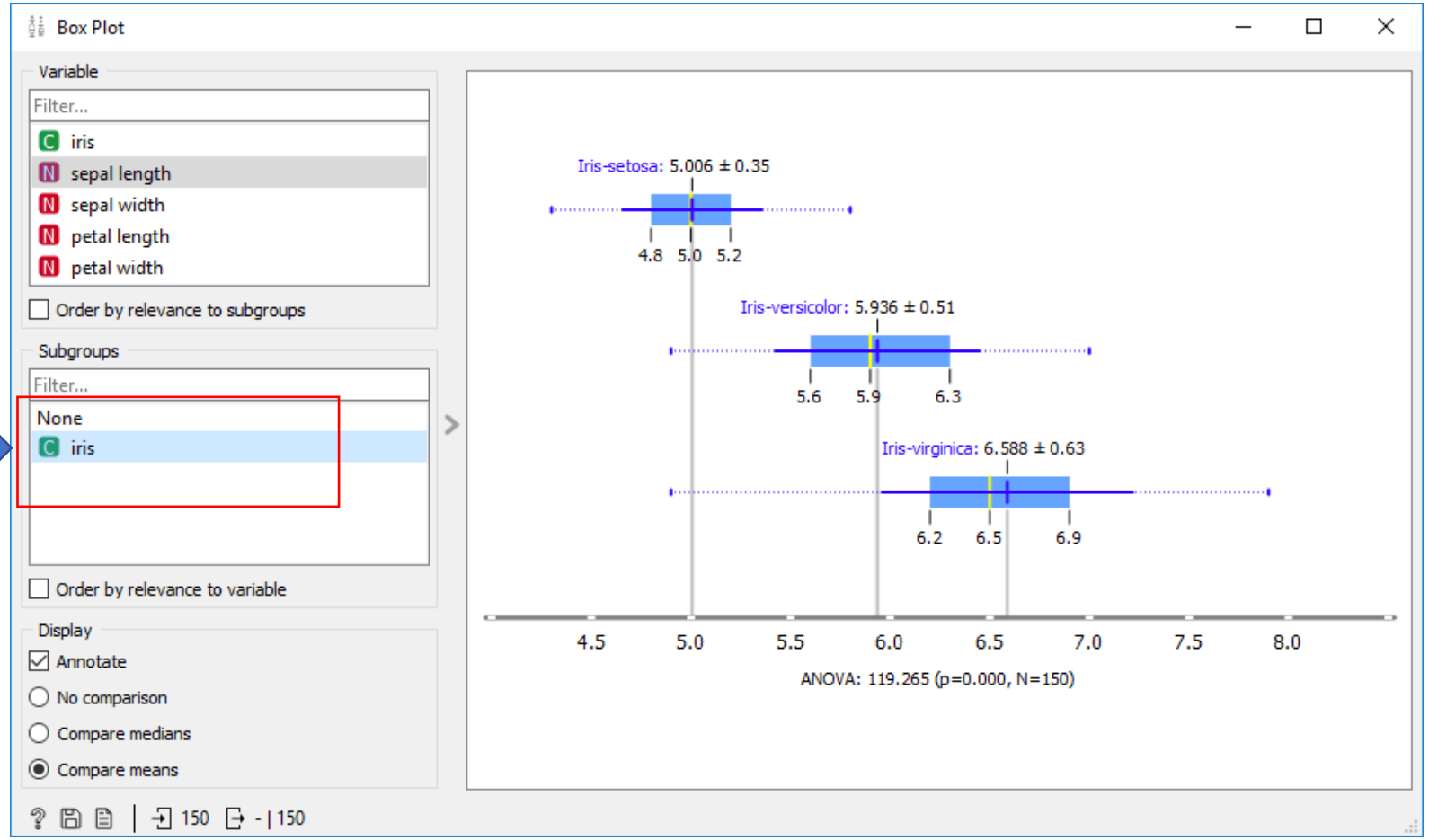
- Graphical presentation for the stats

Orange EDA:

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



Multivariate

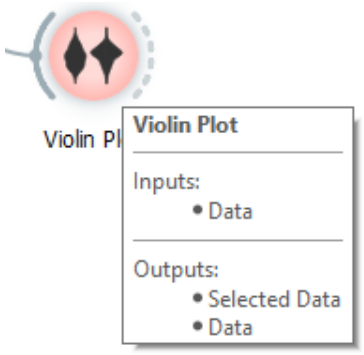


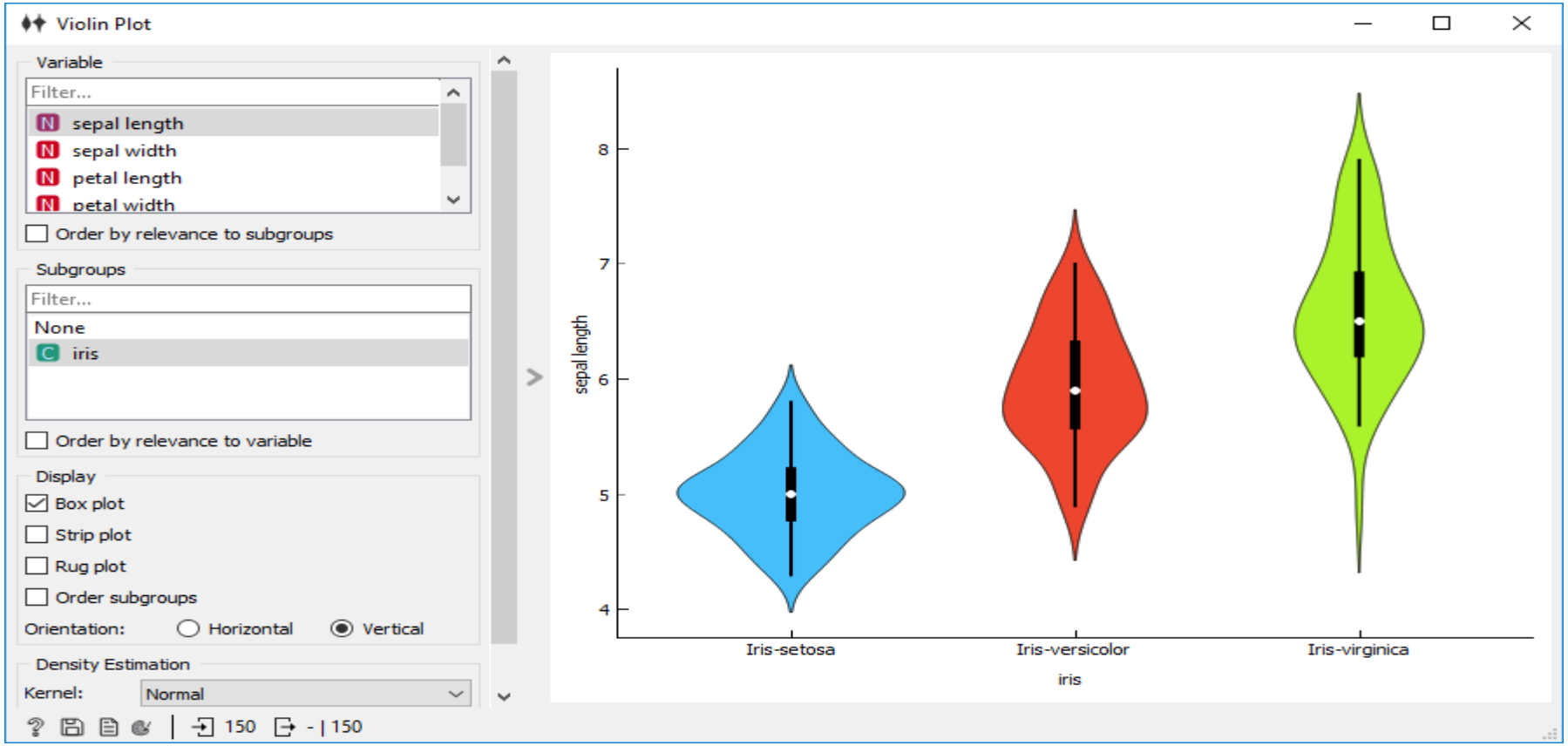
❑What to notice?

