Lab-2 [Manage Data]

**Out date:** Jun 20, 2022

**Due date:** Jun 23, 2022 at 11:59PM

**Submission**

1. Prepare your solutions in Orange and save the workspace (e.g., Lab-2.ows) **[20 points]**
2. Complete the tables given below and save the file (e.g., Lab-2.docx). **[80 points]**
3. Upload the files to the Canvas.

**Objective:** To review and understand the dataset attributes, attribute types, dimensionality and distribution of the attributes.

## **Problem 1/4. [20 points]**

**Data:** For this lab, please download *EIA\_appendixC\_2019.xlsx* from Canvas to your folder.

(**Reference:** The data is from the report titled U.S. Oil and Natural Gas Wells by Production Rate- <https://www.eia.gov/petroleum/wells/>.)

## **Lab Instructions**

1. Launch Orange.
2. Click on the **File** Widget under **Data** to add the widget to your blank Orange canvas. 🡪 Load the *EIA\_appendixC\_2019.xlsx* using the File widget.
3. Open File window by double clicking on **File**. 🡪 Answer the following questions for this data:

|  |  |
| --- | --- |
| How many objects are there in this dataset? | 16796 |
| What is the dimensionality (attribute) of this data? | 2 |
| What are the unique attribute types of this data? | Categorical, Numerical |

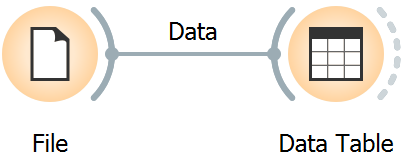
1. If necessary, rename the attribute names after inspecting the header of the Excel data file. For instance, compare the attribute name of column D of the Excel data file with the 4th attribute name in Orange. There is a mismatch. Let’s fix this by double clicking on it and change its name to the correct name.

|  |  |
| --- | --- |
| Before correction |  |
| After  correction |  |

1. Change attribute header of at least two attributes whose names are not matching with the header of the Excel data file. 🡪 Enter the header names before correction and after correction in the following table.

|  |  |  |
| --- | --- | --- |
| Attribute | Before Correction | After Correction |
| 12 | of gas wells | Number of gas wells |
| 5 | of oil wells | Number of oil wells |

1. Observe the time **year** attribute. Its attribute type is numeric. It is more appropriate to have it as datetime. Open the **File** Widget and change the attributes.
2. Click **Apply** button to save changes and close the **File** window.
3. Connect to a **Data Table** widget.



1. Open **Data Table** window by double clicking on **Data Table**. 🡪 Answer the following questions for this data:

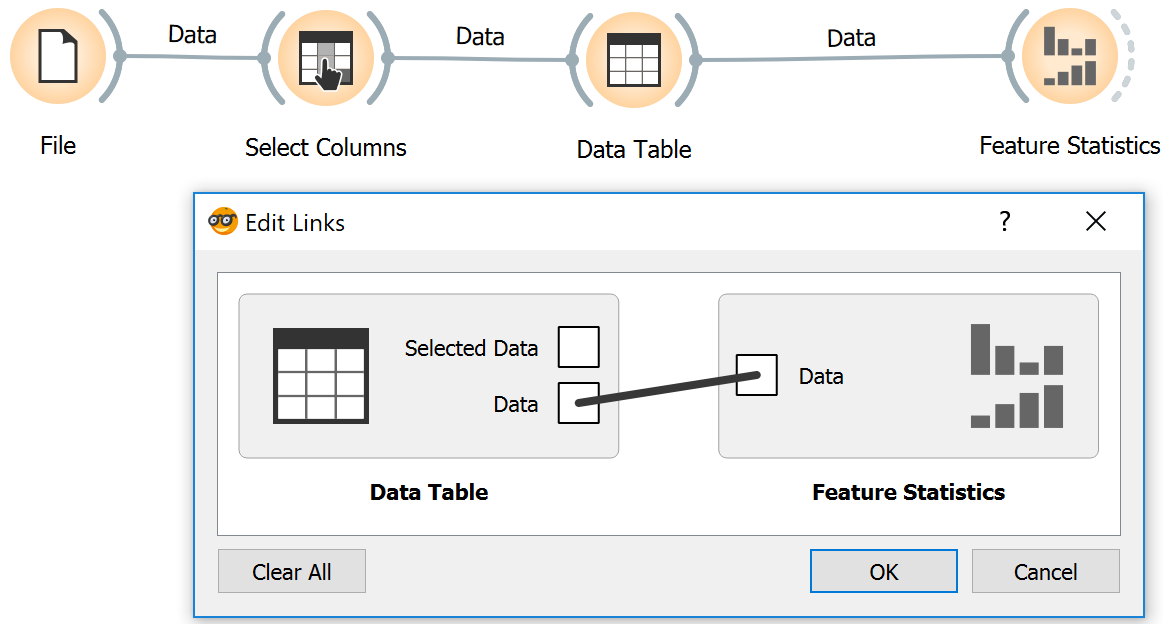
|  |  |
| --- | --- |
| Is this a structured dataset? | Yes |
| Are there any missing values? | No |
| What is the time resolution (frequency) of the dataset? | Annual resolution per State and per category |

