

Internet Studies Lab, Department of Applied
Mathematics and Business Informatics



НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ
УНИВЕРСИТЕТ

INTRODUCTION TO ORANGE SOFTWARE

Анализ баз данных в публичном управлении
Кольцов С.Н.

Saint Petersburg, 07.09.2018

ORANGE DATA MINING



[Screenshots](#) [Download](#) [Docs](#) [Blog](#)

Data Mining Fruitful and Fun

Open source data visualization and data analysis for novice and expert. Interactive workflows with a large toolbox.

[Download Orange](#)

The **old version**, Orange 2.7, is still available.



Можно
скачать по
адресу:
<http://orange.biolab.si/>



Interactive workflows

Create your own interactive workflows to analyse and visualize your data.



Visualization

Orange is packed with different visualizations, from scatter plots, bar charts, trees, to dendrograms, networks and heat maps.



Large Toolbox

Over 100 widgets and growing. We cover most of standard data analysis tasks. Specialized additions available, like Orange-Bioinformatics.

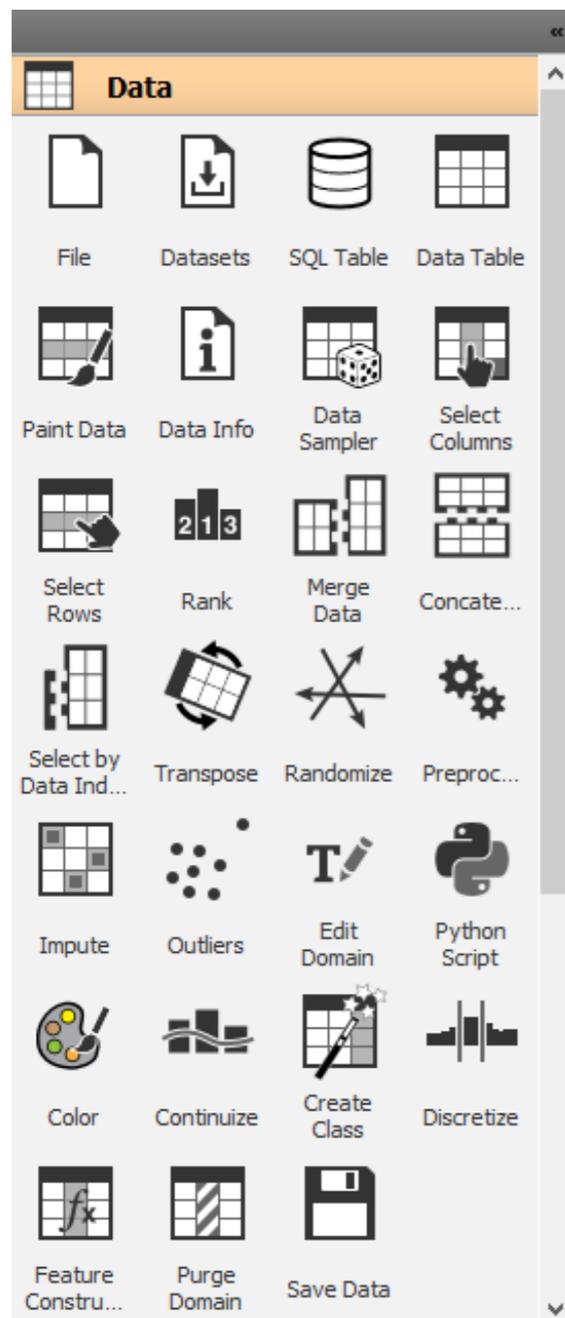
ORANGE CONSISTS OF A CANVAS INTERFACE ONTO WHICH THE USER PLACES WIDGETS AND CREATES A DATA ANALYSIS WORKFLOW

The screenshot displays the Orange data mining software interface. The title bar reads "Hierarchical Clustering". The menu bar includes "File", "Edit", "View", "Widget", "Options", and "Help". On the left, a "Data" widget palette is visible, containing various data processing and analysis widgets. The main canvas shows a workflow diagram with the following components and annotations:

- File**: A widget icon representing a file. An annotation with a red arrow points to it: "Read the data. Try this schema with the brown-selected set (from data sets that come with Orange)." A red arrow also points from this widget to the "Distances" widget.
- Distances**: A central widget icon representing distance calculation. An annotation with a red arrow points to it: "Compute the distances between the data samples." A red arrow also points from this widget to the "Hierarchical Clustering" widget.
- Hierarchical Clustering**: A widget icon representing clustering. An annotation with a red arrow points to it: "Hierarchically cluster the data." A red arrow also points from this widget to the "Distance Map" widget.
- Distance Map**: A widget icon representing a distance heatmap. An annotation with a red arrow points to it: "Visualize the data distances in a heat map." A red arrow also points from this widget to the "Data Table" widget.
- Data Table**: A widget icon representing a data table. An annotation with a red arrow points to it: "Visualize the data distances in a heat map." A red arrow also points from this widget to the "Data Table (1)" widget.
- Data Table (1)**: A widget icon representing a data table. An annotation with a red arrow points to it: "Choose any part of the clustering dendrogram. Then, observe the selected data in a data table, or in any other analysis widget. Open both Hierarchical Clustering and Data Table (1) widget to turn this schema into interactive data analysis."

The workflow is connected by lines labeled "Data" and "Distances". The "Data" widget palette on the left includes categories like "Data", "Visualize", "Classify", and "Regression".

ORANGE CONSISTS OF A CANVAS INTERFACE ONTO WHICH THE USER PLACES WIDGETS AND CREATES A DATA ANALYSIS WORKFLOW



Data: widgets for data input, data filtering, sampling, imputation, feature manipulation and feature selection



Widgets ответственный за открытие файлов с данными.

File



Widgets ответственный за визуализацию данных

Data Table



Widgets позволяет выбирать колонки для анализа

Select Columns



Widgets позволяет выбирать строки для анализа

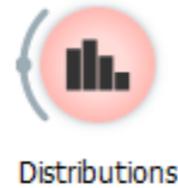
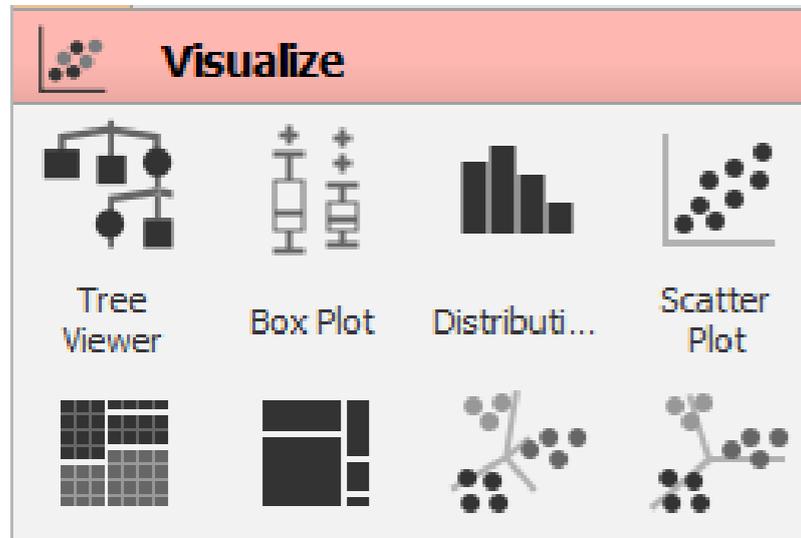
Select Rows