- 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)?



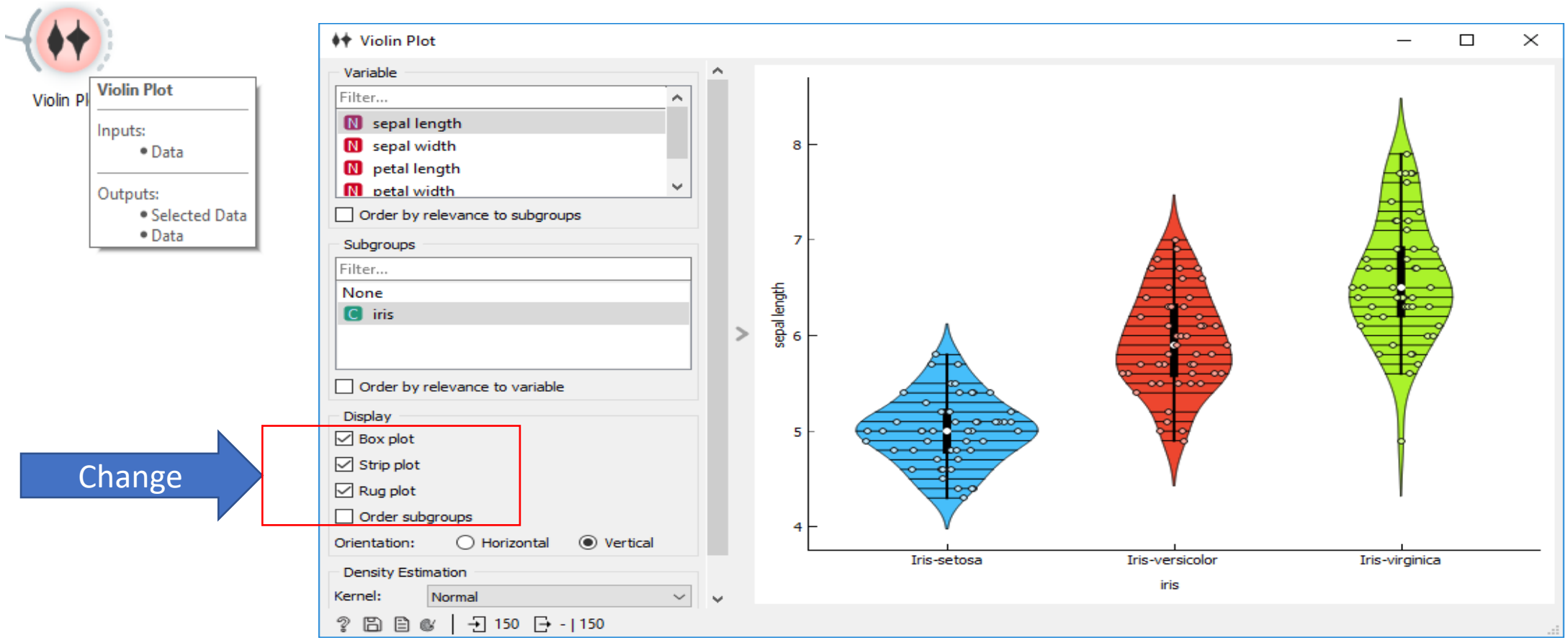


❑What to notice?

- 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)?

