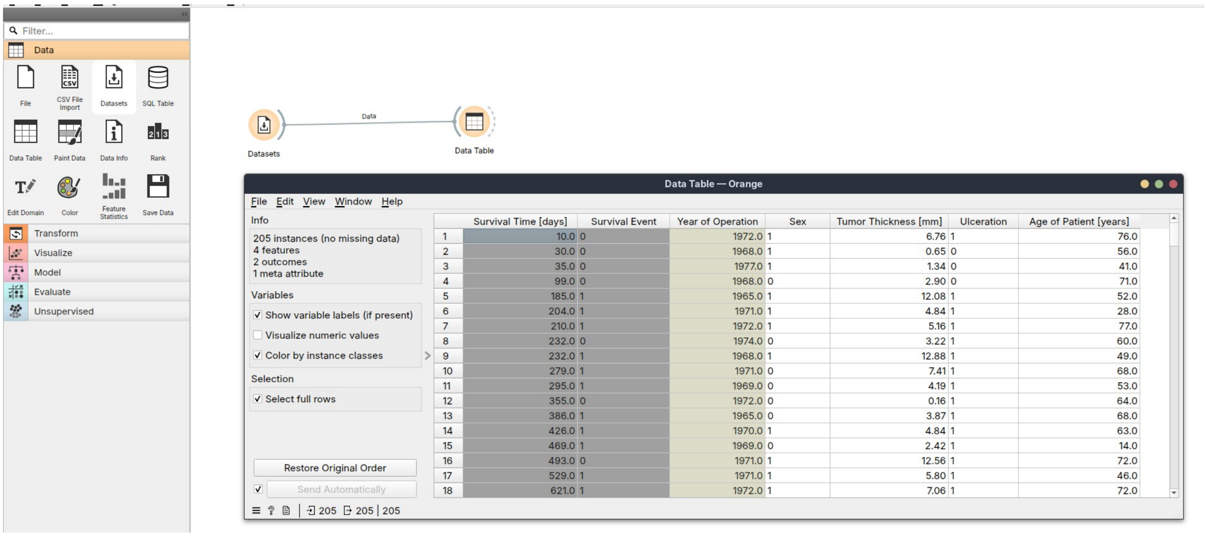
No code lab

Data visualization

1. From the [Data] section select the **Datasets** operator. Double-click it and select the *Melanoma: Survival from Melignant Melanoma* dataset. Double-click on the name of the dataset so that a green dot appears on the left side of the dataset (indicating that the dataset has been succesfully downloaded).
2. Put the **Data Table** operator in the panel. Drag the data connection between **Datasets** and **Data Table** operators. Then, double-click on the **Data Table** operator to see the data.



Info

205 instances (no missing data)
4 features
2 outcomes
1 meta attribute

Variables

☑ Show variable labels (if present)

