Week 2 Practical Lab: Data Pre-processing and Statistical Analysis using KNIME

Useful resources:

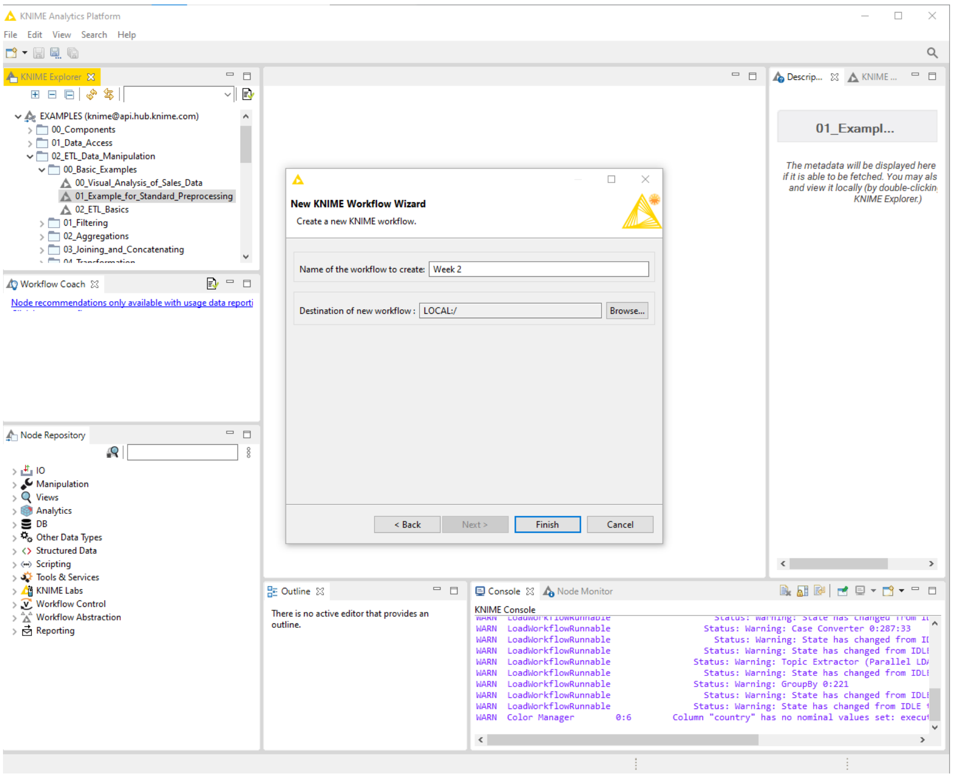
- **KNIME Analytics Platform** download link: https://www.knime.com/downloads/download-knime
  (It is recommender to download the installer version 561 MB for Windows)
- **Getting Started Guide** https://www.knime.com/getting-started-guide
- **KNIME Learning** https://www.knime.com/learning

Objective:

- Perform data preprocessing using KNIME
- Perform Statistical analysis using KNIME

Instructions:

1. Open the KNIME software installed on your computer.
2. Create a new workflow from File-> New -> **New KNIME Workflow** -> Next -> *Enter a project name* -> Finish.



New KNIME Workflow Wizard

Create a new KNIME workflow.

Name of the workflow to create: Week 2

Destination of new workflow : LOCAL:/

The metadata will be displayed here if it is able to be fetched. You may als and view it locally (by double-clickin KNIME Explorer.)