1. Three of the attributes are derived from other attributes.
   * Total number of wells
   * Total wells: Annual gas prod. (Bcf)
   * Total wells: Annual oil prod. (MMbbl)

Most ML algorithms assumes that the attributes are independent. Dependent attributes may not be suitable for ML model building. 🡪 Remove dependent attributes by adding the **Select Columns** widget as shown below.



1. For further data exploration, add the **Feature Statistics** widget as shown below. Double click on the connection line between **Data Table** and **Feature Statistics** 🡪 connect **Data** boxes.



1. Open **Feature Statistics** window by double clicking on **Feature Statistics**. 🡪 Answer the following question:

|  |  |
| --- | --- |
| # | List pieces of information this GUI convey. |
| 1 | Mean |
| 2 | Median |
| 3 | Dispersion |
| 4 | Min |
| 5 | Max |

1. Let’s dig deeper by adding the **Distribution** widget as shown below. Double click on the connection line between **Data Table** and **Distribution** 🡪 connect **Data** boxes.

## 

1. Open **Distribution** window by double clicking on **Distribution**. 🡪 Answer the following questions:

|  |  |
| --- | --- |
| # | What additional information that **Distribution** convey over **Future Statistics**. |
| 1 | Can adjust the width size of histogram |
| 2 | Can filter different distributions |
| 3 | Gives parameters upper right as means and standard deviation |
| 4 | Color variables by, say, state |
| 5 | Stack columns = stack bar graph (like by states…) |
| 6 | Normal, Gamma, Beta, Rayleigh, Exponential, Kernel Density, Pareto |

|  |  |
| --- | --- |
| Attribute Name | Comment on the type of the distributions for three attributes of your interest. |
| Number of Wells | Skewed |
| Number of Oil Wells | Skewed |
| Horizontal Wells Count | Exponential and Skewed |
|  |  |

## **Problem 2/4. [20 points]**

**Data:** For this lab, please download *Log Lithology classification example.xlsx* from Canvas to your folder. We used this dataset for Lab-1.

## **Lab Instructions**

1. The same orange pipeline can be used to inspect various datasets. Let’s bring-in *Log Lithology classification example.xlsx*.
2. Perform the similar inspection for this data and answer the following questions.

|  |  |
| --- | --- |
| How many objects are there in this dataset? | 1001 |
| What is the dimensionality of this data? | 6 |
| What are the unique attribute types of this data? | Cate, Num |

1. In the **File** widget, change Lithology attribute’s Role to target.
2. Click **Apply** to save changes and close the **File** widget window. Use the **Data Table** widget to answer the following questions:

|  |  |
| --- | --- |
| Is this a structured dataset? | Yes |
| Are there any missing values? | No |
| What is the depth resolution of the dataset? | 0.5 ft per Lithology |