□ Visualize numeric values

☑ Color by instance classes

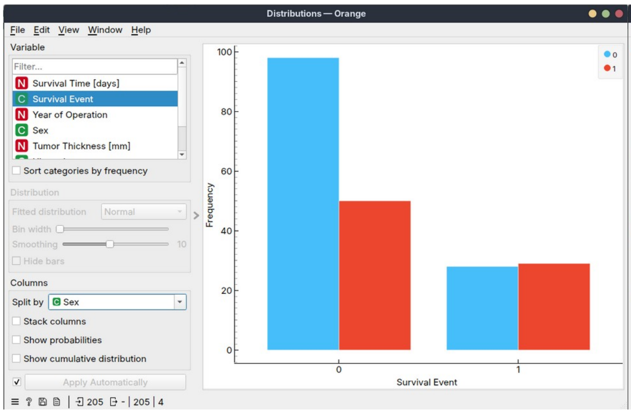
Selection

☑ Select full rows

	Survival Time [days]	Survival Event	Year of Operation	Sex	Tumor Thickness [mm]	Ulceration	Age of Patient [years]
1	10.0	0	1972.0	1	6.76	1	76.0
2	30.0	0	1968.0	1	0.65	0	56.0
3	35.0	0	1977.0	1	1.34	0	41.0
4	99.0	0	1968.0	0	2.90	0	71.0
5	185.0	1	1965.0	1	12.08	1	52.0
6	204.0	1	1971.0	1	4.84	1	28.0
7	210.0	1	1972.0	1	5.16	1	77.0
8	232.0	0	1974.0	0	3.22	1	60.0
9	232.0	1	1968.0	1	12.88	1	49.0
10	279.0	1	1971.0	0	7.41	1	68.0
11	295.0	1	1969.0	0	4.19	1	53.0
12	355.0	0	1972.0	0	0.16	1	64.0
13	386.0	1	1965.0	0	3.87	1	68.0
14	426.0	1	1970.0	1	4.84	1	63.0
15	469.0	1	1969.0	0	2.42	1	14.0
16	493.0	0	1971.0	1	12.56	1	72.0
17	529.0	1	1971.0	1	5.80	1	46.0
18	621.0	1	1972.0	1	7.06	1	72.0

As you can see, the dataset presents the on patients being treated from melanoma, with attributes representing the year of operation, the sex of a patient (0=female, 1=male), the size of the tumor and its potential ulceration, the age of a patient, and two special attributes representing the survival time and the survival event (outcome of the operation, 0=survived, 1=died).

3. From the [Visualize] section drag the **Distributions** operator and send the data to the operator. To do this, simply drag a line from the output of the **Datasets** operator (the grey arc) to the **Distributions** operator. Open the operator and compare the display of continuous (numerical) attributes with the display of categorical (discrete) attributes.
   1. Display the *Survival Event*, are patients are more likely to survive the operation?
   2. Split the *Survival Event* by *Sex* (look for “Split by” option in the lower left part of the window). Are women more likely to survive the operation then men?
   3. Display the *Age of Patient* and see what happens when you slide the “Bin width” slider. Are there more patients in the age range between 40-50 or 50-60?
   4. Fit the normal distribution to the display (look for “Fitted distribution” dropdown list). What is the mean age of patients?



Distributions — Orange

File Edit View Window Help

Variable

- N Survival Time [days]
- C Survival Event
- N Year of Operation
- C Sex
- N Tumor Thickness [mm]

□ Sort categories by frequency

Distribution

Fitted distribution: Normal

Bin width

Smoothing 10

□ Hide bars

Columns

Split by: C Sex

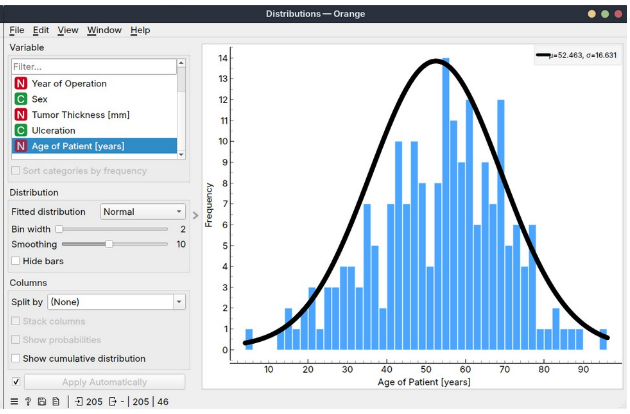
□ Stack columns

□ Show probabilities

□ Show cumulative distribution

☑ Apply Automatically

205 | 205 | 4



Distributions — Orange

File Edit View Window Help

Variable

- N Year of Operation
- C Sex
- N Tumor Thickness [mm]
- C Ulceration
- N Age of Patient [years]

□ Sort categories by frequency

Distribution

Fitted distribution: Normal

Bin width 2

Smoothing 10

□ Hide bars

Columns

Split by: (None)