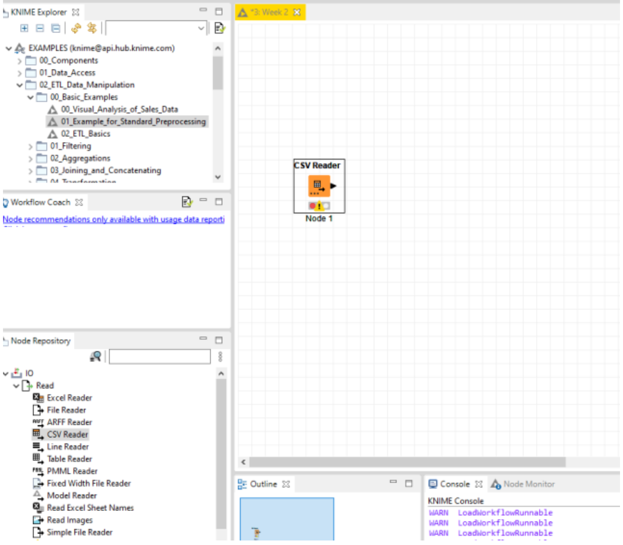
There is no active editor that provides an outline.

KNIME Console

```
WARN  LoadWorkflowRunnable           Status: Warning: Case Converter 0:287:33
WARN  LoadWorkflowRunnable            Status: Warning: State has changed from II
WARN  LoadWorkflowRunnable           Status: Warning: State has changed from IDLE
WARN  LoadWorkflowRunnable          Status: Warning: Topic Extractor (Parallel LDA
WARN  LoadWorkflowRunnable           Status: Warning: State has changed from IDLE
WARN  LoadWorkflowRunnable          Status: Warning: GroupBy 0:221
WARN  LoadWorkflowRunnable           Status: Warning: State has changed from IDLE
WARN  LoadWorkflowRunnable          Status: Warning: State has changed from IDLE t
WARN  Color Manager       0:6      Column "country" has no nominal values set: execut
```

3. Select a data set of your choice from the given link (https://archive.ics.uci.edu/ml/index.php) and open the file using appropriate node reader. (In this example, the following data set is chosen: https://archive.ics.uci.edu/ml/datasets/Early+stage+diabetes+risk+prediction+dataset.)

a. Drag and drop the appropriate node from the Node reposity in the workflow (for this example, CSV reader.)

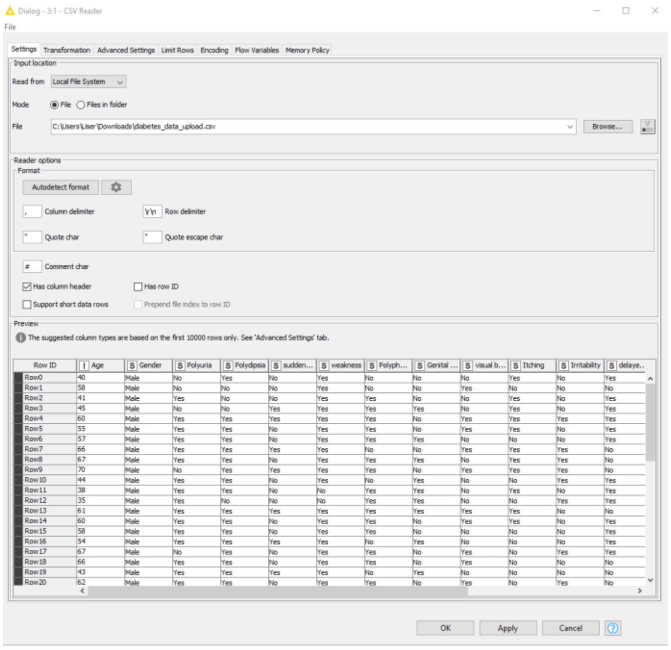


KNIME Console

```
WARN  LoadWorkflowRunnable
WARN  LoadWorkflowRunnable
WARN  LoadWorkflowRunnable
```

b. To configure the node, right click on the node and select **Configure….**

c. Browse and select your chosen file and click Okay if the data preview is accurate.



