Open for Innovation
KNIME

[L4-DV] Low Code Data Extraction and Visualization

17 April, 2023

Course objectives

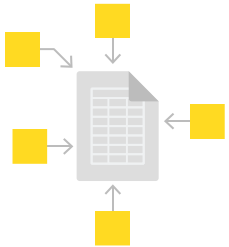
Once you have completed this course, you will be able to:

- Collect data via **REST APIs**, **web text scraping**, and an **interactive** data collection **tool**
- Explore and visualize data
- Extract data and images from **PDF** documents
- Write regular expressions (**regex**)
- Identify and correct errors in data via **outlier detection**
- Build effective and **beautiful visuals**

Who is this course for? Data and business analysts

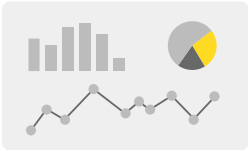
Data Analyst	Data Scientist	Data Engineer
Data acquisition, cleaning, analysis, visualizations, descriptive statistics, reporting, dashboards.	Data pre-processing, training machine learning and statistics algorithms, modeling, predicting.	Integrating various data sources, building data pipelines (ETL, ELT), databases, data lakes, data warehouses, file systems, and/or data mart maintenance, monitoring and testing.

Data analytics within an organization



Data Collection

Gathering data from different sources

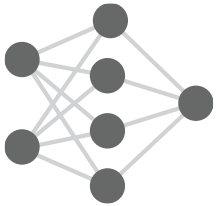


Data Visualization

Data visualization and statistics

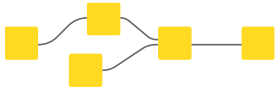
Data Preprocessing

Cleaning the data; checking quality



Machine Learning

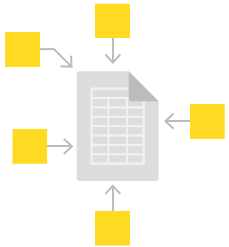
Modeling and evaluation



Deployment for Users

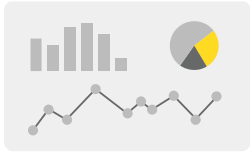
Production model

Data analytics for this course



Data Collection

Gathering data from different sources



Data Visualization

Data visualization and statistics

Data Preprocessing

Cleaning the data; checking quality

Structure of the course

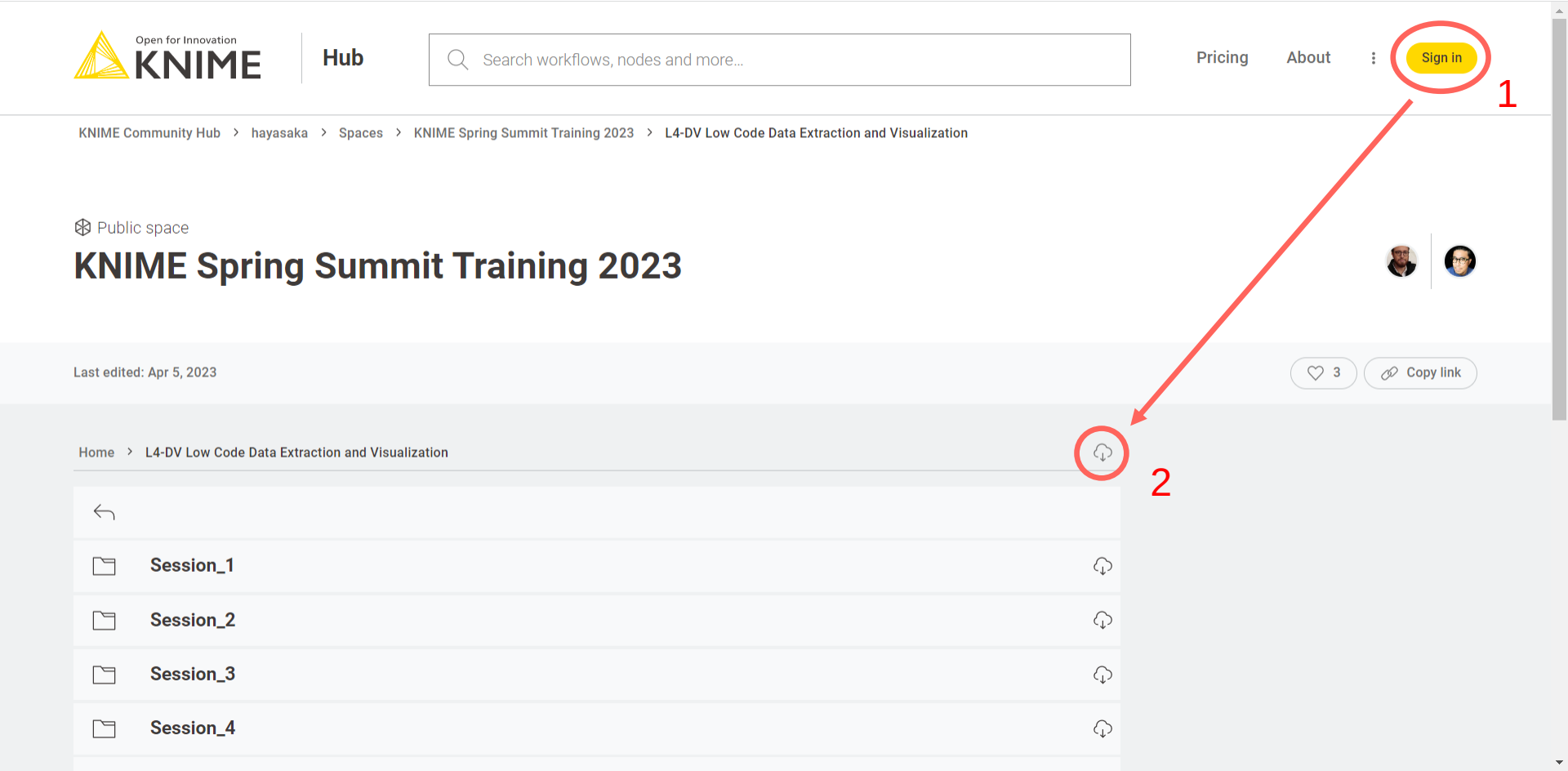
This course consists of five sections:

1. Data Collection
2. Data Visualization
3. Data Extraction
4. Data Quality and Visualization Best Practices

Each section offers demo workflows and exercises.

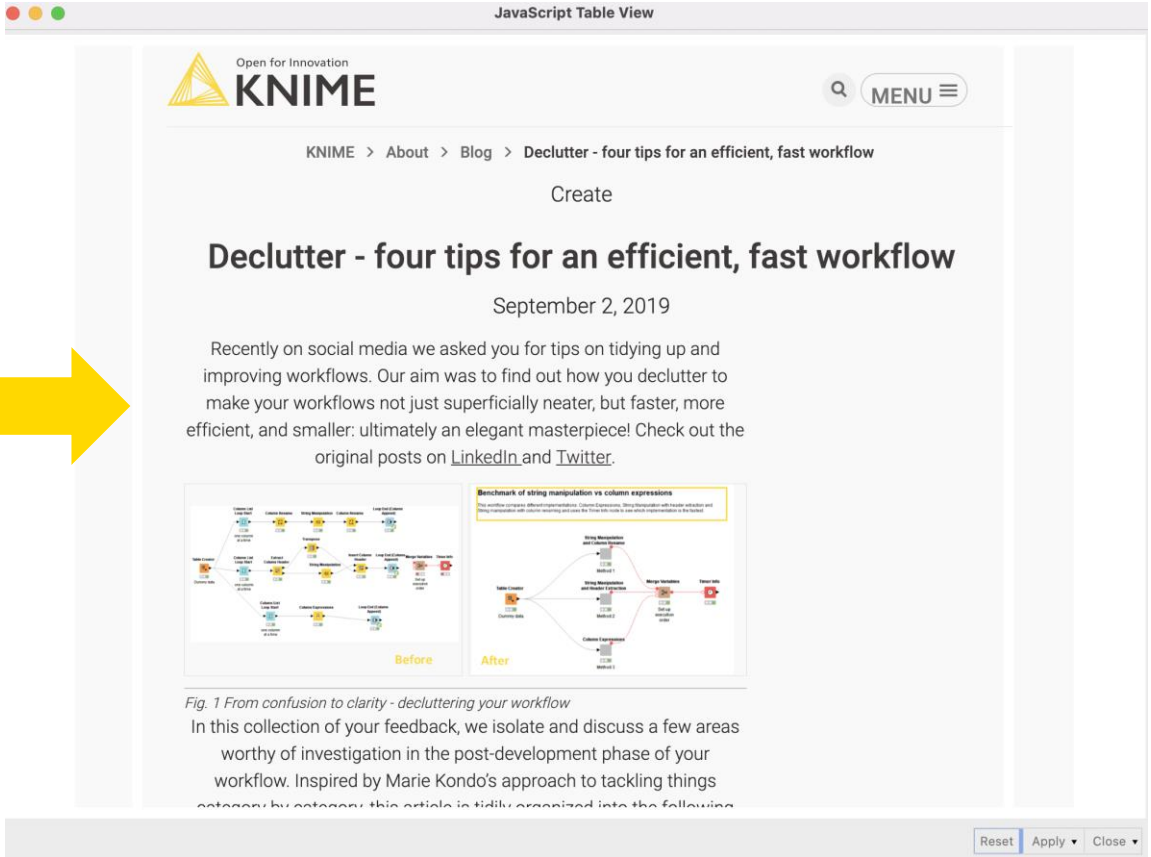
Find all materials at: *tinyurl.com/L4DV-SpringSummit*

How to Download Materials

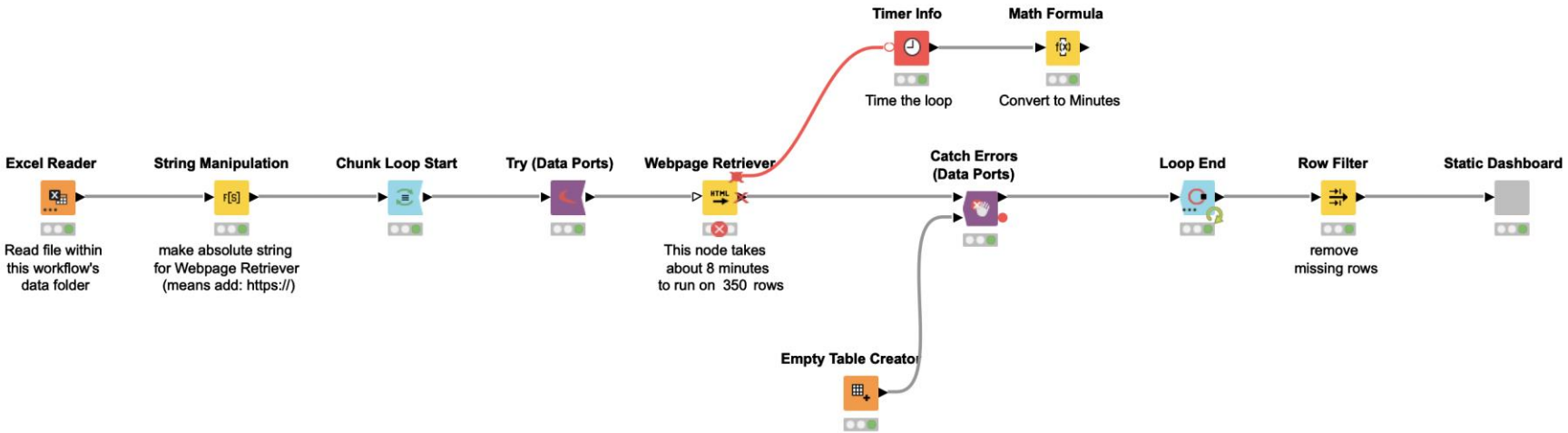


Section 1: From links to data

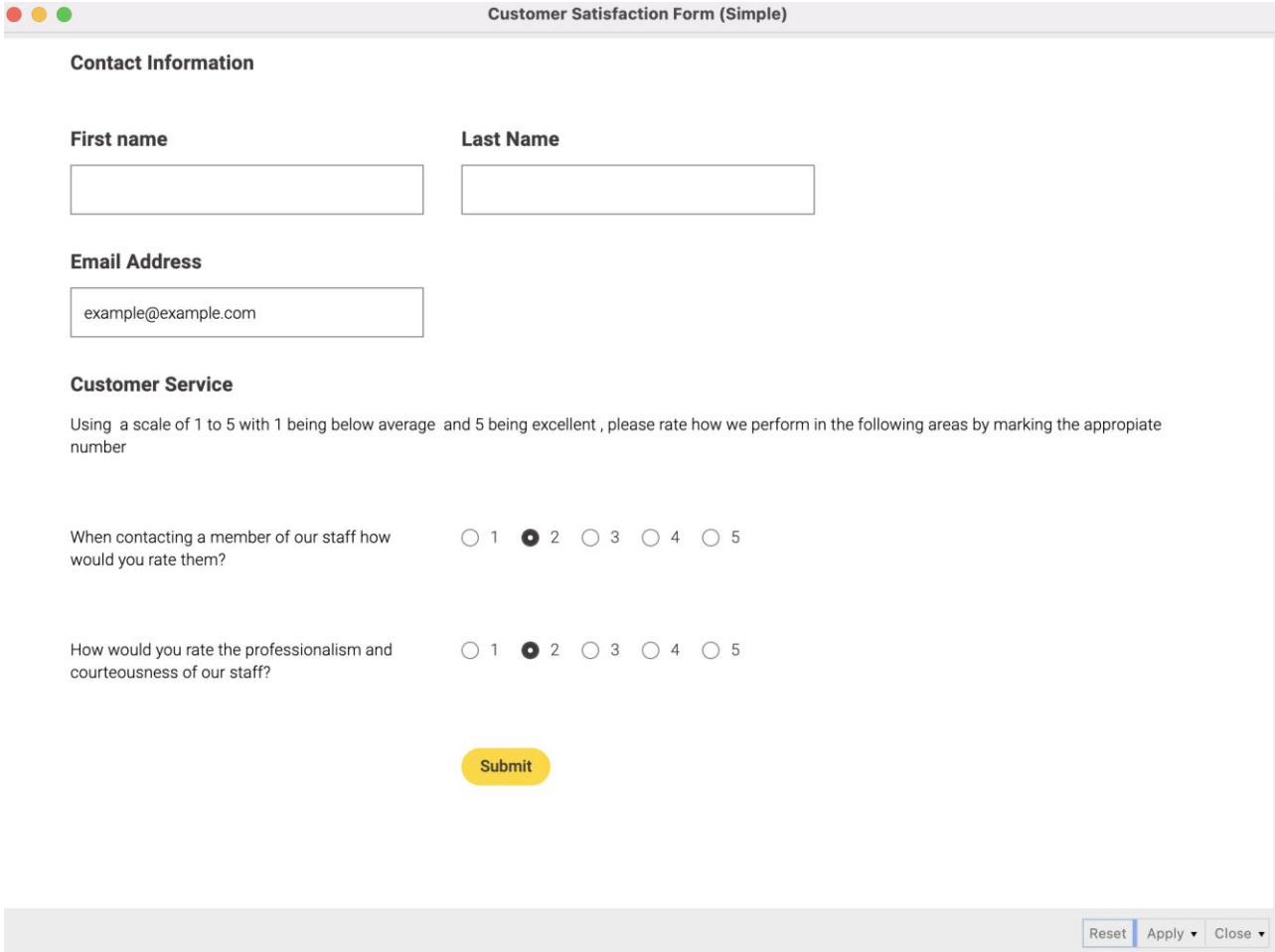
www.knime.com/blog/declutter-four-tips-for-an-efficient-fast-workflow