□ Stack columns

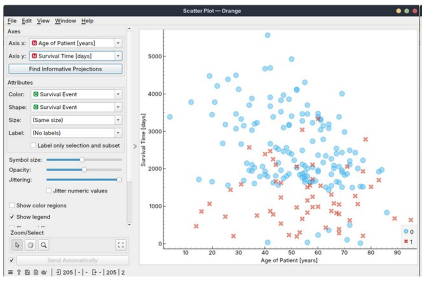
□ Show probabilities

□ Show cumulative distribution

☑ Apply Automatically

205 | 205 | 46

4. From the [Visualize] section drag the **Scatter Plot** operator. Drag the mouse from the **Datasets** to the **Scatter Plot** operator. Try to answer the following questions:
    1. How does age of a patient and survival time relate to the probability of surviving the melanoma? *Hint*: plot *Age of Patient* on X-axis, plot *Survival Time* on Y-axis, and use *Survival Event* to set the color and shape of points.
    2. Is there a difference between men and women when it comes to surviving the illness with respect to the age of the patient? *Hint*: plot *Age of Patient* on X-axis, plot *Sex* on Y-axis, and use *Survival Event* to set the color and shape of points. Since *Sex* is a discrete attribute, in order to make the figure more readable you will have to use the jittering option. You may also want to see what happens when you check the “Show color regions” checkbox.
    3. How do ulceration and tumor thickness influence the survival of the disease? *Hint*: plot *Tumor Thickness* on X-axis, plot *Ulceration* on Y-axis, and use *Survival Event* to set the color and shape of points.
5. Click on the button “Find Informative Projections” in the upper left corner of the window. The tool will propose the combinations of features that will produce the most informative displays of the data.



Scatter Plot — Orange

File Edit View Window Help

Axes

Axis x: N Age of Patient [years]

Axis y: N Survival Time [days]

Find Informative Projections

Attributes

Color: C Survival Event

Shape: C Survival Event

Size: (Same size)

Label: (No labels)

□ Label only selection and subset

Symbol size:

Opacity:

Jittering:

□ Jitter numeric values

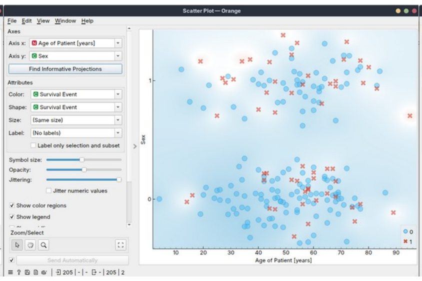
□ Show color regions

☑ Show legend

Zoom/Select

☑ Send Automatically

205 | - | - | 205 | 2



Scatter Plot — Orange

File Edit View Window Help

Axes

Axis x: N Age of Patient [years]

Axis y: C Sex

Find Informative Projections

Attributes

Color: C Survival Event

Shape: C Survival Event

Size: (Same size)

Label: (No labels)

□ Label only selection and subset

Symbol size:

Opacity:

Jittering:

□ Jitter numeric values

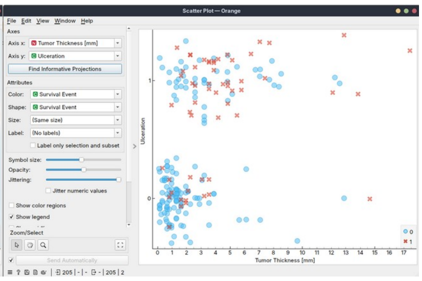
☑ Show color regions

☑ Show legend

Zoom/Select

☑ Send Automatically

205 | - | - | 205 | 2



Scatter Plot — Orange

File Edit View Window Help

Axes

Axis x: N Tumor Thickness [mm]

Axis y: C Ulceration

Find Informative Projections

Attributes

Color: C Survival Event

Shape: C Survival Event

Size: (Same size)

Label: (No labels)

□ Label only selection and subset

Symbol size:

Opacity:

Jittering:

□ Jitter numeric values

□ Show color regions

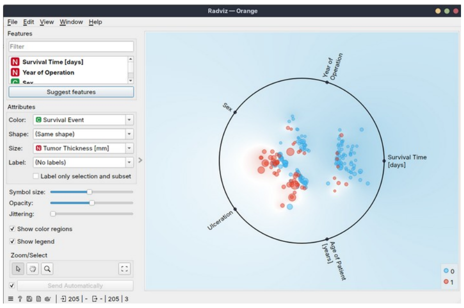
☑ Show legend

Zoom/Select

☑ Send Automatically

205 | - | - | 205 | 2

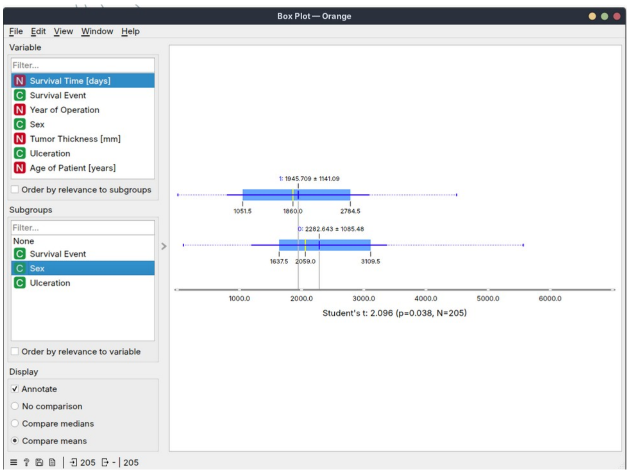
6. Send the data to the **RadVis** operator. Use color to distinguish between patients who survived and who did not. Use size of a data point to represent the tumor thickness. Display color regions to better understand, how these 6 features together influence the survival probability. Think of each patient as a point connected to features with rubber bands. The larger the value of a given attribute, the stronger the rubber band pulls in the direction of the attribute. Experiment with adding or removing features (simply click on a feature in the upper left part of the window). Do you find this type of visualization useful?



Survival Time [days]
Year of Operation
Sex

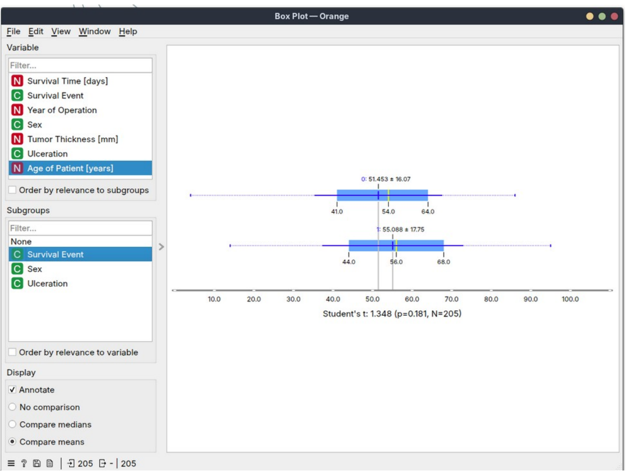
Color: Survival Event
Shape: (Same shape)
Size: Tumor Thickness [mm]
Label: (No labels)
□ Label only selection and subset
Symbol size:
Opacity:
Jittering:
☑ Show color regions
☑ Show legend

7. Drag the **Box Plot** operator to the panel and send the data to the operator. Box Plot is useful for visualizing simple data statistics. The blue line in the middle represents the mean, the values at the ends of the blue shaded area are the values of the 1st and 3rd quartile (25% and 50% of data), the yellow blue line represents the median, and the horizontal blue line is the standard deviation around the mean.
Try to answer the following questions:
    1. Is there a significant difference in survival time between men and women? *Hint*: use *Survival Time [days]* as the main variable in the upper left part of the window. Mark *Sex* as the Subgroups variable in the lower left part of the window. Select the "Compare means" radiobox in the lower left part of the window and make sure that "Annotate" checkbox is checked. You can see that the average survival for men is 1945 days, and 2282 days for women. In addition, the difference is statistically significant, because the p-value of the Student's t-test is below 0.05.
    2. Is there a significant difference in patients age between patients who survived and those who did not? Hint: use *Age of Patient [years]* as the main variable, and *Survival Event* as the subgroup variable. Is there enough data to say that the difference of two years (51.6 vs 53.9) is statistically significant?



□ Order by relevance to subgroups
□ Order by relevance to variable
☑ Annotate
○ No comparison
○ Compare medians
● Compare means

Student's t: 2.096 (p=0.038, N=205)



□ Order by relevance to subgroups
□ Order by relevance to variable
☑ Annotate
○ No comparison
○ Compare medians
● Compare means

Student's t: 1.348 (p=0.181, N=205)

Assignment

Use the **Datasets** operator to download the *German Credit Data*. Using visualization tools try to answer the following questions:

- Does age influence the credit score rating of customers?
- Is there a relationship between sex/marital status and credit score rating?
- What can be said about credit score rating from the credit history and status of existing checking account?