Input location

Read from: Local File System

Mode: ● File ○ Files in folder

File: C:\Users\User\Downloads\diabetes_data_upload.csv

Reader options

Format

, Column delimiter \r\n Row delimiter

" Quote char " Quote escape char

# Comment char

☑ Has column header □ Has row ID

□ Support short data rows □ Prepend file index to row ID

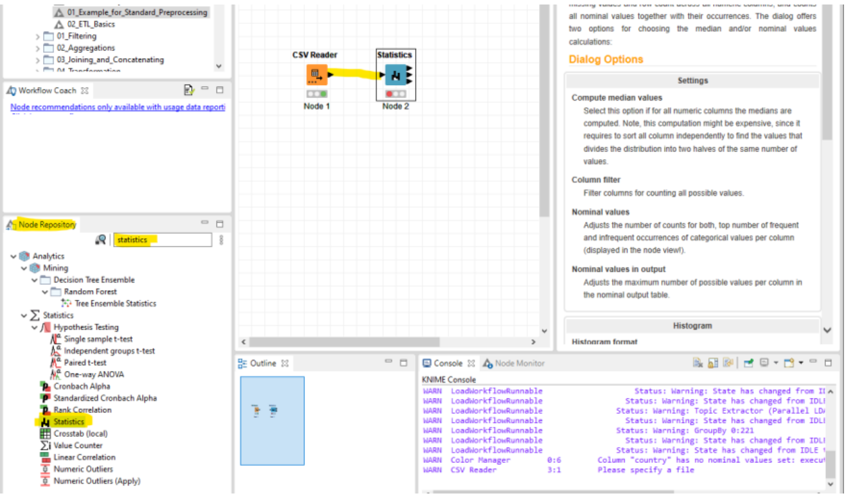
Preview

The suggested column types are based on the first 10000 rows only. See 'Advanced Settings' tab.

Row ID	Age	Gender	Polyuria	Polydipsia	sudden...	weakness	Polyph...	Genital ...	visual b...	Itching	Irritability	delaye..
Row0	40	Male	No	Yes	No	Yes	No	No	No	Yes	No	Yes
Row1	58	Male	No	No	No	Yes	No	No	Yes	No	No	No
Row2	41	Male	Yes	No	No	Yes	Yes	No	No	Yes	No	Yes
Row3	45	Male	No	No	Yes	Yes	Yes	Yes	No	Yes	No	Yes
Row4	60	Male	Yes	Yes	Yes	Yes	Yes	No	Yes	Yes	Yes	Yes
Row5	55	Male	Yes	Yes	No	Yes	Yes	No	Yes	Yes	No	Yes
Row6	57	Male	Yes	Yes	No	Yes	Yes	Yes	No	No	No	Yes
Row7	66	Male	Yes	Yes	Yes	Yes	No	No	Yes	Yes	Yes	No
Row8	67	Male	Yes	Yes	No	Yes	Yes	Yes	No	Yes	Yes	No
Row9	70	Male	No	Yes	Yes	Yes	Yes	No	Yes	Yes	Yes	No
Row10	44	Male	Yes	Yes	No	Yes	No	Yes	No	No	Yes	Yes
Row11	38	Male	Yes	Yes	No	No	Yes	Yes	No	Yes	No	Yes
Row12	35	Male	Yes	No	No	No	Yes	Yes	No	No	Yes	Yes
Row13	61	Male	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	No	No
Row14	60	Male	Yes	Yes	No	Yes	Yes	No	Yes	Yes	No	Yes
Row15	58	Male	Yes	Yes	No	Yes	Yes	No	No	No	No	Yes
Row16	54	Male	Yes	Yes	Yes	Yes	No	Yes	No	No	No	Yes
Row17	67	Male	No	Yes	No	Yes	Yes	No	Yes	No	Yes	Yes
Row18	66	Male	Yes	Yes	No	Yes	Yes	No	Yes	No	No	No
Row19	43	Male	Yes	Yes	Yes	Yes	No	Yes	No	No	No	No
Row20	62	Male	Yes	Yes	No	Yes	Yes	No	Yes	No	Yes	No

d. Execute the node by clicking right on the node and then select **Execute**.