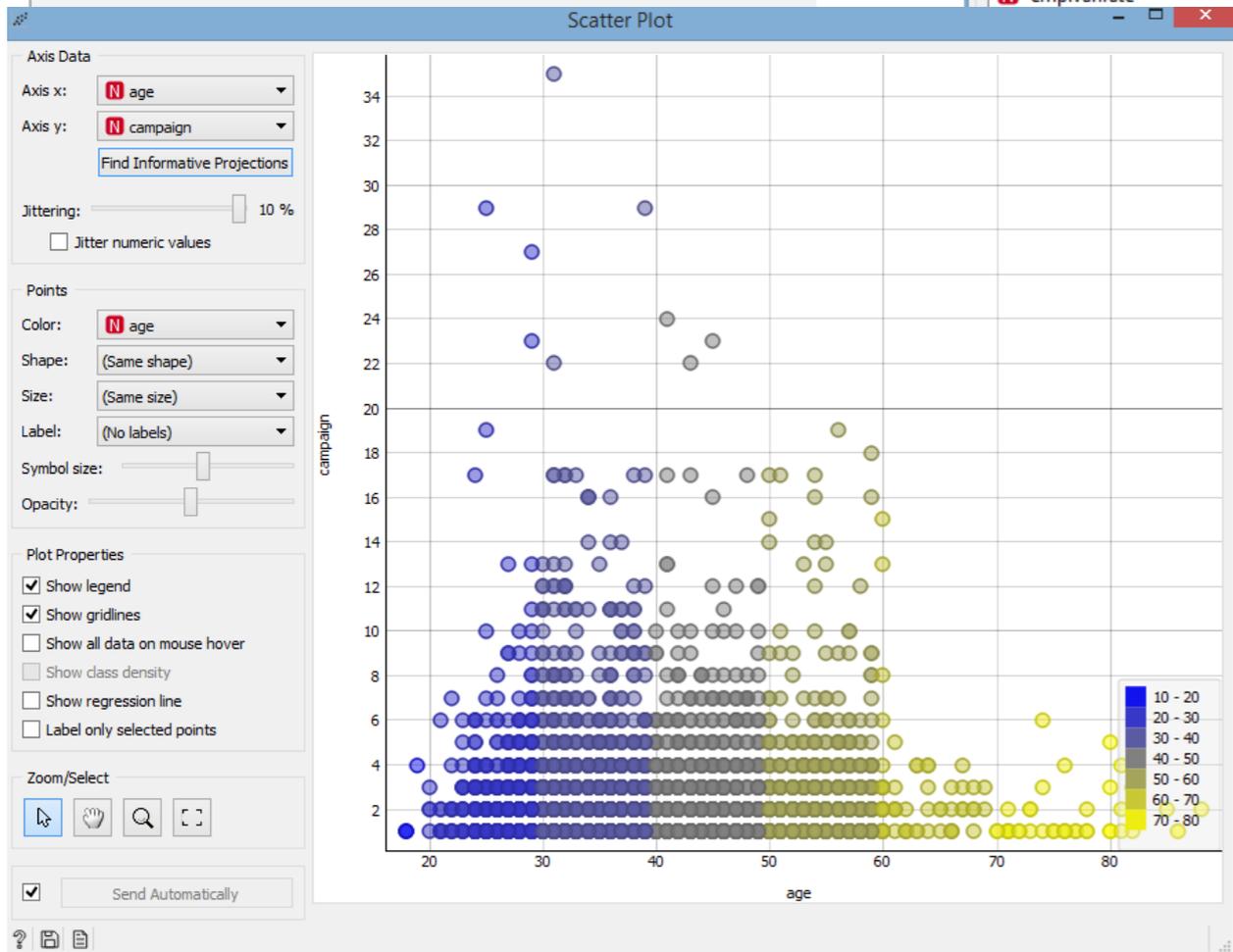
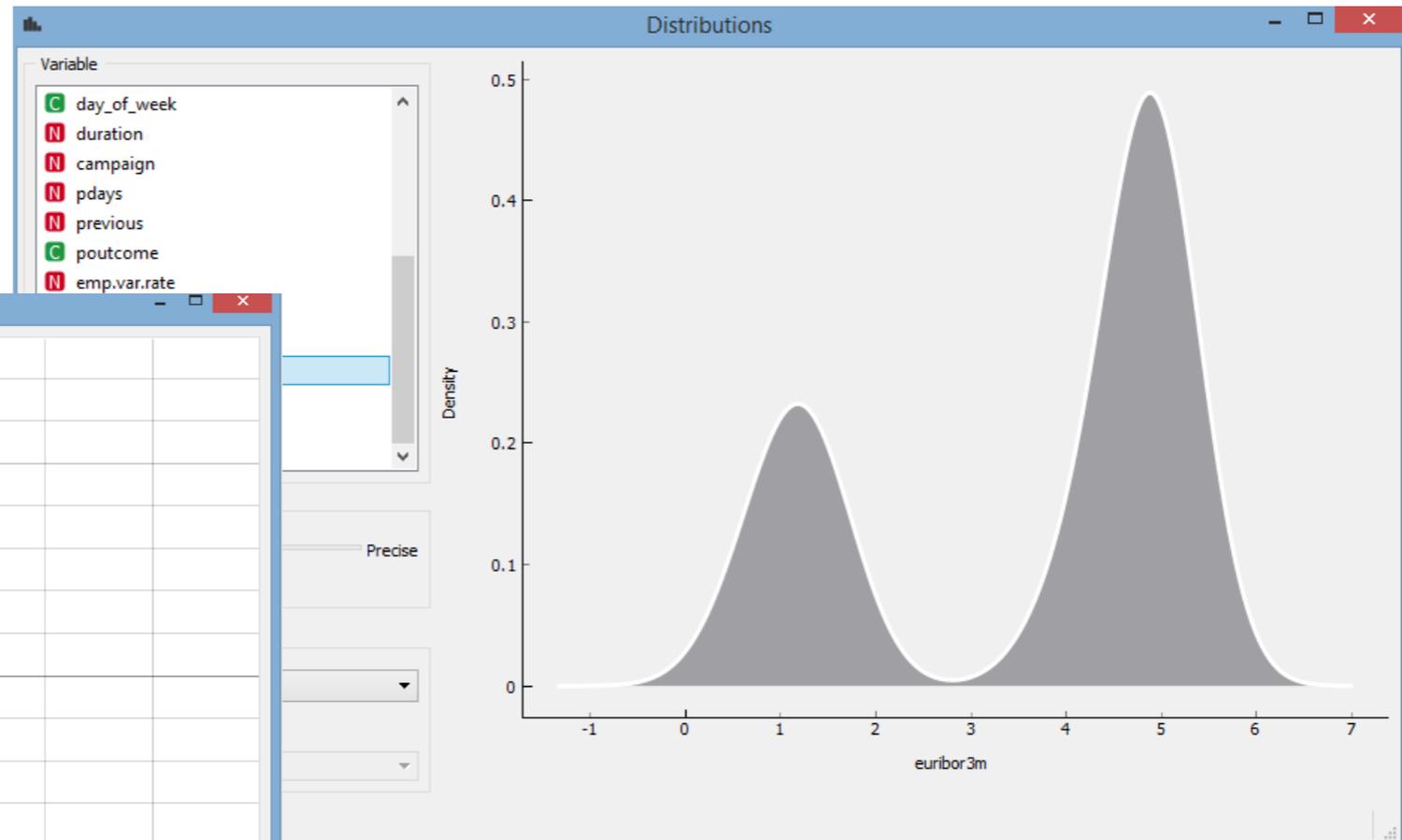
Widgets позволяет выбирать датасеты из предустановленного набора