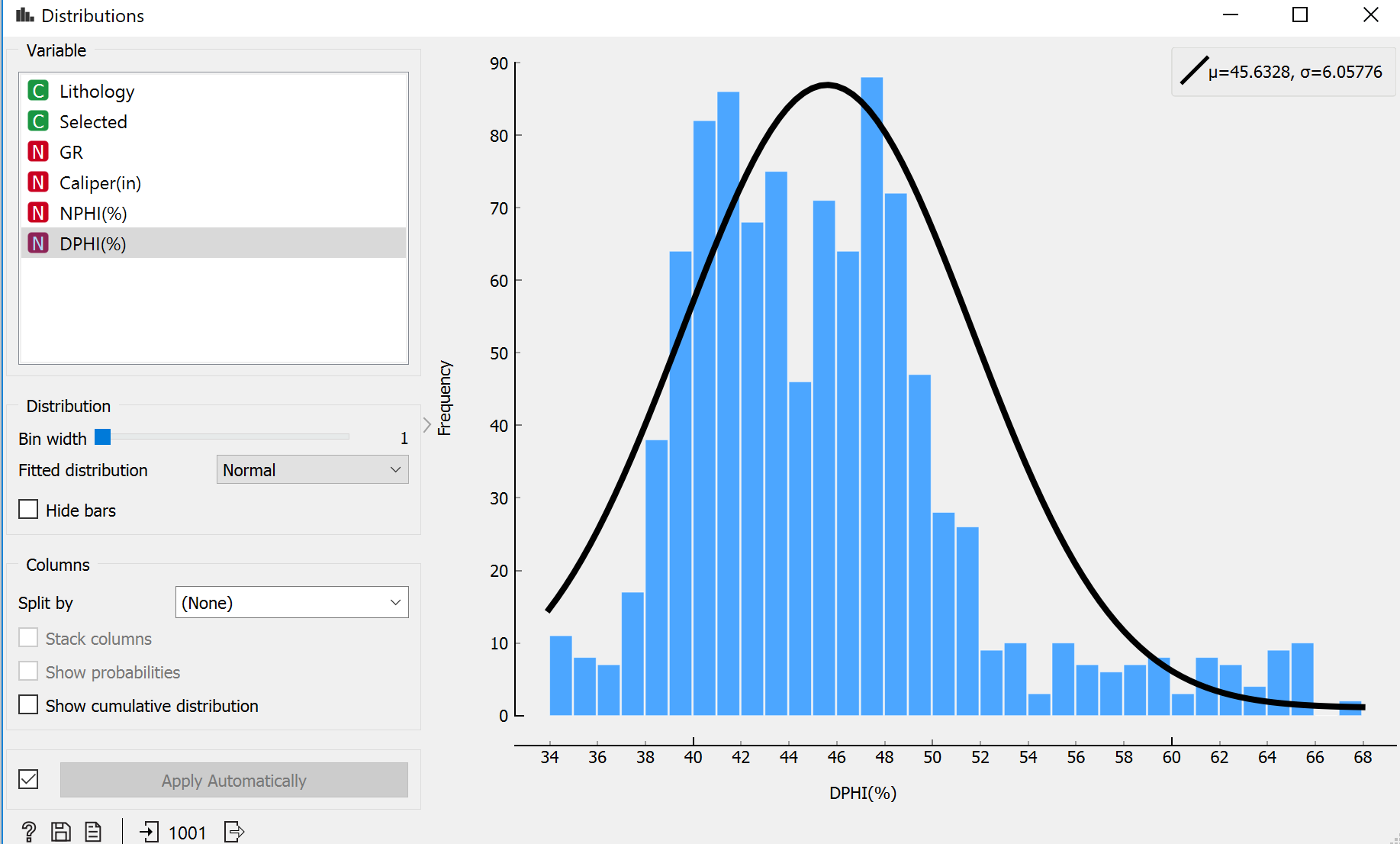
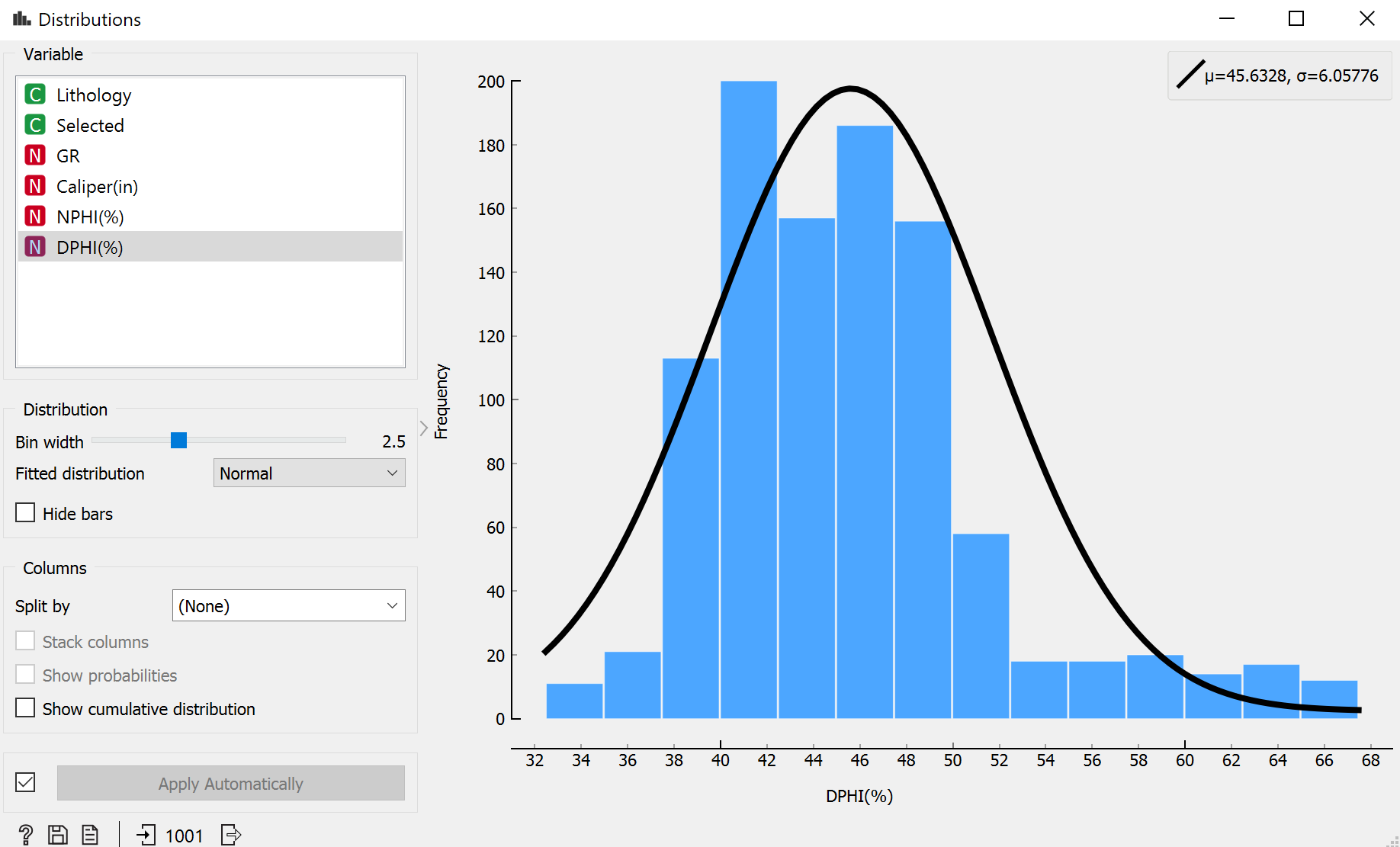
1. If necessary, remove any dependent attributes
2. Open **Feature Statistics** window by double clicking on **Feature Statistics**. 🡪 Answer the following question:

|  |  |
| --- | --- |
| Attribute Name | Write attribute range. |
| DPHI% |  |
| GR |  |
| Caliper(in) |  |
| NPHI(%) |  |

1. Open **Distribution** window by double clicking on **Distribution**. 🡪 Answer the following questions:

|  |  |
| --- | --- |
| Attribute Name | Comment on the type of the distributions for three attributes of your interest. |
| DPHI(%) | Skewed Distribution |
| GR | Skewed Distribution |
| Caliper | Skewed |
|  |  |

Sample visualizations of DPHI% distribution.

## **Problem 3/4. [20 points]**

**Data:** For this lab, please download *58-32\_xray\_diffraction\_data.csv* from Canvas to your folder.

(**Reference**: Utah FORGE Well Data, <https://gdr.openei.org/submissions/1111>)

## **Lab Instructions**

1. Load the *58-32\_xray\_diffraction\_data.csv*.
2. Perform the similar inspection for this data and answer the following questions.

|  |  |
| --- | --- |
| How many objects are there in this dataset? | 79 |
| What is the dimensionality of this data? | 15 |
| What are the unique attribute types of this data? | Cate, Num, Text |

1. Click **Apply** to save changes and close the **File** widget window. Use the **Data Table** widget to answer the following questions:

|  |  |
| --- | --- |
| Is this a structured dataset? | Yes |
| Are there any missing values? | Yes |
| What is the depth resolution of the dataset? | 30.5M |

1. If necessary, remove any dependent attributes
2. Open **Feature Statistics** window by double clicking on **Feature Statistics**. 🡪 Answer the following question:

|  |  |
| --- | --- |
| Attribute Name | Write attribute range of 4 attributes of your interest. |
|  |  |
|  |  |
|  |  |
|  |  |

1. Open **Distribution** window by double clicking on **Distribution**. 🡪 Answer the following questions:

|  |  |
| --- | --- |
| Attribute Name | Comment on the type of the distributions for three attributes of your interest. |
| Quartz | Normal Distribution |
| Plagioclase | Normal Distribution |
| K-feldspar | Normal Distribution |
|  |  |

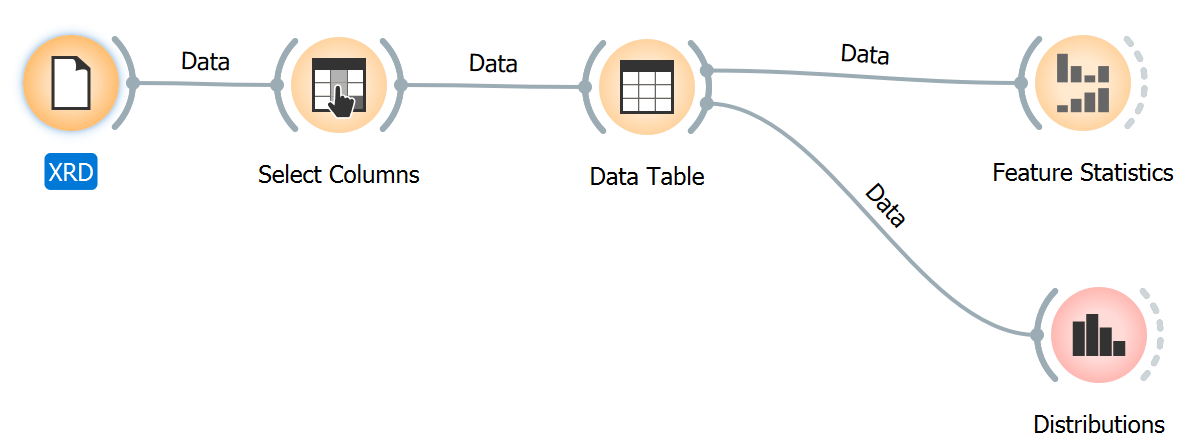
## **Problem 4/4. [20 points]**

**Data:** For this lab, please download *58-32\_thermal\_conductivity\_data.csv* from Canvas to your folder. We will use two datasets for this problem:

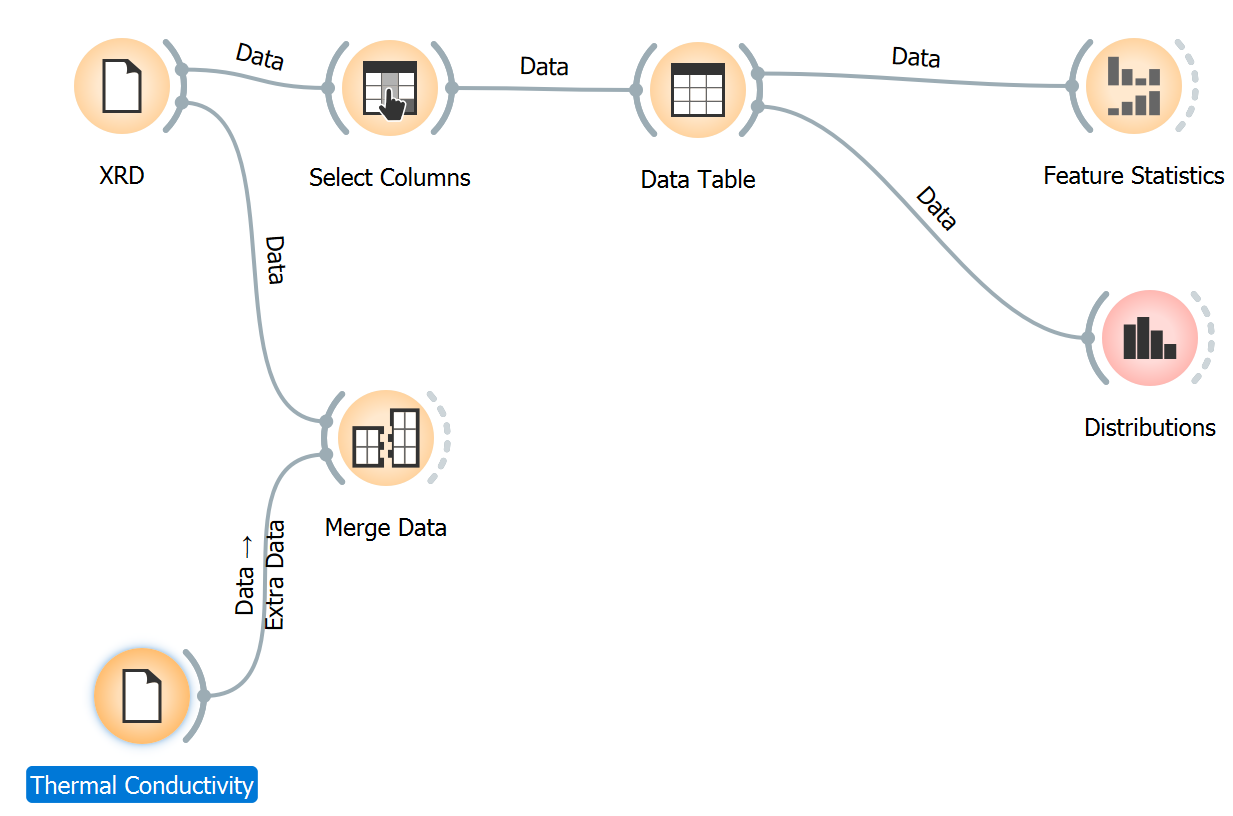
1. *58-32\_xray\_diffraction\_data.csv, and*
2. *58-32\_thermal\_conductivity\_data.csv*