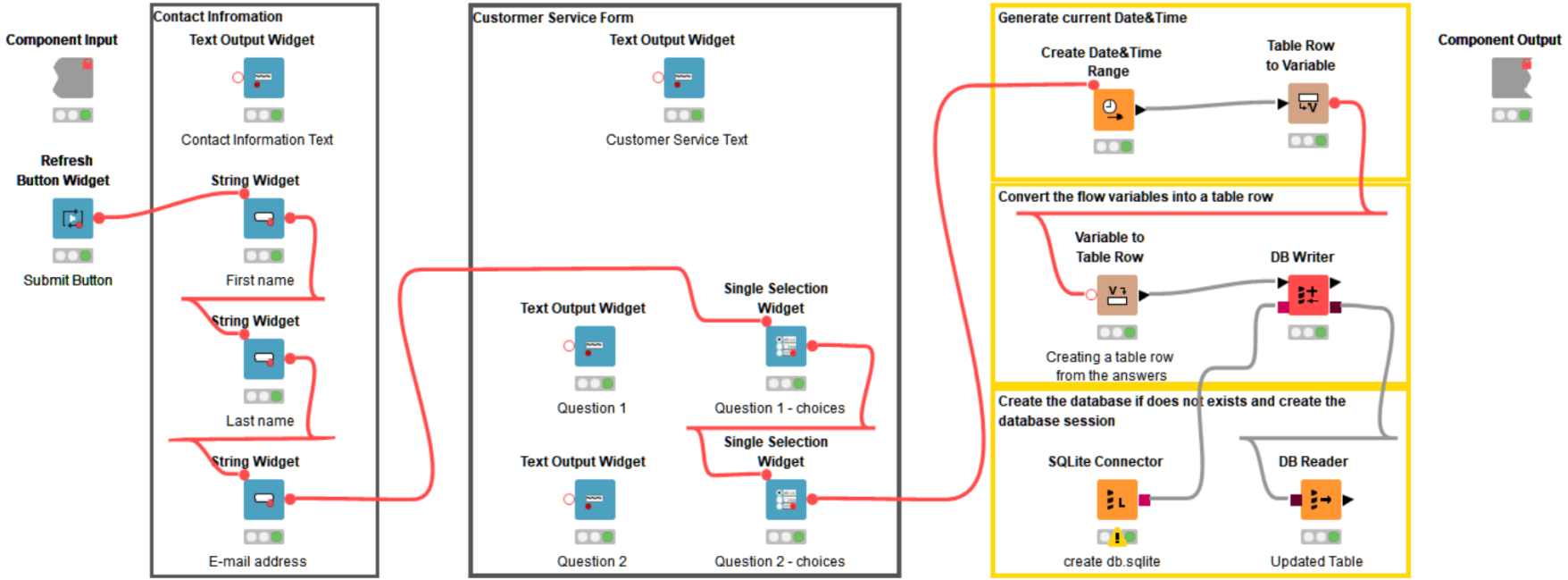
Section 1: From links to data



Section 1: Customer satisfaction form

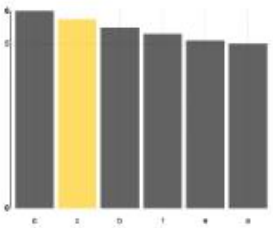


Section 1: Customer satisfaction form

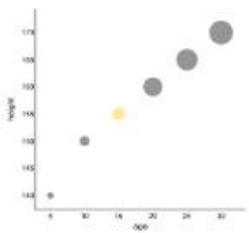


Section 2: Common visualization tasks

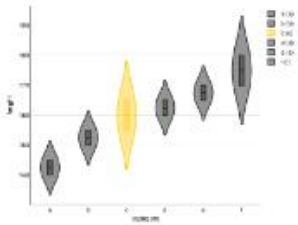
Comparison



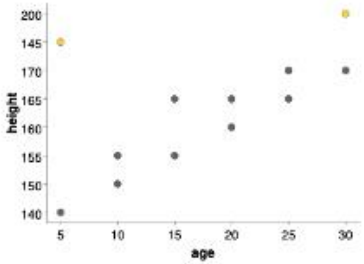
Correlation



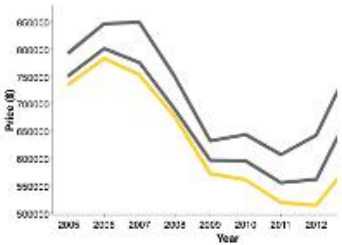
Distribution



Outliers

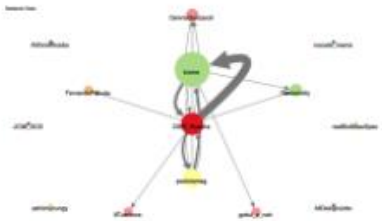


Time



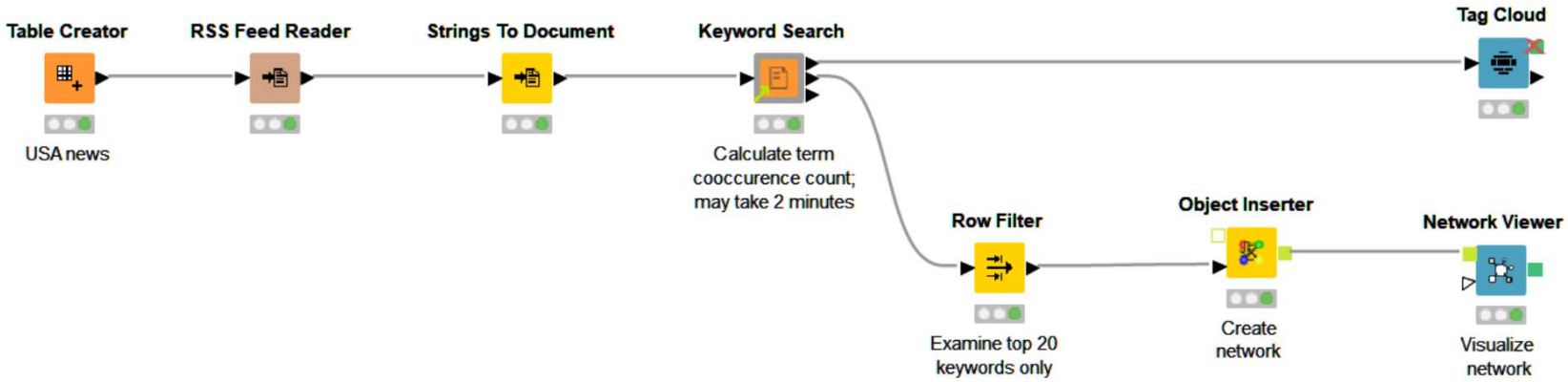
Text

Networks



Geography

Section 2: Common visualization tasks



Section 3: PDF parsing with Regex

Extracting with Regex

Choose column with name "Text" and any other columns to display.

Excludes

Includes

Document
Text

Type Regex Below:

[0-9]{4}-[0-9]{1,2}-[0-9]{1,2}

Click to re-execute regex

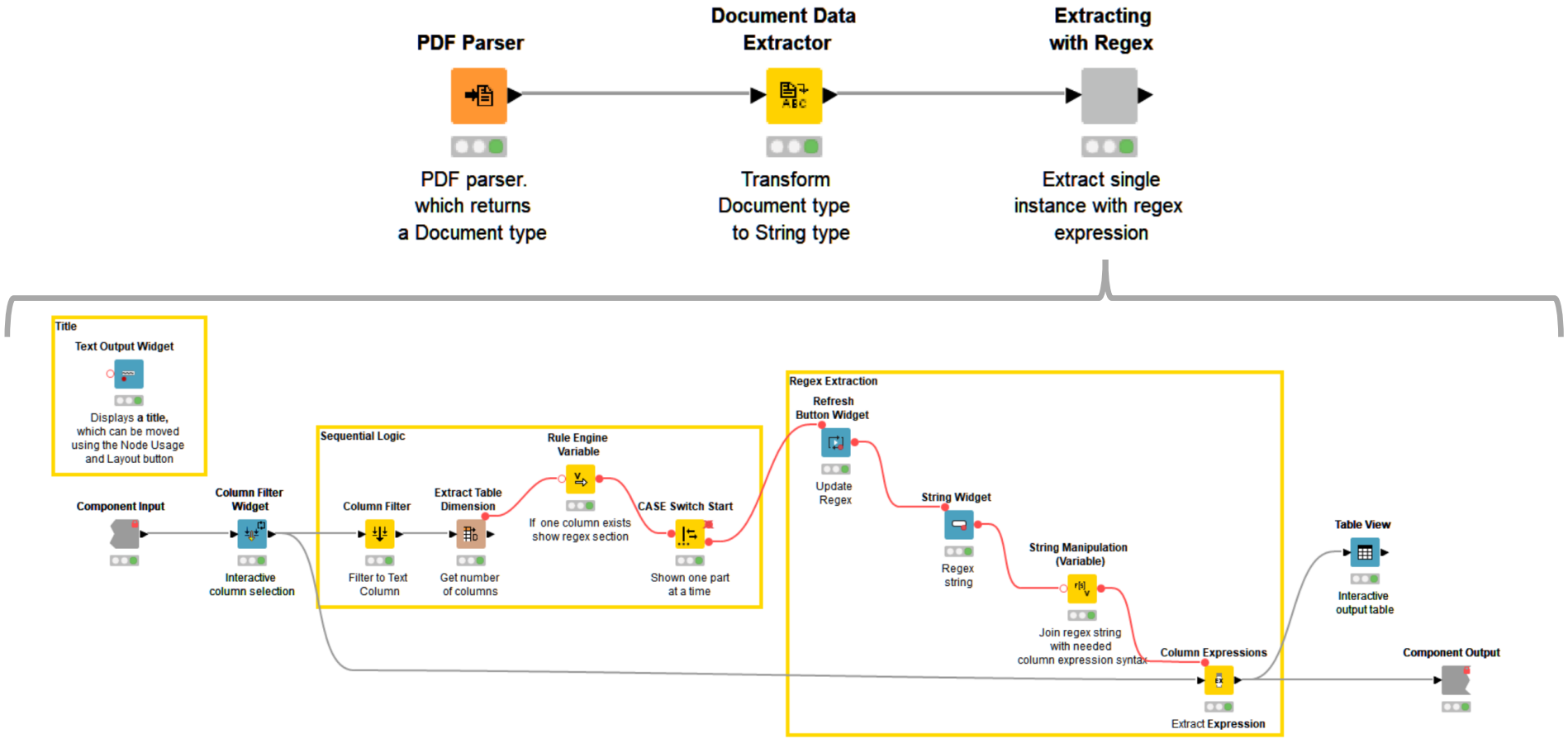
Show 10 entries

Search:

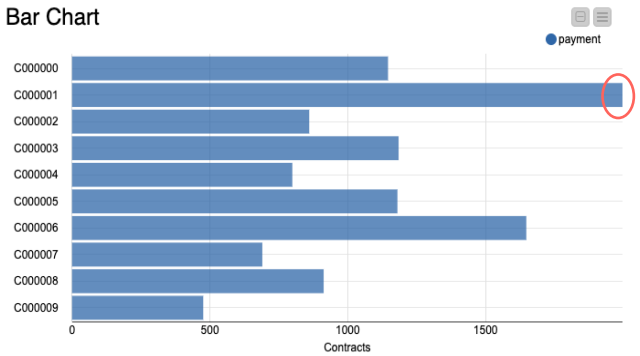
Document	Text	Captured Expression
"C:\Users\daria.liakh\knime-workspace\L4-DV Low Code Data Extraction and Visualization\Session_3\00_Demos\03_Parsing_PDFs_Demo\..\..\..\data\pdfs\11012.pdf"	Investment Agreement C000004 KNOW ALL MEN BY THESE PRESENTS : This Contract is entered by and between : Bank A GmbH , whose principal place of address is at Debussystr . 18,Konstanz , Germany , 78460 , (hereinafter referred to as "	2016-04-05