Datasets

ORANGE : VISUALIZE WIDGETS



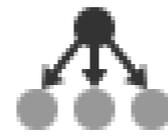
Widgets ответственный за рисование распределений по данным



Widgets ответственный за рисование распределений по данным

ORANGE : MODELS

Model			
Constant	CN2 Rule Induction	kNN	Tree
Random Forest	SVM	Linear Regression...	Logistic Regression...
Naive Bayes	AdaBoost	Neural Network	Stochastic Gradient...
Save Model	Load Model		



Naive Bayes

Widgets отвечающий за **наивный байесовский классификатор.**



SVM

Widgets осуществляющий классификацию на основе **Метода опорных векторов.**



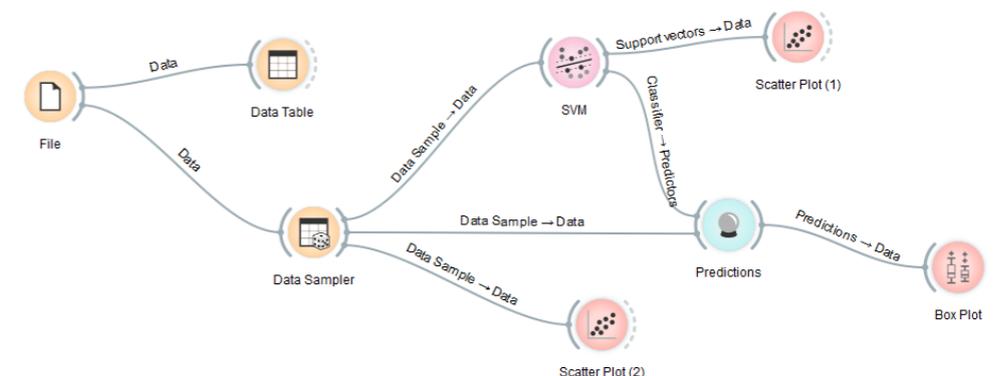
Logistic Regression...