## **Lab Instructions**

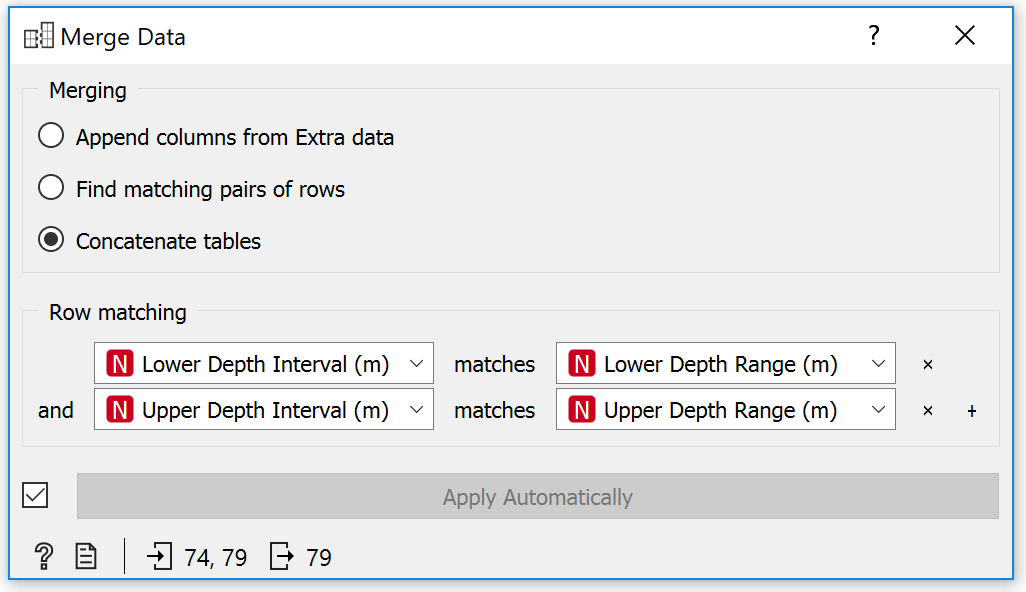
1. Right click on the **File** Widget and rename to *XRD*.



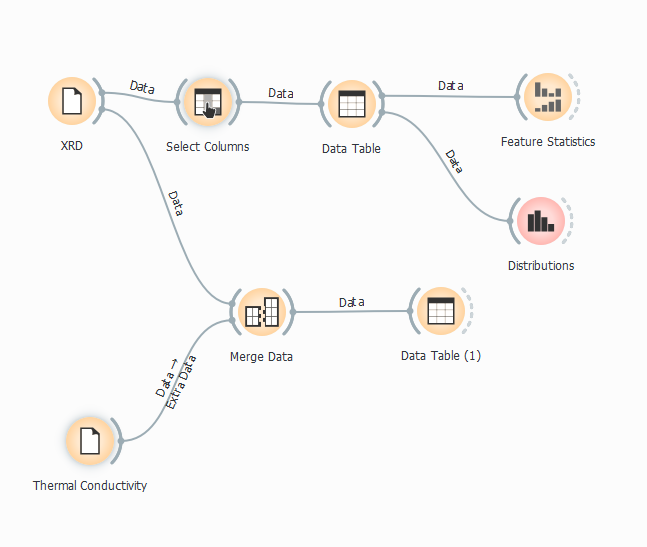
1. Click on the **File** Widget under **Data** to add the widget to your Orange canvas. 🡪 Load the *58-32\_xray\_thermal\_conductivity\_data.csv*.
2. Right click on the **File** Widget and rename to *Thermal Conductivity*.
3. Add **Merge** Data Widget 🡪 Connect
   1. *XRD* to **Merge** Data
   2. *Thermal Conductivity* to **Merge** Data



1. Make the following changes in the **Merge Data** Widget by double clicking on it. 🡪 Concatenate tables as shown below.



1. Add **Data Table** as shown below.



1. Open **Data Table** window by double clicking on **Data Table**. 🡪 Answer the following questions for this data:

|  |  |
| --- | --- |
| # | Write three observations from the Data Table. |
| 1 | 79 data instances |
| 2 | 18 Features |
| 3 | 36.6% missing data |
| 4 | 7 meta-attributes (65.3% missing data) |
| 5 | No target Variable |