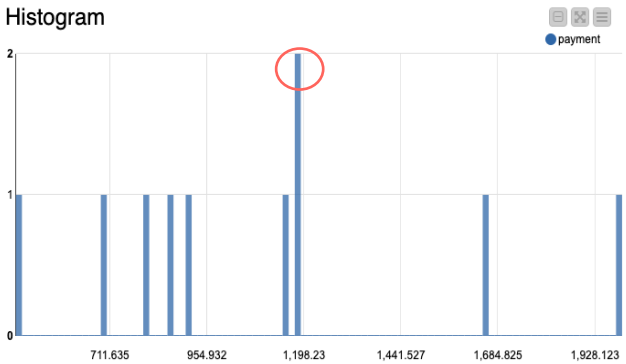
Open for Innovation KNIME

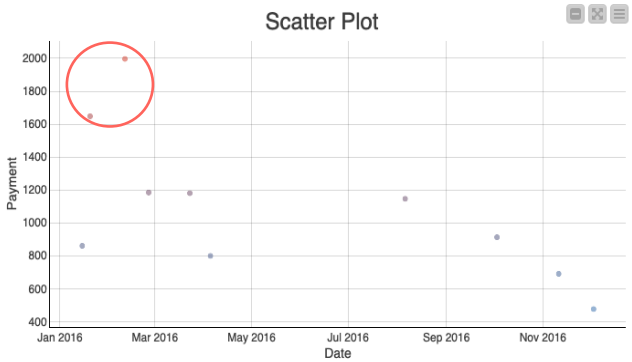
Section 3: PDF parsing with Regex

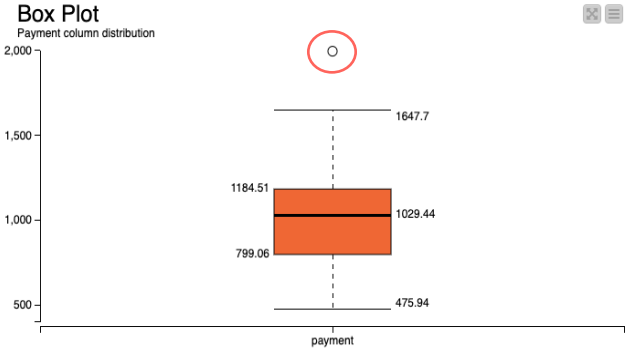


Section 4: Anomaly detection

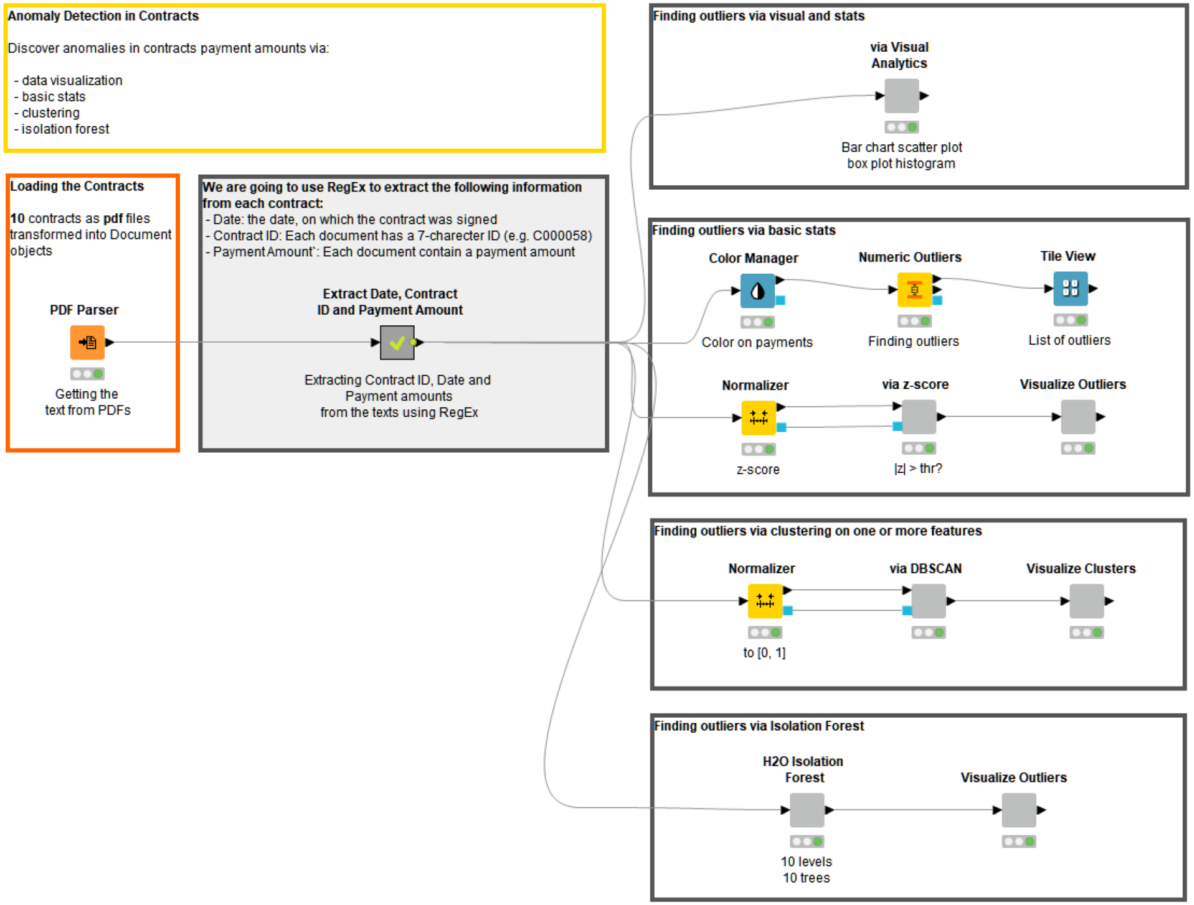




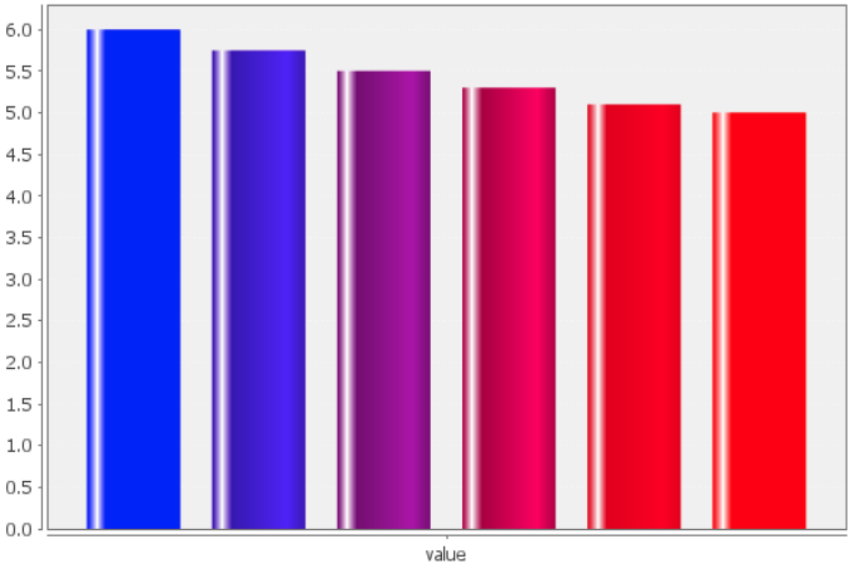


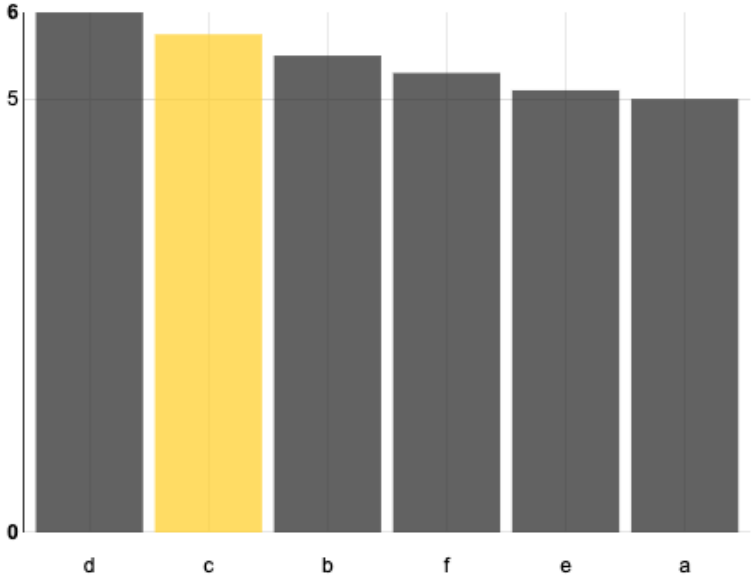


Section 4: Anomaly detection



Section 4: Visualization best practices





Open for Innovation
KNIME

[L4-DV] Low Code Data Extraction and Visualization

Section 1

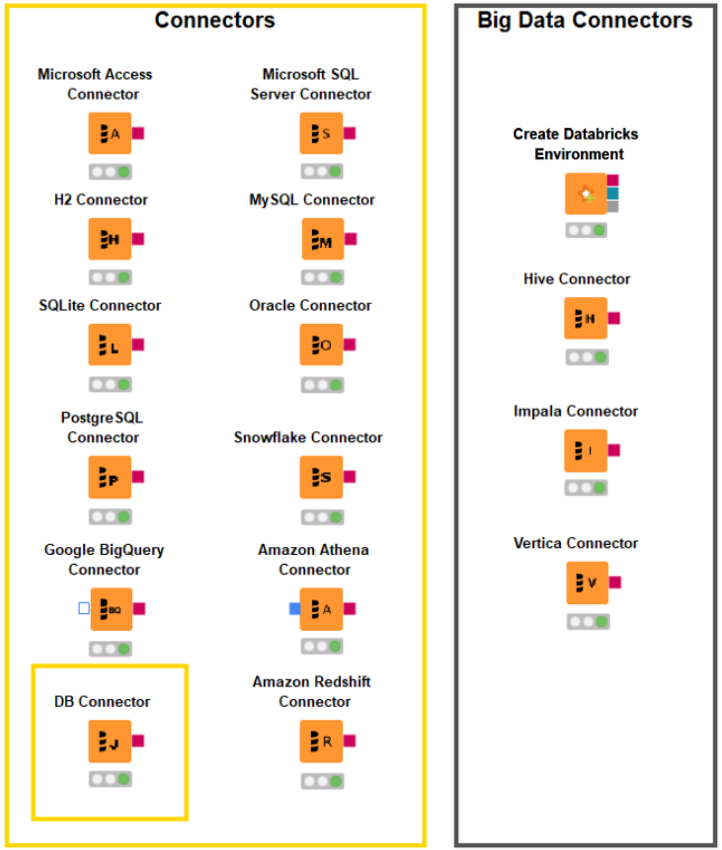
Data Collection

At the end of this section, you will be able to:

1. Recognize data access nodes.
2. Perform webpage retrieval.
3. Differentiate between widgets.
4. Build a data collection tool.

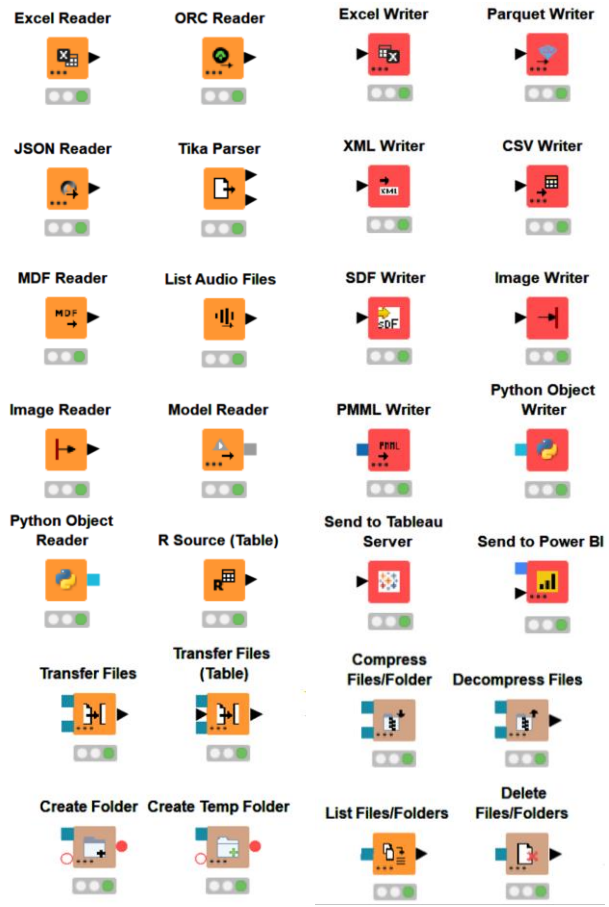
Database nodes

- Dedicated nodes to connect to specific Databases
- Hive and Impala connector part of the KNIME Big Data Connectors extension
- General Database Connector
  - Register new JDBC driver via File -> Preferences -> KNIME -> Databases



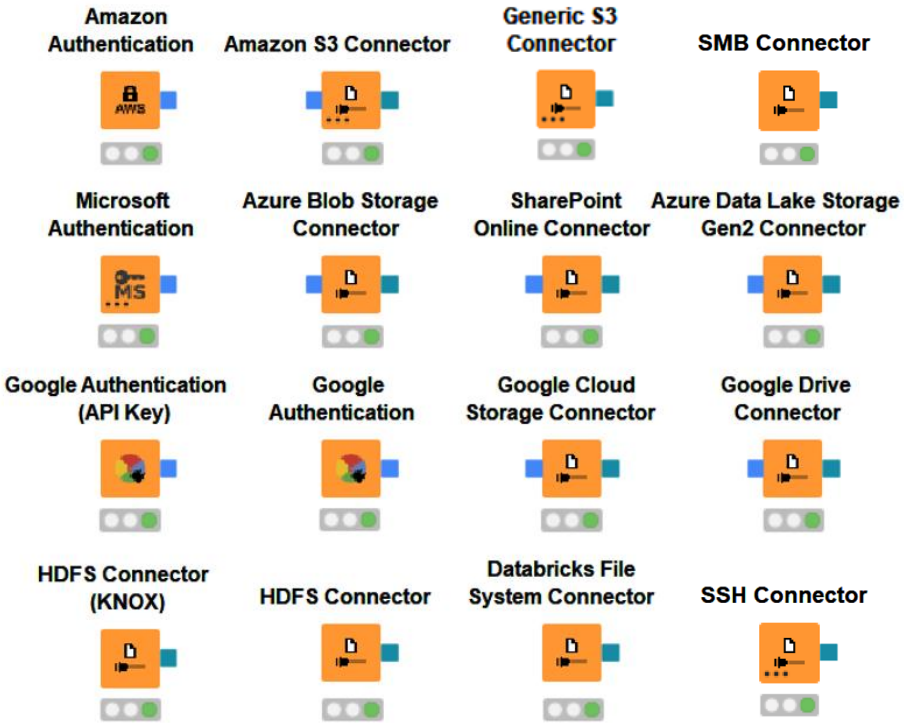
Reader/Writer/Utility nodes

- Reading/writing **tabular, structured, textual**, chemical data, audio, image, and model files
- Reading one or **multiple** files
- Support of integrations: **Python**, R, H2O, **PowerBI, Tableau**



Authentication/Connector nodes

- Connected file systems
  - File systems with external authentication
    - Amazon
    - Microsoft
    - **Google**
  - File systems without external authentication
    - Databricks
    - HDFS, HttpFS
    - SSH, HTTP(S), KNIME Business Hub

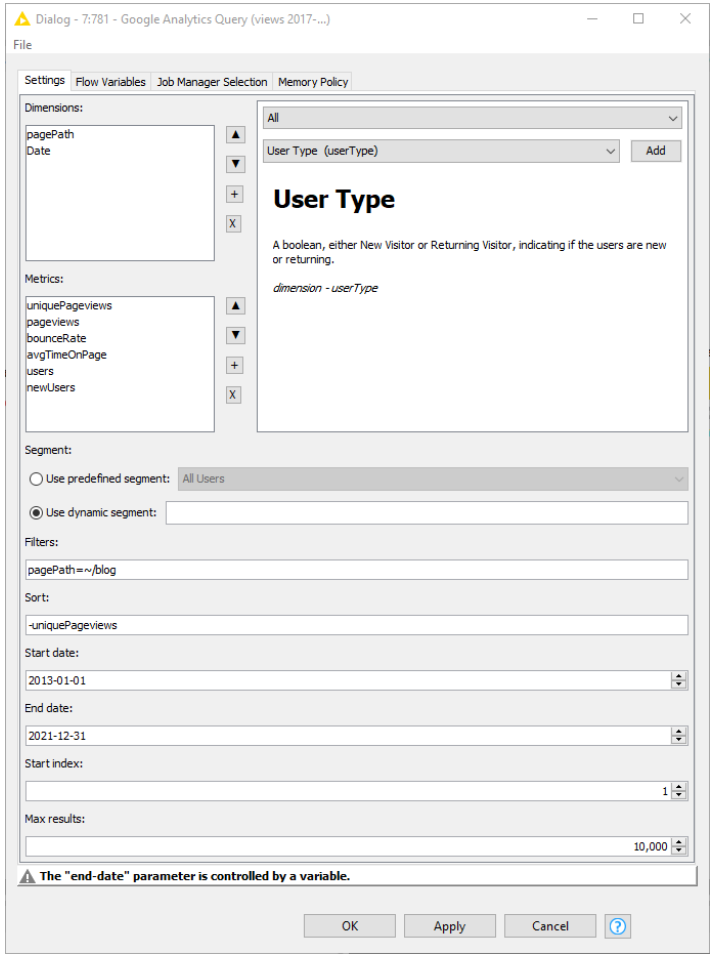


Google analytics extension

- Access Google Analytics data
  - Assemble the queries to extract metrics and dimensions of interest
- No need to write API requests

Select metrics and dimensions, filter, restrict period, sort, etc.