Widgets осуществляющий классификацию на основе **логистической регрессии.**

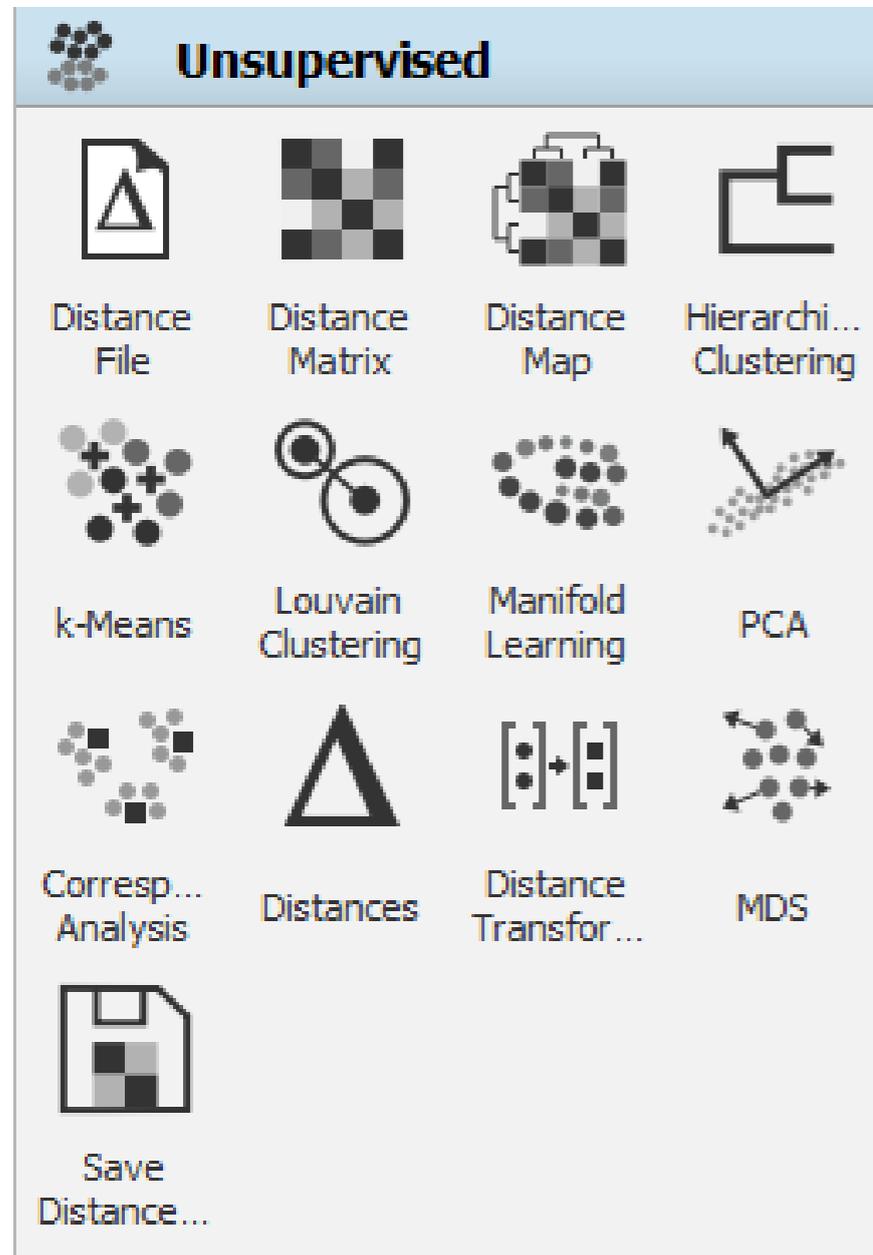


Nearest Neighbors

Widgets производящий классификацию по **методу ближайших соседей.**

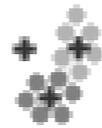


ORANGE : UNSUPERVISED



Hierarchi...
Clustering

Widgets отвечающий за
иерархическую кластеризацию.