4. Select the **Statistics** node from the **Node Reposity** and connect with the **File Reader node** to understand on which columns or rows you need to perform pre-precessing. Then execute the node.



all nominal values together with their occurrences. The dialog offers two options for choosing the median and/or nominal values calculations:

Dialog Options

Settings

Compute median values
Select this option if for all numeric columns the medians are computed. Note, this computation might be expensive, since it requires to sort all column independently to find the values that divides the distribution into two halves of the same number of values.

Column filter
Filter columns for counting all possible values.

Nominal values
Adjusts the number of counts for both, top number of frequent and infrequent occurrences of categorical values per column (displayed in the node view!).

Nominal values in output
Adjusts the maximum number of possible values per column in the nominal output table.

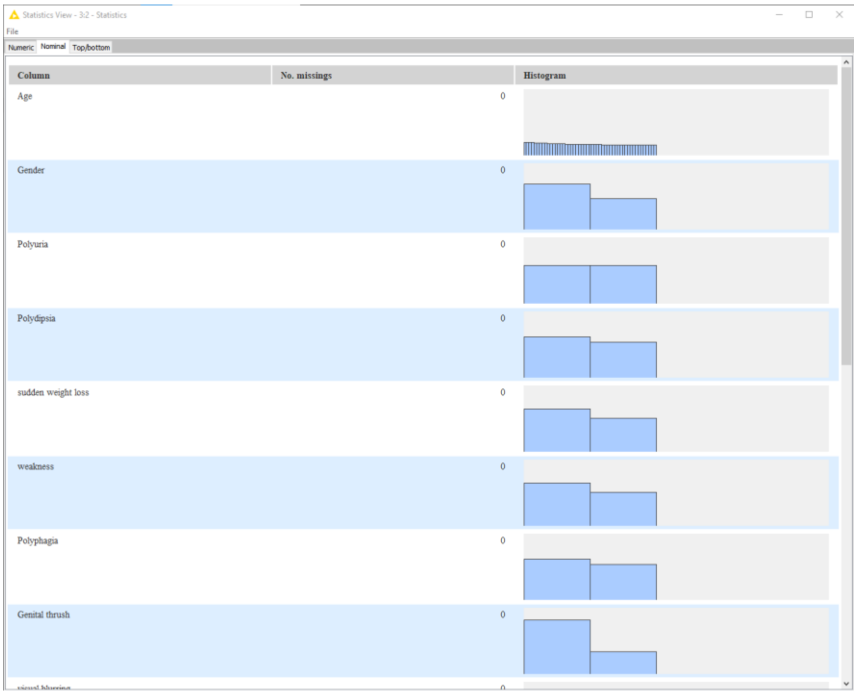
Histogram

Histogram format

KNIME Console

```
WARN  LoadWorkflowRunnable              Status: Warning: State has changed from II
WARN  LoadWorkflowRunnable             Status: Warning: State has changed from IDLI
WARN  LoadWorkflowRunnable            Status: Warning: Topic Extractor (Parallel LD/
WARN  LoadWorkflowRunnable             Status: Warning: State has changed from IDLI
WARN  LoadWorkflowRunnable            Status: Warning: GroupBy 0:221
WARN  LoadWorkflowRunnable             Status: Warning: State has changed from IDLI
WARN  LoadWorkflowRunnable            Status: Warning: State has changed from IDLE 1
WARN  Color Manager        0:6       Column "country" has no nominal values set: execu
WARN  CSV Reader           3:1       Please specify a file
```

5. After executing the node, right click and select "View: Statistics View" to explore the statistical measures.



Column	No. missings	Histogram
Age	0	
Gender	0	
Polyuria	0	
Polydipsia	0	
sudden weight loss	0	
weakness	0	
Polyphagia	0	
Genital thrush	0	
visual blurring	0	

6. If there is any missing value in your data set, then select the "Missing Value" node as part of the pre-processing and connect it with the File Reader node.
   a. Go to Configuration mode for the missing value node and select appropriate rules for different data types.
   b. For column specific rules, Select the Column Settings tab.

7. Explore other pre-processing nodes (Convert & Replace, Filter, Split & Combine, Normaliser etc.) and discuss the suitability with your Lecturer.