Authenticate via pop-up window (Oauth2)



Google sheets

- Access data stored in Google Services:
  - Read data from Google Sheets
  - Transform in KNIME
  - Modify existing sheets
- No need to write API requests

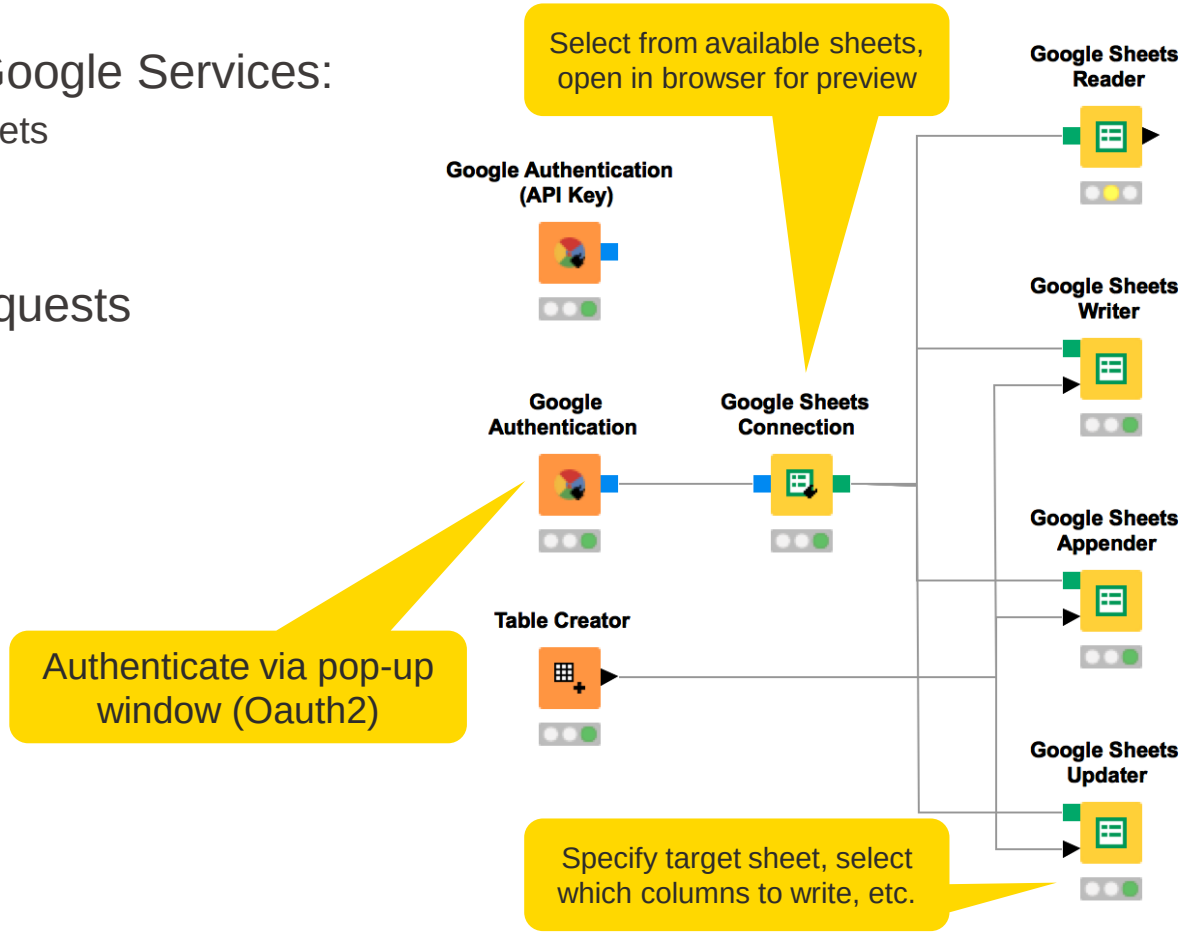
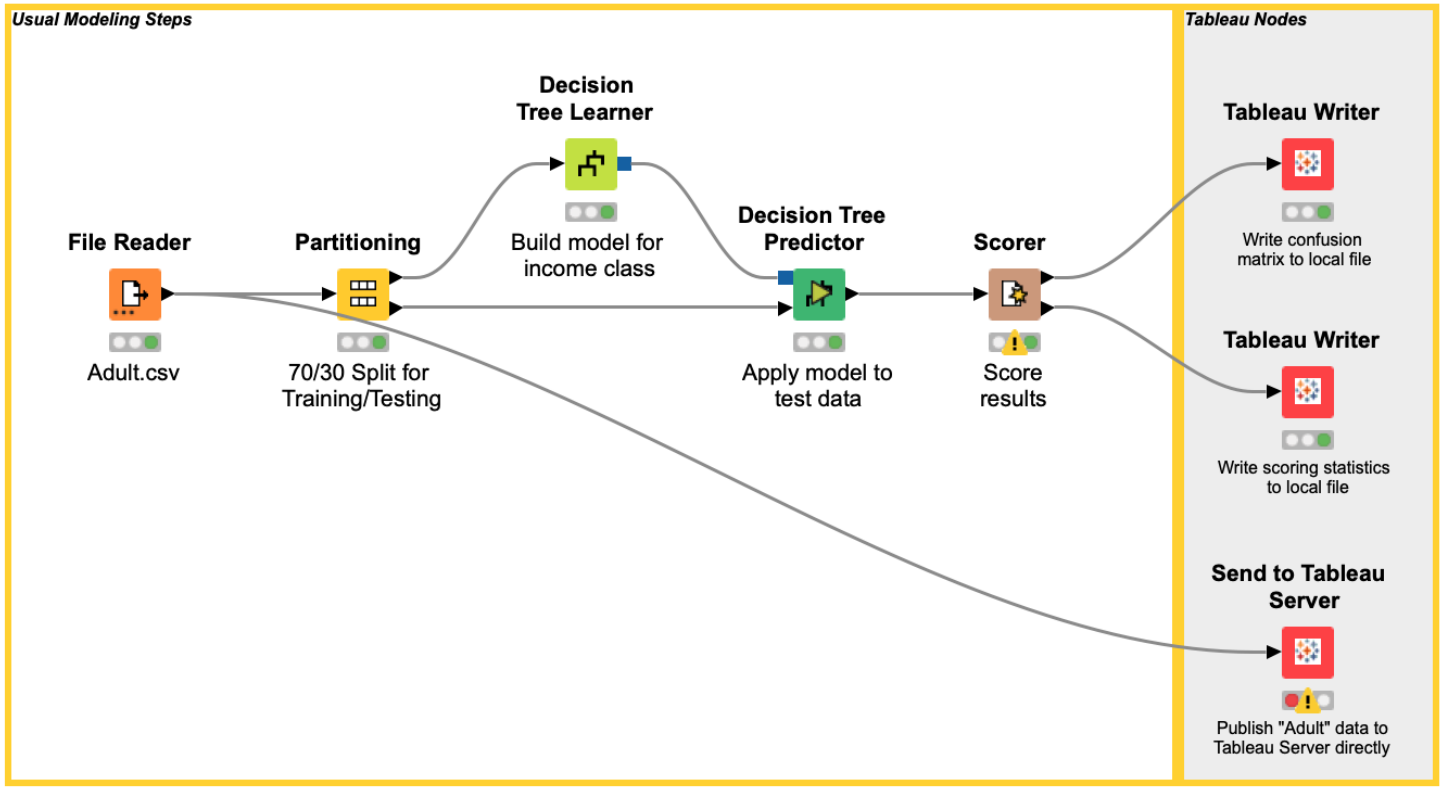


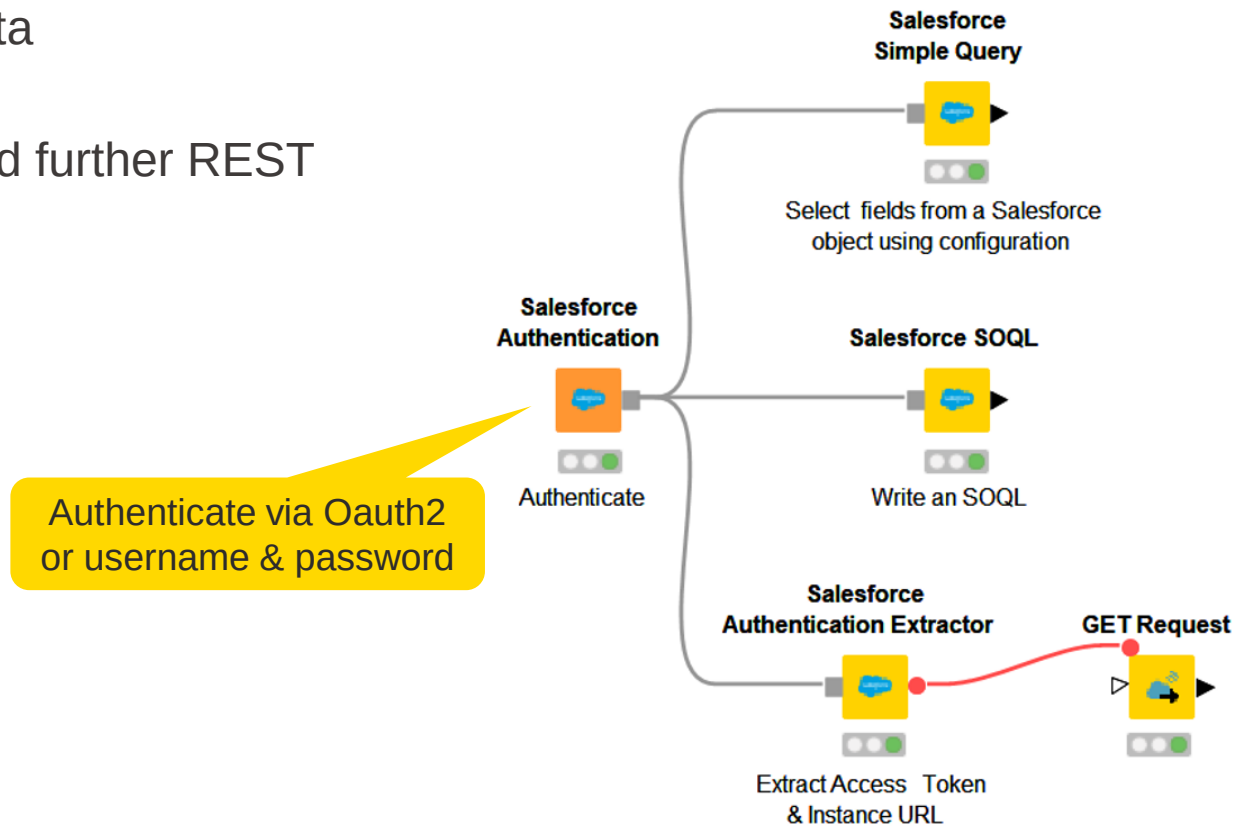
Tableau extension

- Write tableau files/send to Tableau server



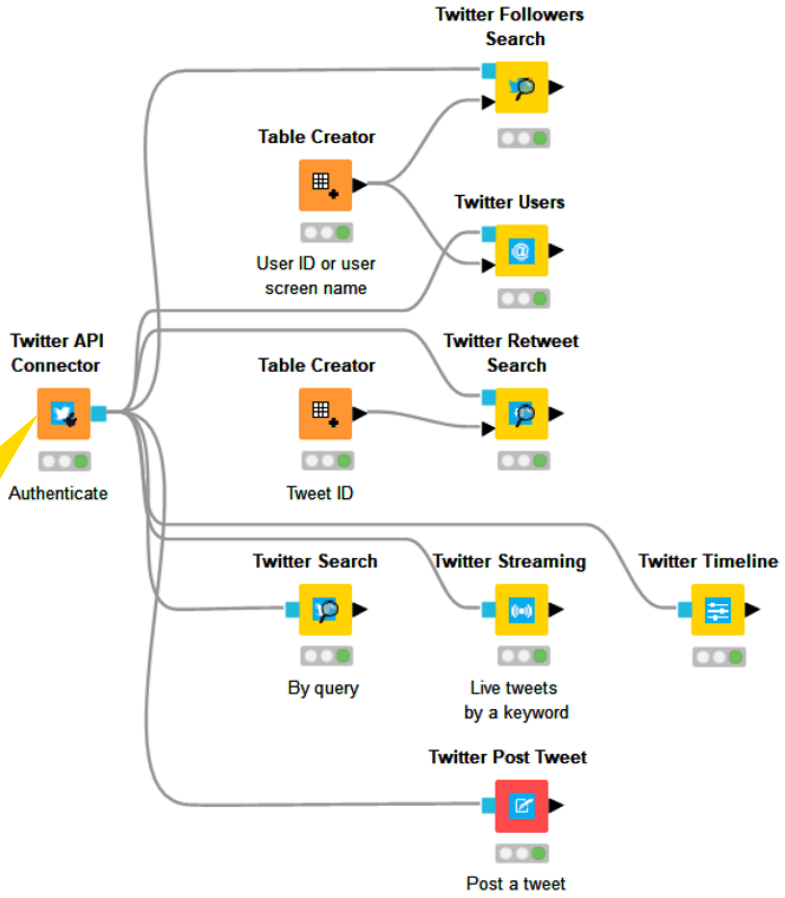
Salesforce extension

- Access Salesforce data
- Write queries and build further REST requests



Twitter extension

- Connect to Twitter Developer Account and
  - Search users and followers by user ID or screen name
  - Search retweets by tweet ID
  - Return the timeline
- No need to write API requests

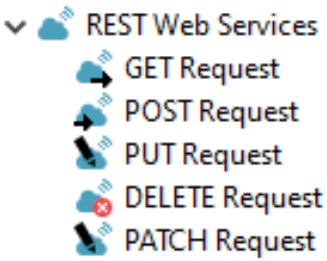


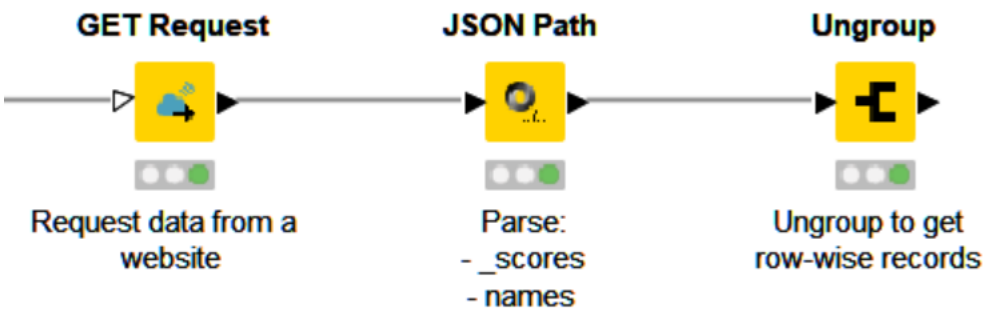
Provide API key, API secret, Access token, & Access token secret