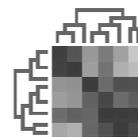
k-Means

Widgets позволяющий проводить
кластеризацию методом **К-Means.**



PCA

Widgets позволяющий проводить
principal component analysis



Distance
Map

Widgets визуализирующий
матрицу расстояний.



Louvain Clustering

Widgets позволяющий провести кластеризацию сети при помощи луванского алгоритма.

ORANGE : КАК ЗАГРУЗИТЬ И ПРОСМОТРЕТЬ ДАННЫЕ

1. Запускаем Orange и создаем пустой проект.
2. Открываем опцию 'Data', и кликаем на widget 'File'
3. Выбираем widget 'Data Table'.

Соединяем 'File'
'Data Table' линией.



File



Data Table



File

Data Table

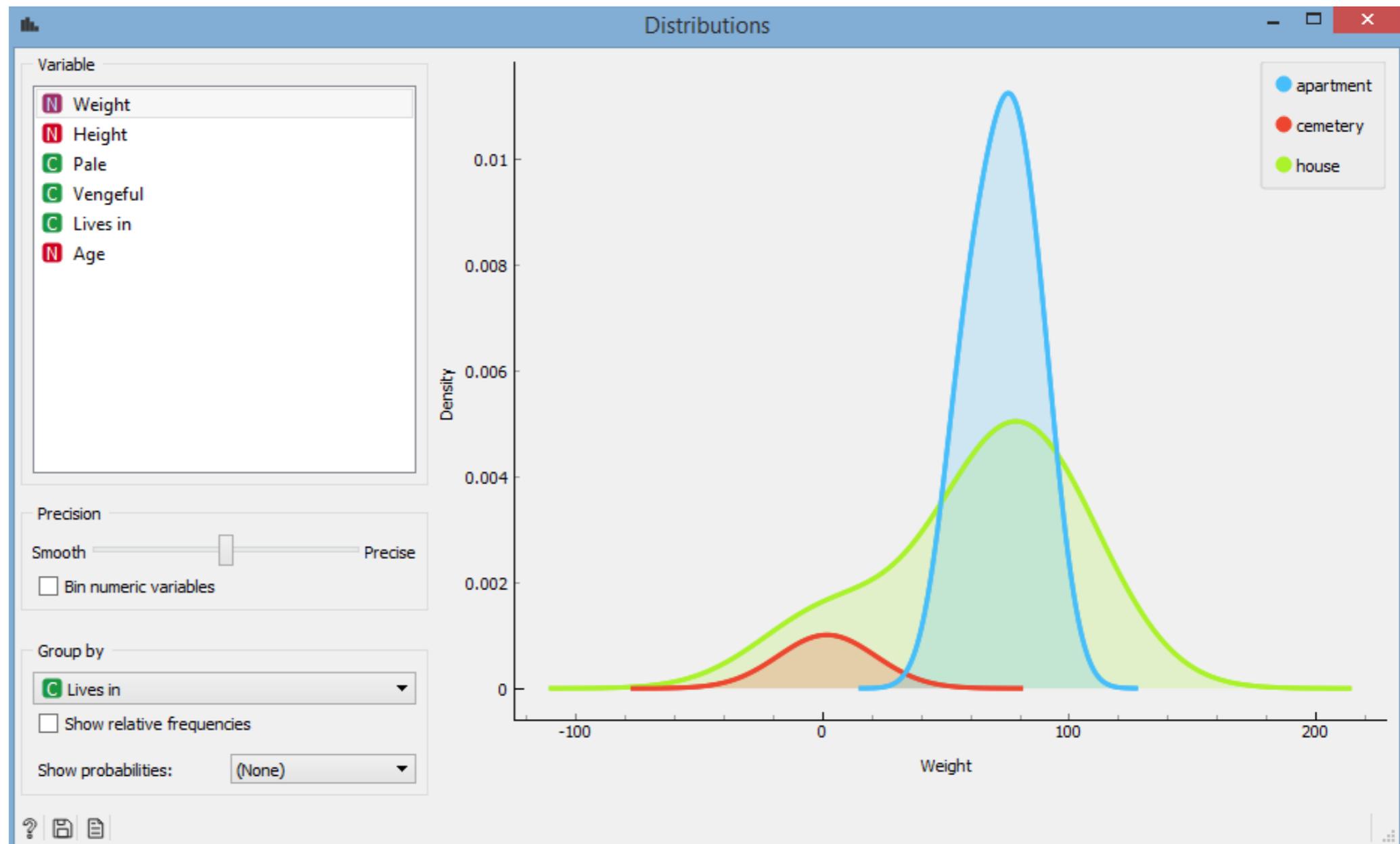
widgets 'File' – отвечает за то какой файл нужно загрузить. Кликаем на нем и указываем имя файла, например, **Paranormal distribution.csv**. Если все нормально, то что бы посмотреть что находится в этом файле нужно кликнуть на widgets 'Data Table'. В результате получим следующее:

	Weight	Height	Pale	Vengeful	Lives in	Age	Name
1	55.000	167.000	yes	no	house	45.000	Anne
2	89.000	180.000	no	no	apartment	19.000	Jack
3	78.000	175.000	yes	yes	apartment	43.000	Mark
4	2.000	166.000	yes	yes	house	50.000	Ghoul
5	70.000	150.000	yes	yes	house	32.000	Marie
6	102.000	190.000	no	no	house	44.000	Clark
7	60.000	155.000	yes	no	apartment	61.000	Rebecca
8	2.000	190.000	yes	yes	cemetery	120.000	Spirit
9	79.000	161.000	no	yes	apartment	21.000	Helen
10	81.000	181.000	no	yes	house	37.000	Nick
11	63.000	171.000	no	no	apartment	35.000	Sarah
12	92.000	178.000	yes	no	house	41.000	Oliver
13	78.000	168.000	yes	no	house	28.000	Valerie
14	2.000	171.000	yes	yes	house	92.000	Phantom
15	72.000	185.000	yes	yes	apartment	58.000	Bill
16	54.000	153.000	yes	no	apartment	18.000	Claire
17	90.000	178.000	no	no	house	47.000	Ian

ORANGE : КАК ПОСТРОИТЬ РАСПРЕДЕЛЕНИЯ ДАННЫХ НА ГРАФИКЕ

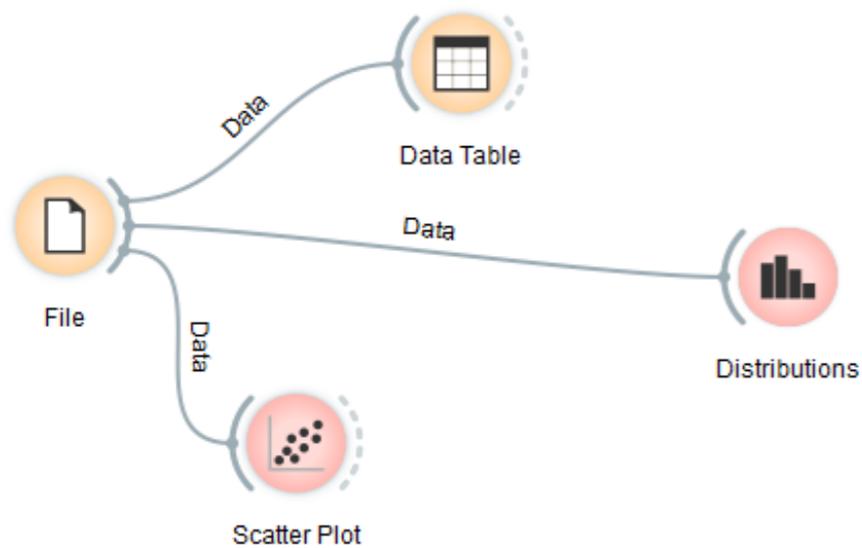
К существующему проекту добавляем widgets ‘Distribution’. Он находится в опции ‘Visualize’. Соединяем widget ‘File’ с widget ‘Distribution’.

Можно объединить несколько графиков на графике. На данном примере показаны распределения веса в зависимости от места проживания.