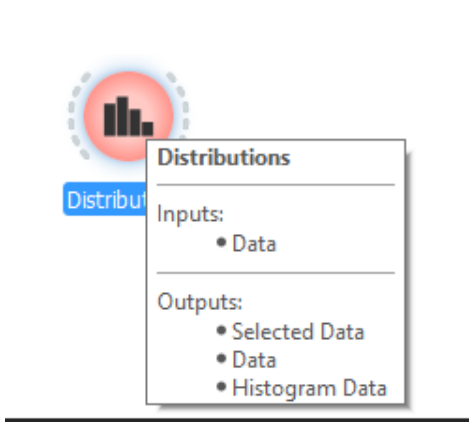


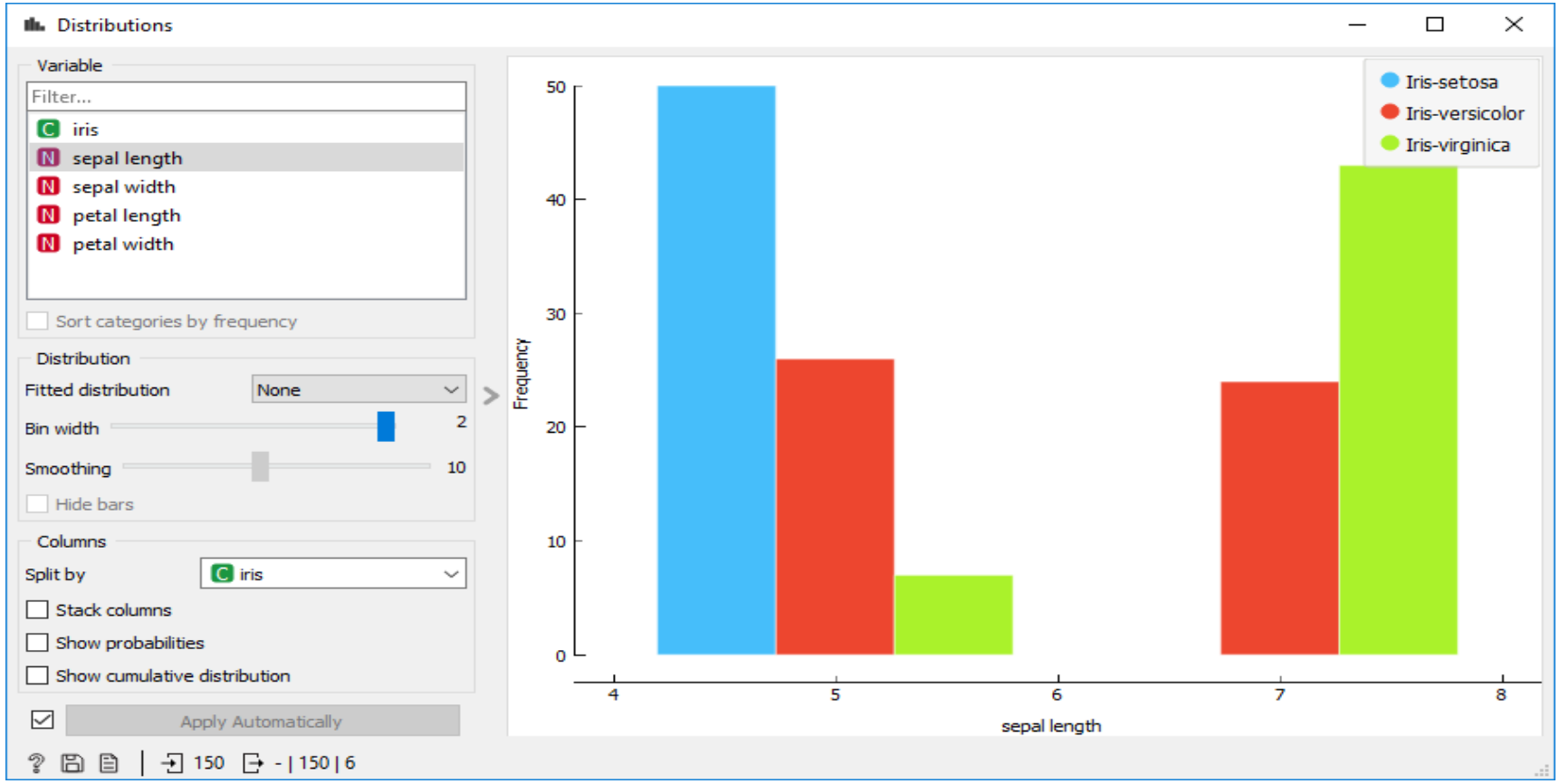
❑What to notice?

- 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?