RESTful API extension

- Execute **Re**presentational **S**tate **T**ransfer commands
  - RESTful APIs are Web Service APIs that adhere to the REST constraints
  - One the most predominant **architectures for obtaining and managing data** across applications

- Existing KNIME nodes





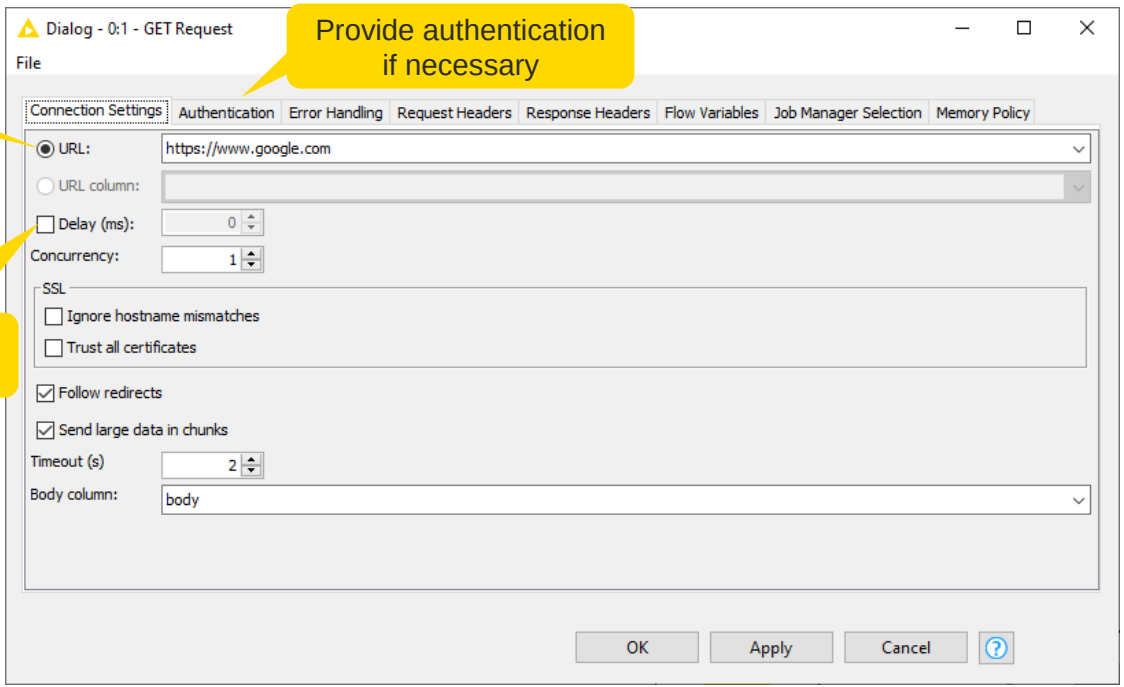
RESTful web services / API

GET Request

Enter URL, or use from column

Add delay between individual requests

Provide authentication if necessary



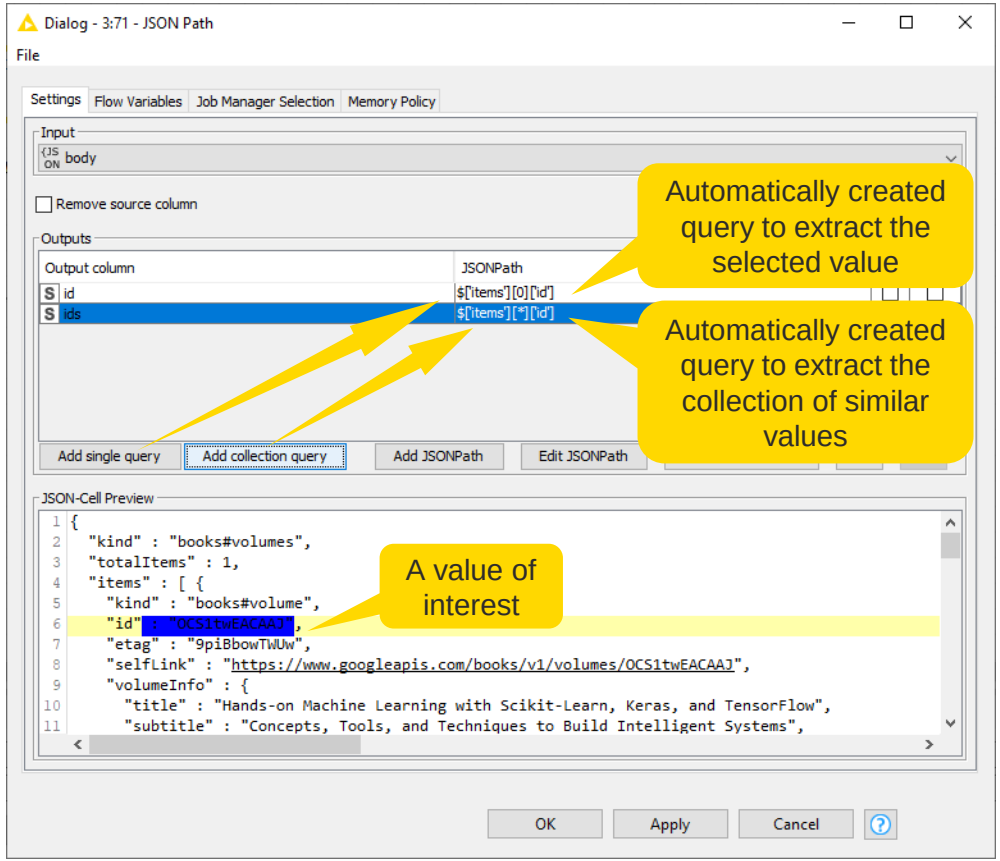
https://www.youtube.com/watch?v=HeE7tEgUdq0

https://www.knime.com/blog/a-restful-way-to-find-and-retrieve-data

https://www.knime.com/blog/OSM-meets-CSV-file-and-Google-API

JSON and XML parsing tips

- Use the JSON Path node to query the JSON file and extract parameters
- Editor window simplifies construction of JSON queries by auto-generating them
  - Select the value of interest in the JSON-Cell Preview and use the buttons to automatically add a query to extract this single value or a collection of similar values
  - OR write a JSONPath query manually
- Analogously with Xpath node for XML



KNIME knowledge check 1

- All the extensions mentioned today (Salesforce, Google, Twitter, etc.) come automatically installed in the KNIME Analytics Platform.
  - True
  - False

01: Get Request node demo

Live
Demo

Course exercises

- All exercises are available here: tinyurl.com/L4DV-SpringSummit

Education

Last edited: 25 Jun 2020

Home > Courses

- L1-DS KNIME Analytics Platform for Data Scientists - Basics
- L1-DW KNIME Analytics Platform for Data Wranglers - Basics
- L1-LS KNIME Analytics Platform for Data Scientists - Life Sciences - Basics
- L2-DS KNIME Analytics Platform for Data Scientists - Advanced
- L2-DW KNIME Analytics Platform for Data Wranglers - Advanced
- L2-LS KNIME Analytics Platform for Data Scientists - Life Sciences - Advanced
- L3-PC KNIME Server Course - Life Sciences - Productionizing and Collaboration
- L3-PC KNIME Server Course - Productionizing and Collaboration
- L4-BD Introduction to Big Data with KNIME Analytics Platform
- L4-CA Machine Learning for Chemical Applications
- L4-CH Introduction to Working with Chemical Data
- L4-DE Best Practices for Data Engineering
- L4-DL Introduction to Deep Learning
- L4-DV Low Code Data Extraction and Visualization
- L4-ML Introduction to Machine Learning Algorithms
- L4-TP Introduction to Text Processing

Public space

KNIME Spring Summit Training 2023

Last edited: Apr 5, 2023

Home > L4-DV Low Code Data Extraction and Visualization

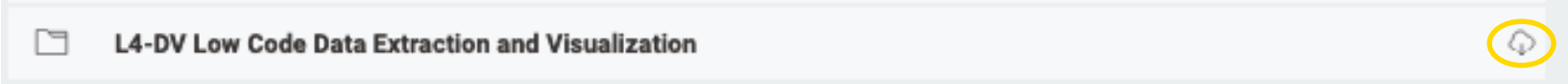
- Session_1
- Session_2
- Session_3
- Session_4
- ata
- Agenda-L4-DV.pdf
- Extension_Requiring_Nodes
- Slides-L4-DV.pdf

Open for Innovation KNIME

Download material

- Download the course material from the KNIME Community Hub here:

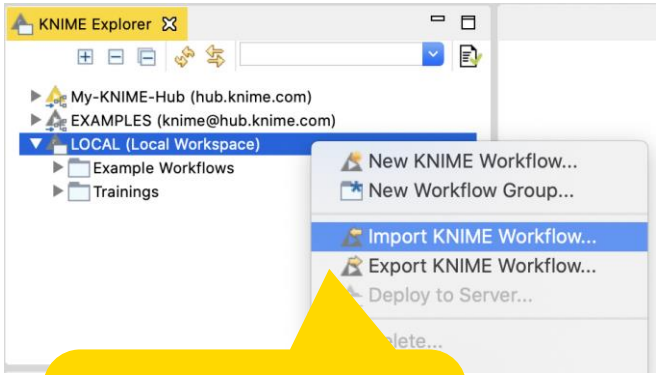
  tinyurl.com/L4DV-SpringSummit



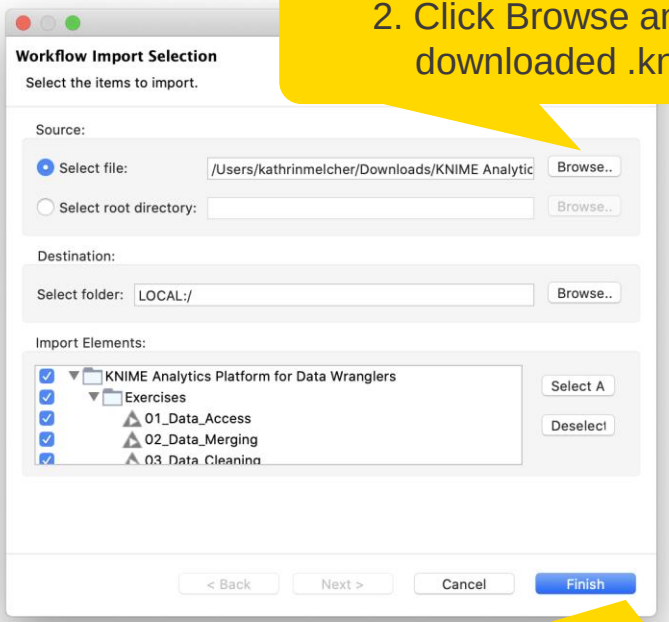
Note: You must be logged into the KNIME Hub to see that download icon for all course material

Import material

- Import the course material to KNIME Analytics Platform



1. Right click LOCAL and select Import KNIME Workflow….



2. Click Browse and select downloaded .knar file

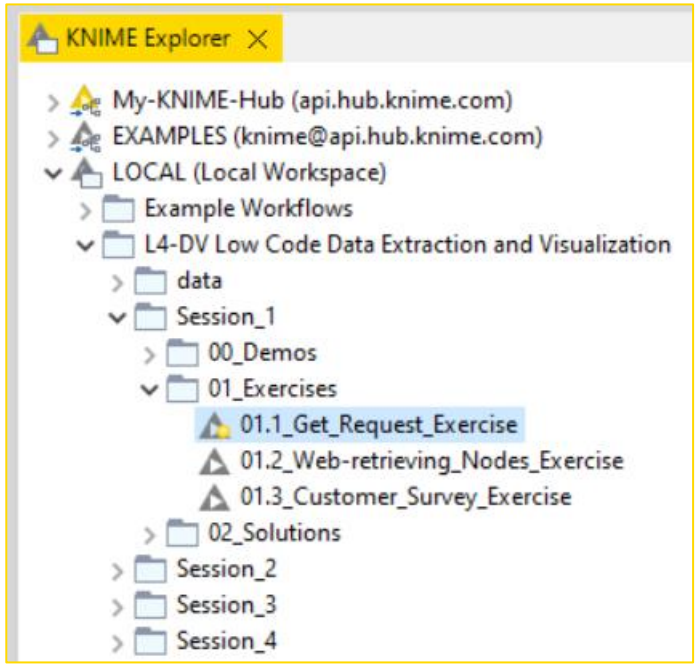
3. Click Finish

Exercises: Section 1

1. Get Request Exercises
   Given a table with retrieve information with the Get Request node.

2. From Links to Data Exercise
   Given a list of URLs, acquire the text from each link.

3. Customer Survey Exercise
   Create your own customer feedback form.



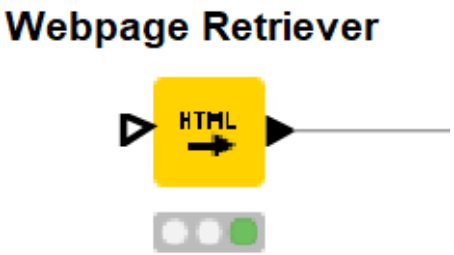
Data Collection

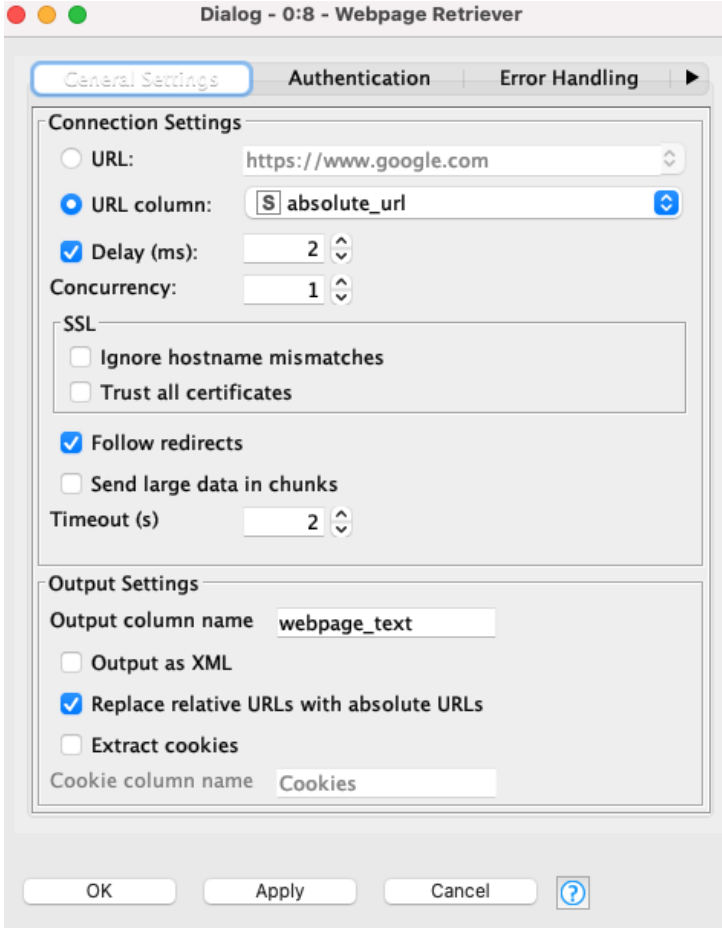
At the end of this section, you will be able to:

1. Recognize data access nodes.
2. Perform webpage retrieval.
3. Differentiate between widgets.
4. Build a data collection tool.