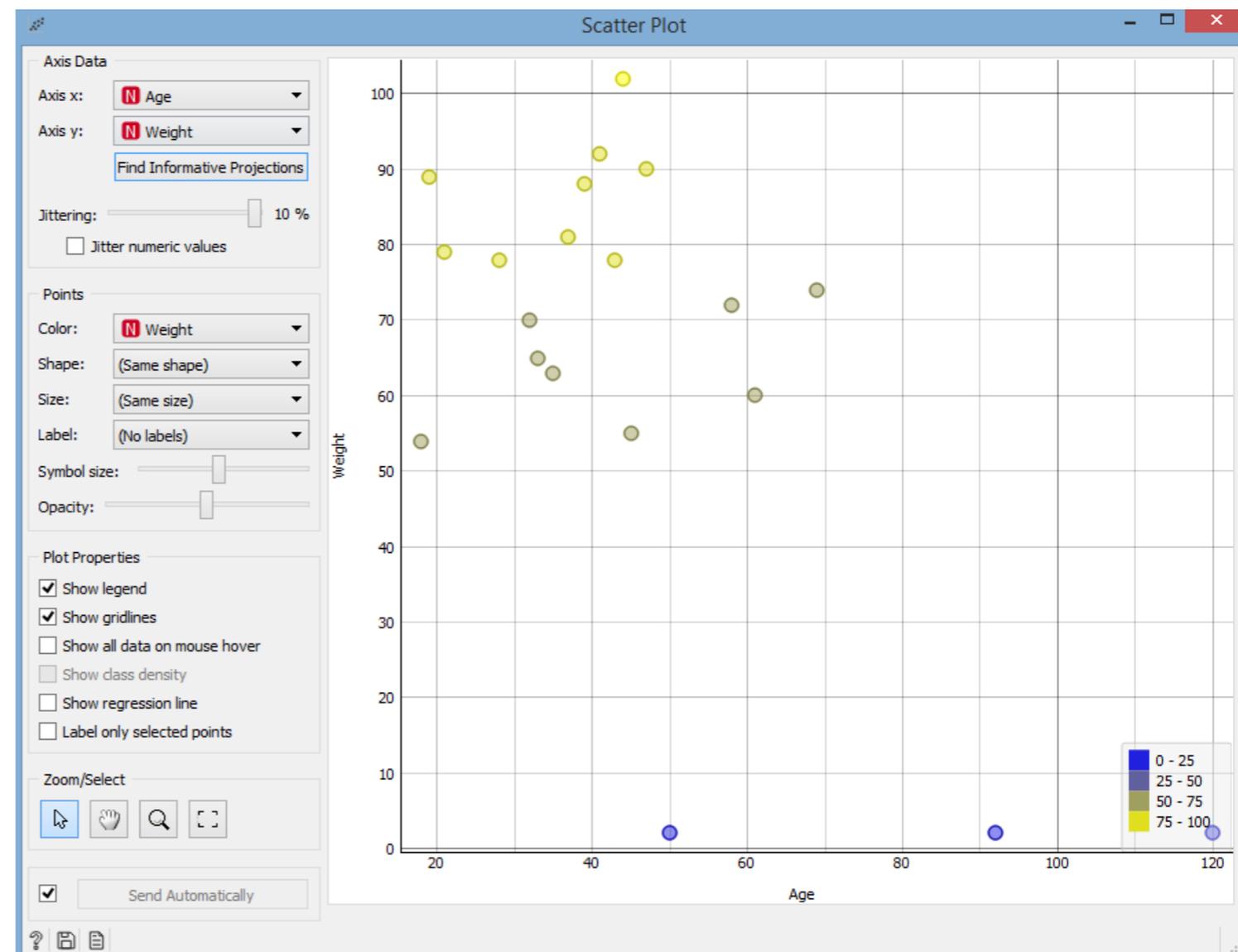


ORANGE : КАК ПОСТРОИТЬ SCATTERING PLOT

К существующему проекту добавляем widgets ‘Scatter plot’. Он находится в опции ‘Visualize’. Соединяем widget ‘Scatter plot’ с widget ‘File’.



Для того что бы построить график нужно кликнуть на widget ‘Scatter plot’. В появившемся окне указать какие данные будут использоваться для осей X, Y и по какой переменной раскрашивать данные.





НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ
УНИВЕРСИТЕТ

<https://linis.hse.ru/>

Phone: +7 (911) 981 9165

Email: skoltsov@hse.ru