Webpage Retriever node

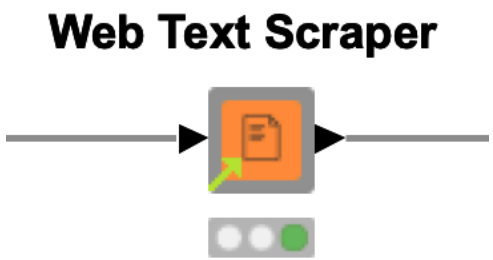
- Get *whole* webpage when no (useful) API exists



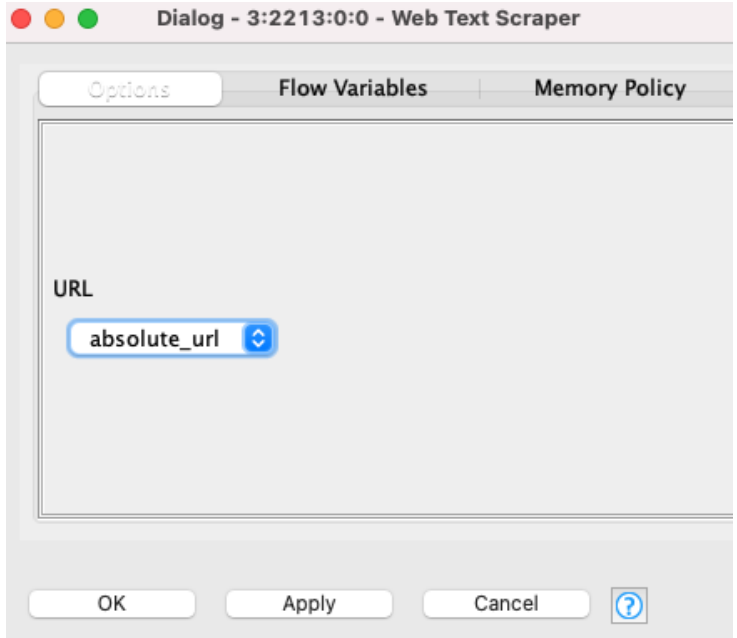


Web Text Scraper Verified Component

- Get *only* text from webpage

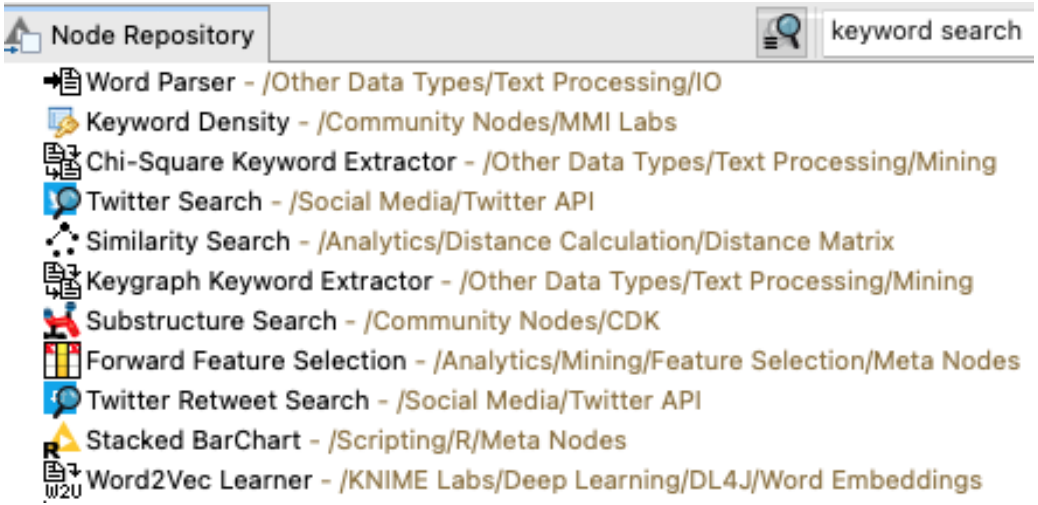


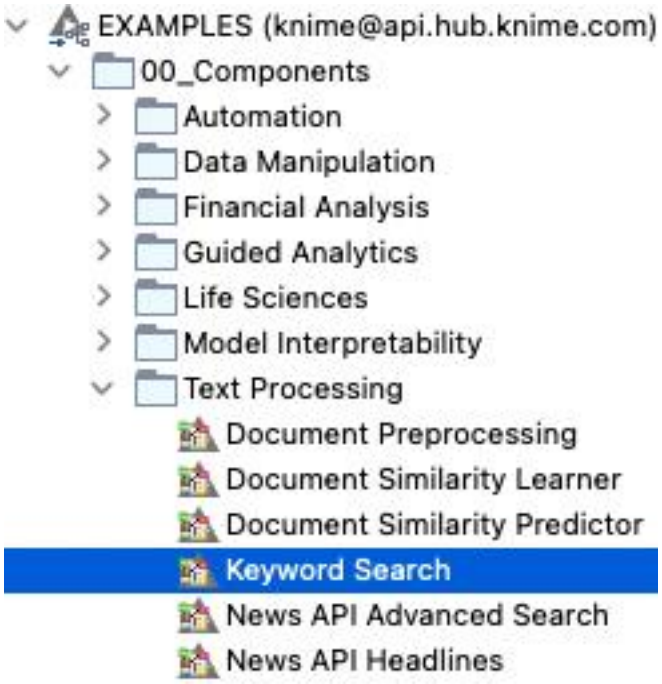
S	text
	Guided Labeling Blog Series – Episode 6: Comparing Active L...
	Create
	Guided Labeling Blog Series – Episode 6: Comparing Active L...
	July 27, 2020 — by Paolo Tamagnini & ...
	Recap: In the last episode we made an analogy with a numbe...
	Let's pick up where we left off.
	You can blend friends' movies opinions in a single model, but...
	Weak Supervision instead of Active Learning
	The key feature that differentiates active learning from weak ...
	Unique vs Flexible



Where can I find Verified Components?

- They are not in the Node Repository, because it's a Verified Component

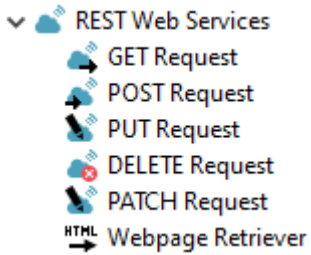


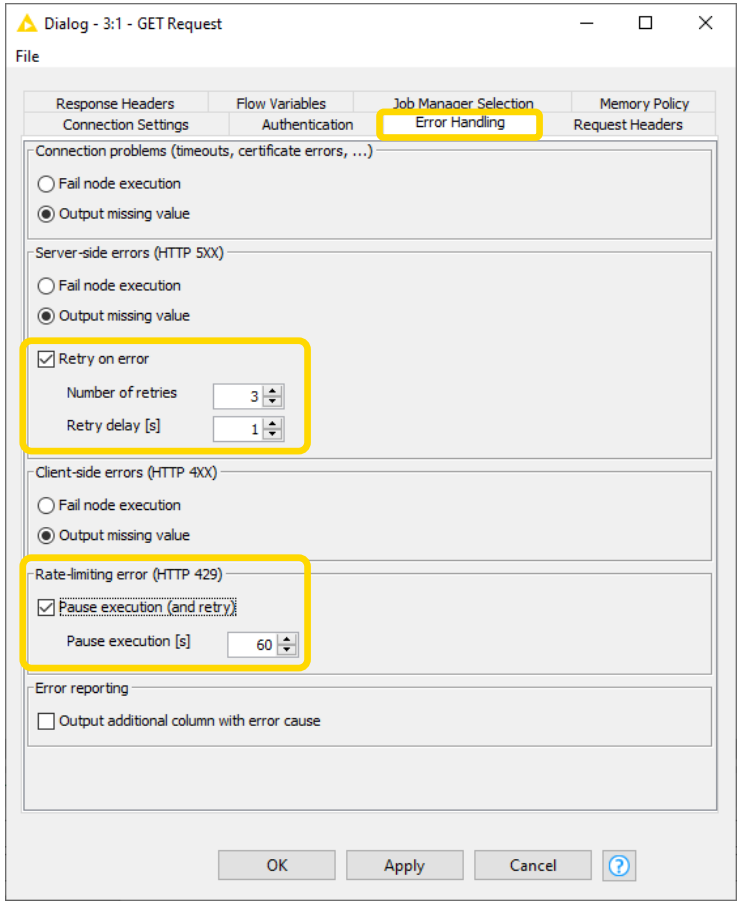


See more verified components here

Error handling on REST Web Service Nodes

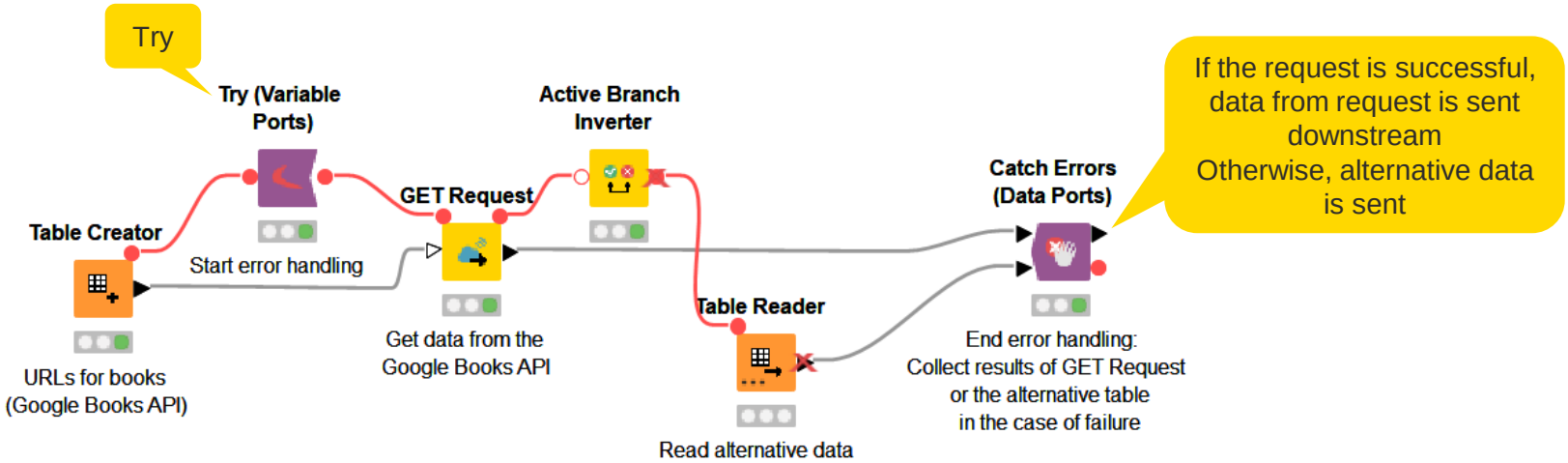
- Many errors might occur
  - **client**-side errors, **server**-side errors or **rate**-limiting conditions
- Individually configurable
- Works with all the nodes in KNIME REST Client Extension





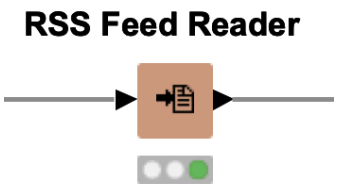
Custom error handling

- Try & Catch nodes handle errors
- **Prevent whole workflow failure** even when a single node fails in between



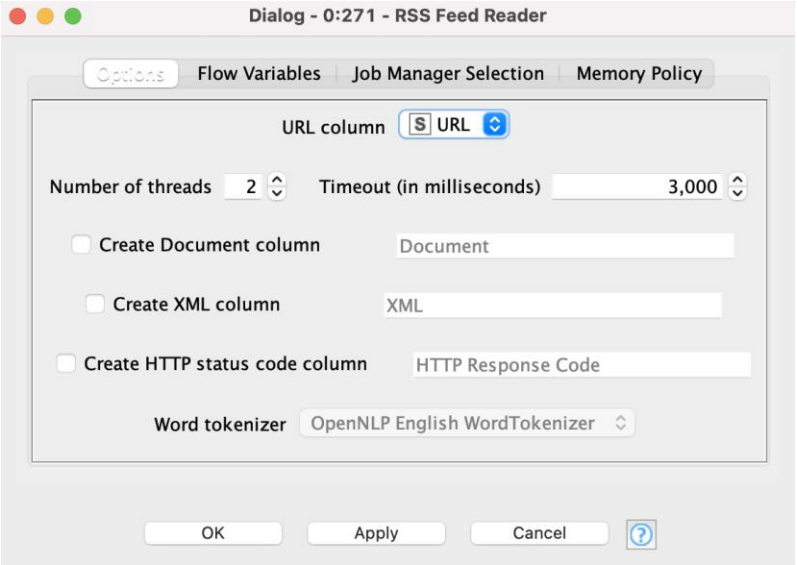
RSS Feed Reader: Another way to read web data

- Download up-to-date information from a particular website such as a news website



S Title	S Description
Worry and fear as US faces baby fo...	With stocks running low across the country,
Camille Vasquez: Johnny Depp's la...	The young lawyer has caught the internet's
Oklahoma passes bill banning most...	The state's ban, its third in recent months,

- What's RSS?
  - RSS (Really Simple Syndication) is a **content** distribution method that allows access to **updates** to websites in a standardized format.



02: From links to data demo

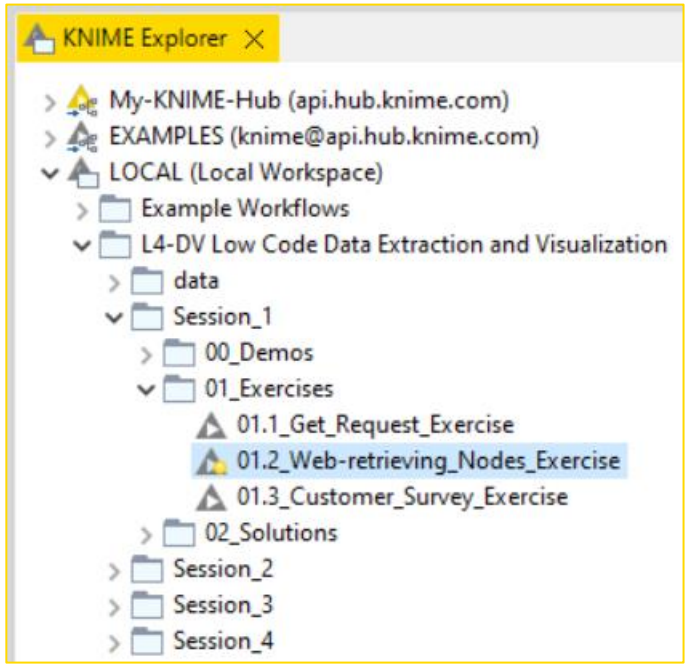
Live
Demo

Exercises: Section 1

1. Get Request Exercises
   Given a table with retrieve information with the Get Request node.

2. From Links to Data Exercise
   Given a list of URLs, acquire the text from each link.

3. Customer Survey Exercise
   Create your own customer feedback form.

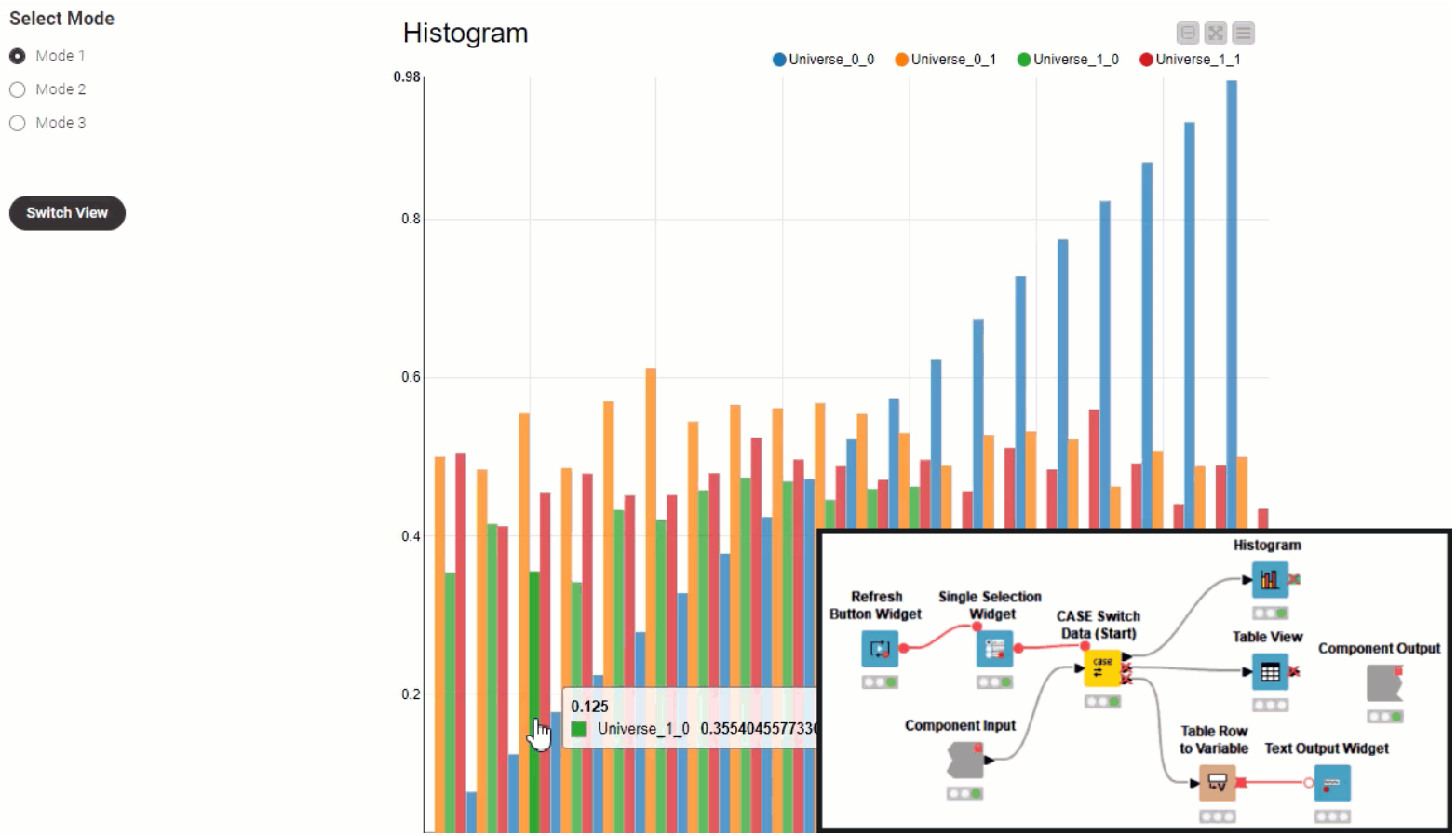


Data Collection

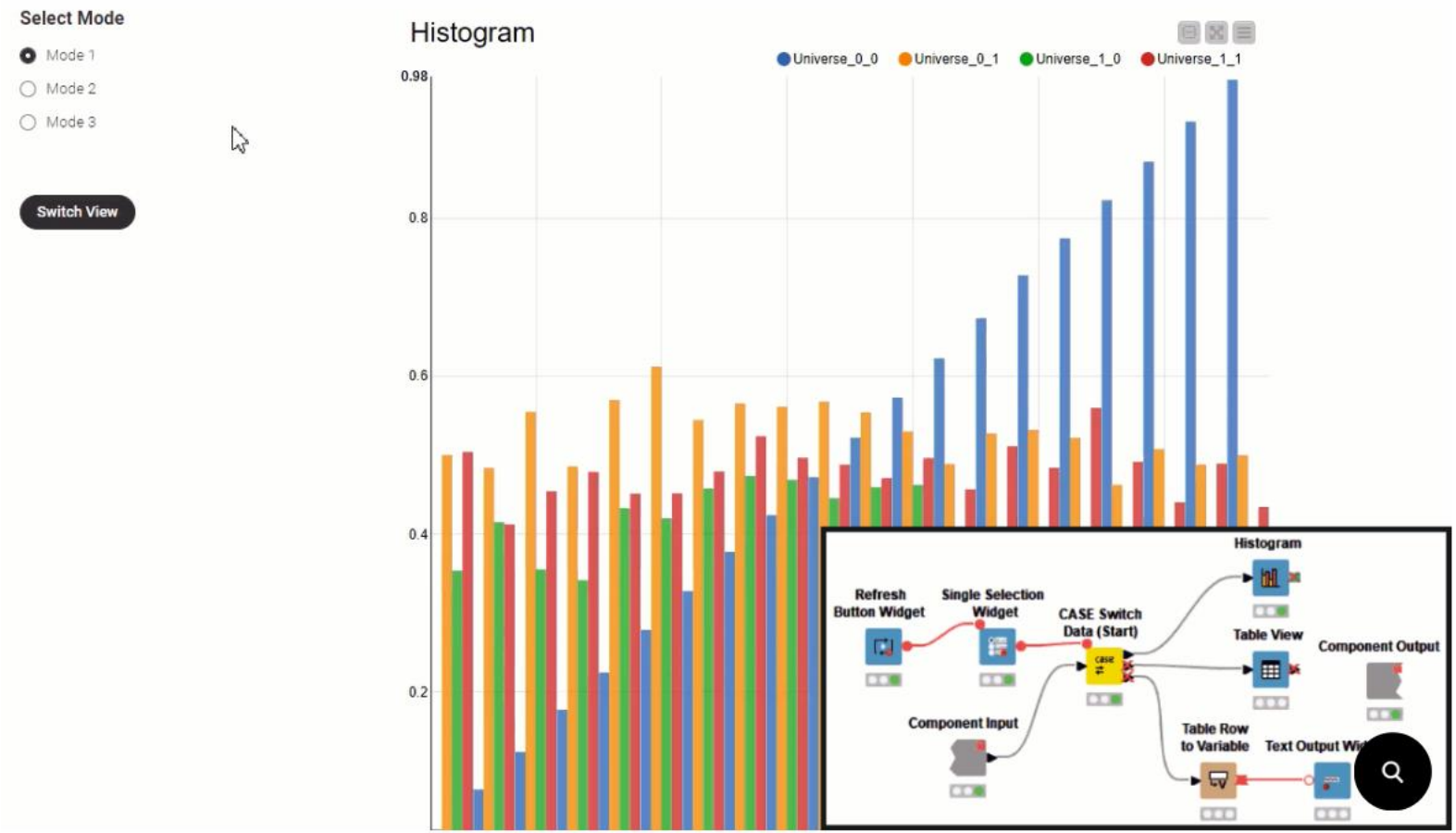
At the end of this section, you will be able to:

1. Recognize data access nodes.
2. Perform webpage retrieval.
3. Differentiate between widgets.
4. Build a data collection tool.

What’s a widget? An interaction enabler!

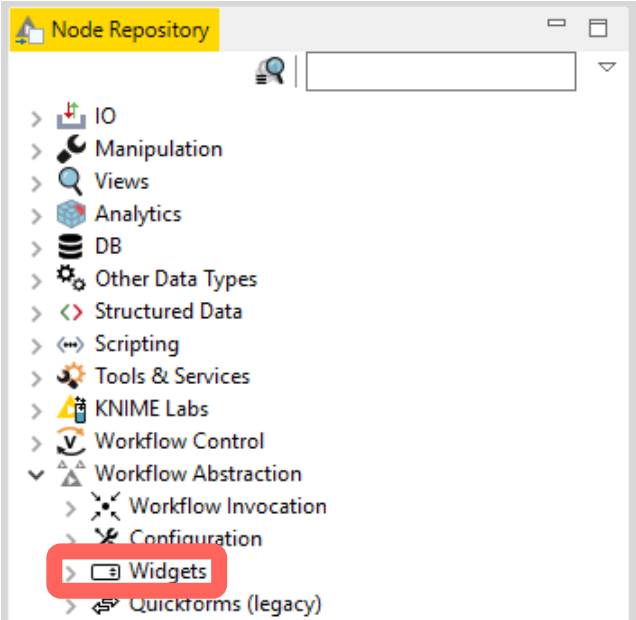


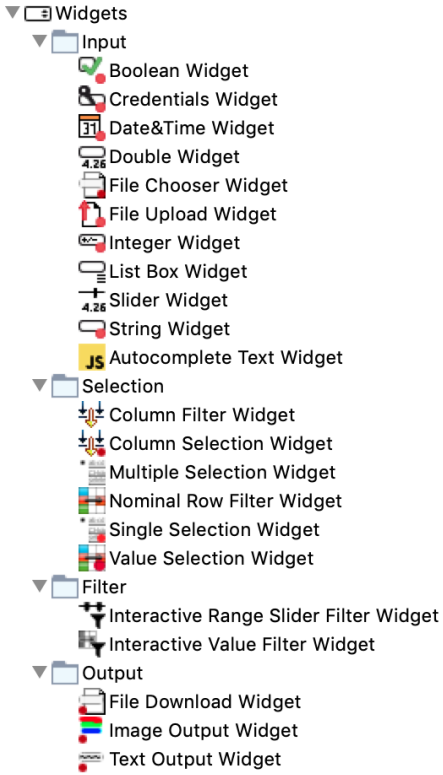
What’s a widget? An interaction enabler!



Widget nodes within a component

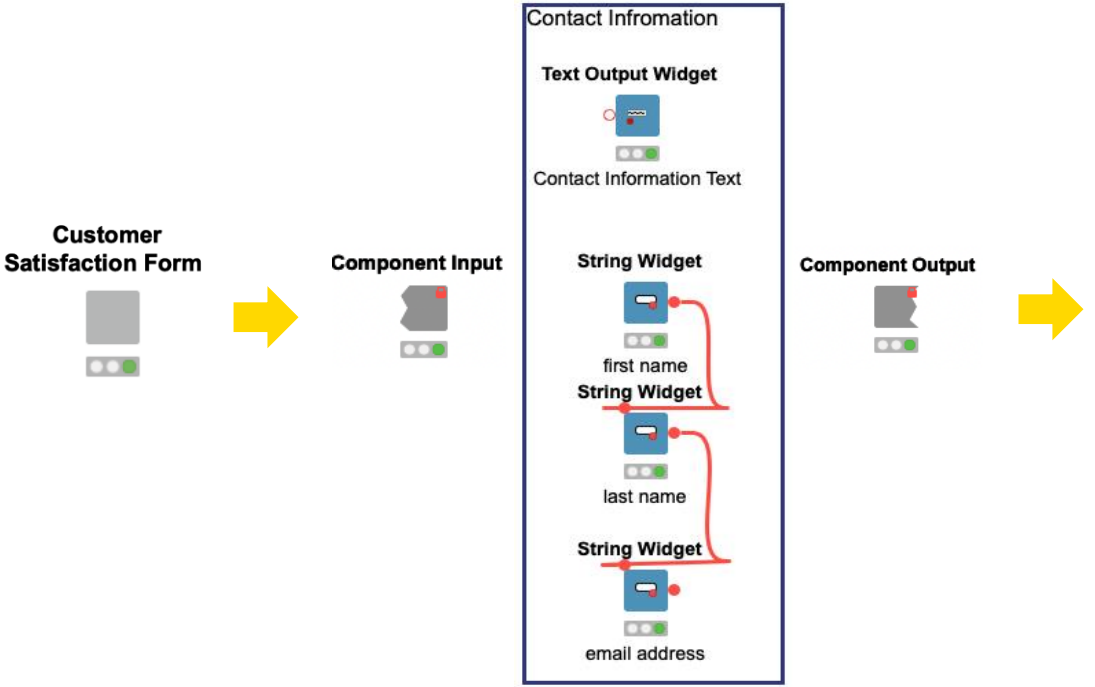
- Enable interaction by **selecting, filtering, and entering values** in the interactive view.





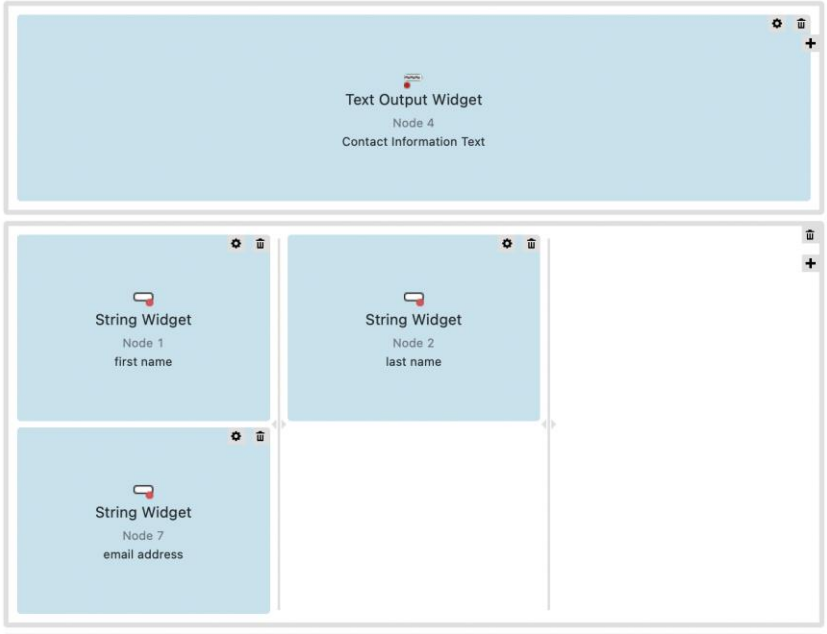
Widgets within a component create interactivity

- Add custom interactions in the component's interactive view.



Component Node Usage and Layout button

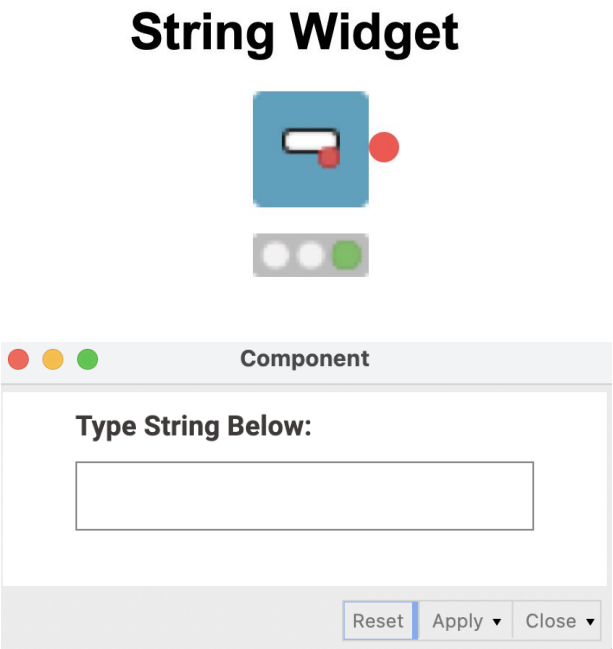
- Layout editor of the composite view:
  1. Node Usage
  2. **Visual Layout**
  3. Basic Layout
  4. Advanced Layout

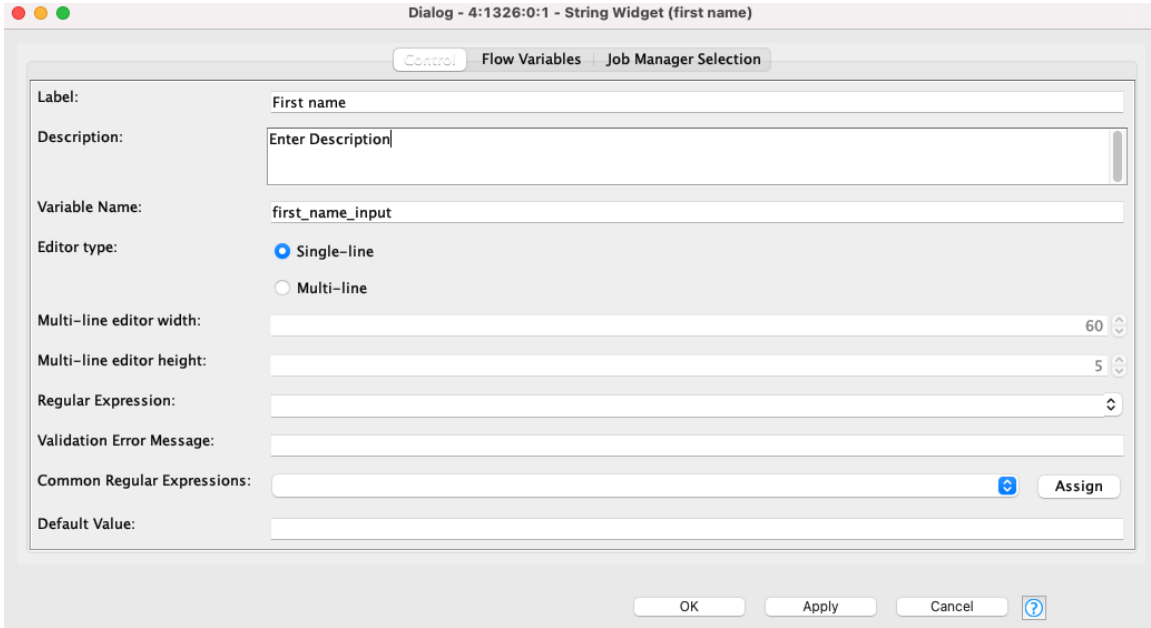


https://docs.knime.com/2022-12/analytics_platform_components_guide/index.html#introduction

Input widget node: String Widget

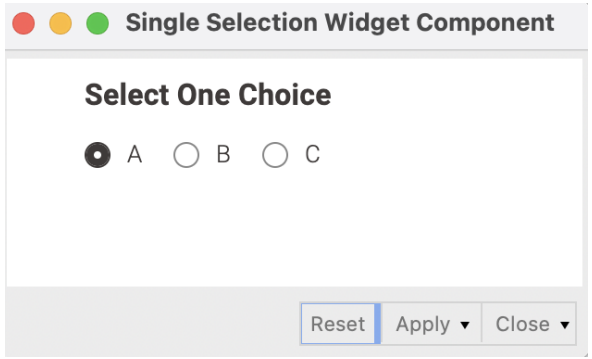
- Allow user to type.

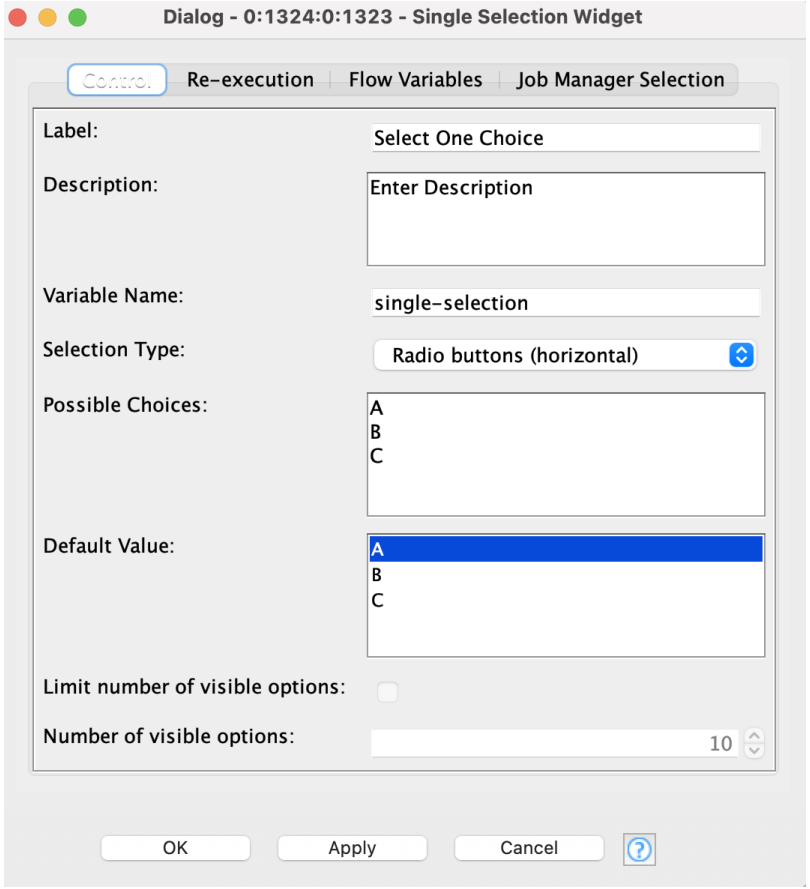




Selection widget node: Single Selection Widget

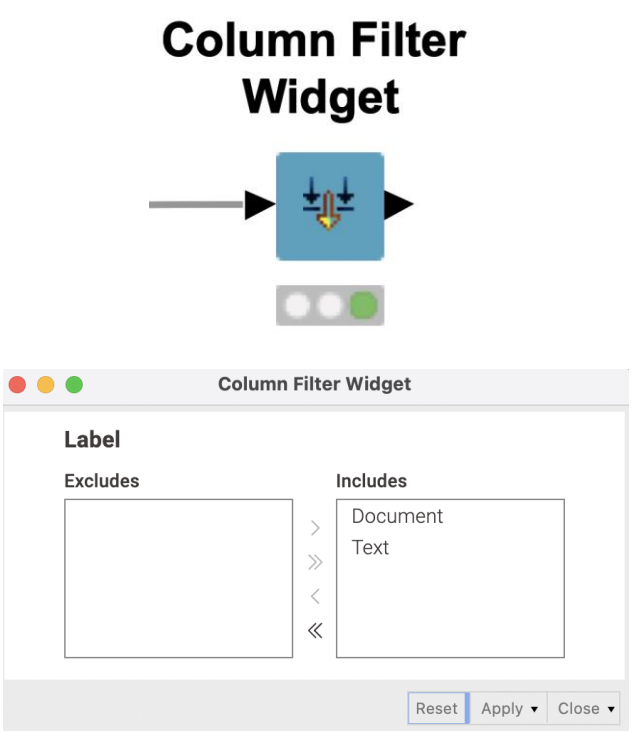
- Allow user to choose an option.

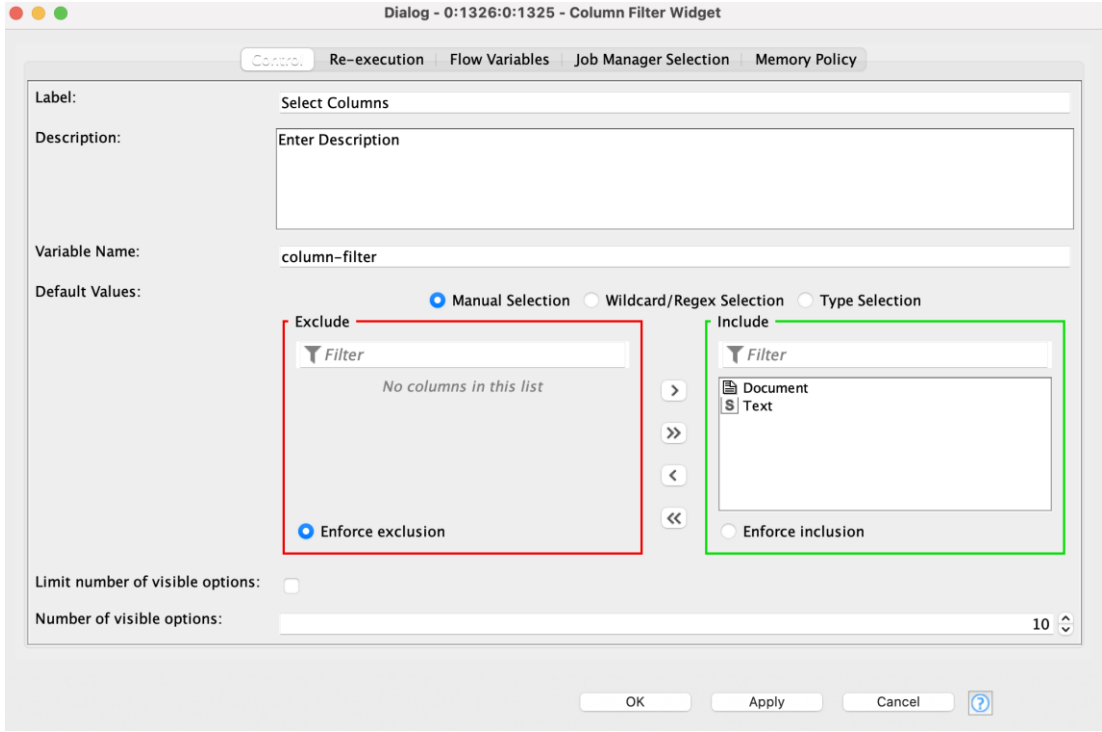




Filter widget node: Column Filter Widget

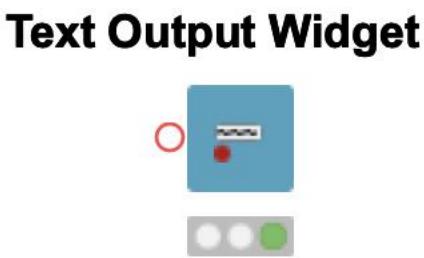
- Allow user to choose which columns to include and exclude.

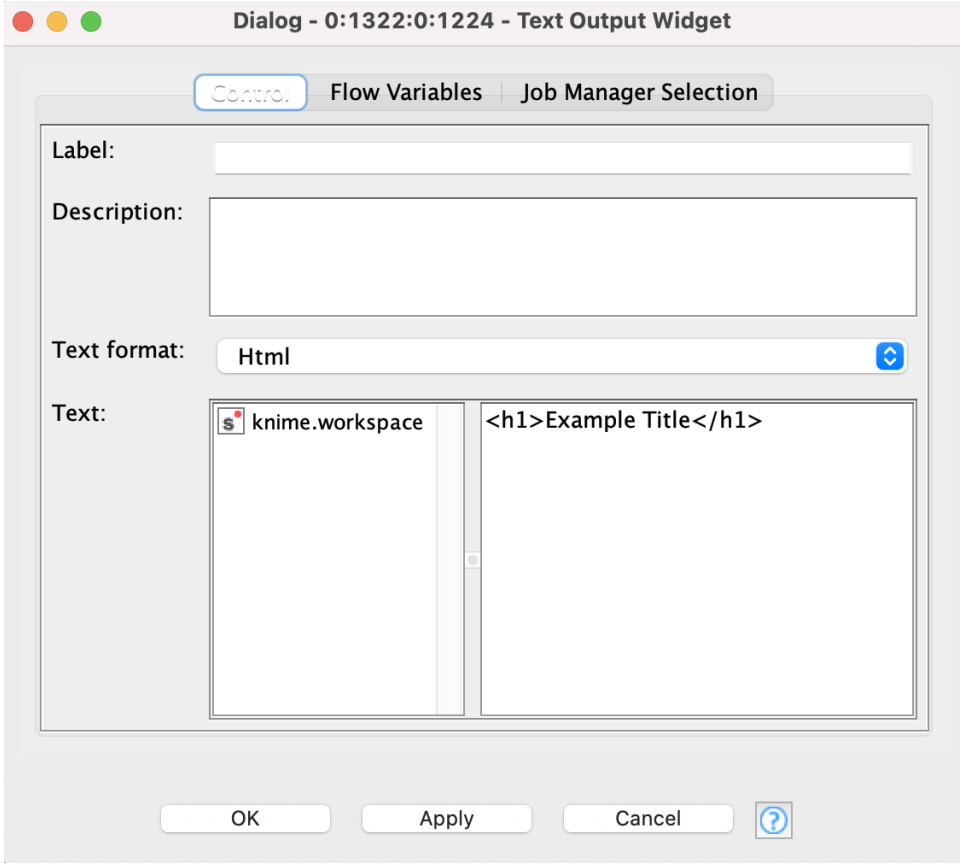




Output widget node: Text Output Widget

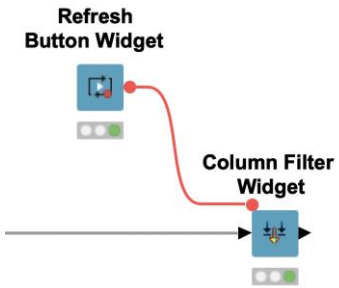
- Generates non-interactive text.

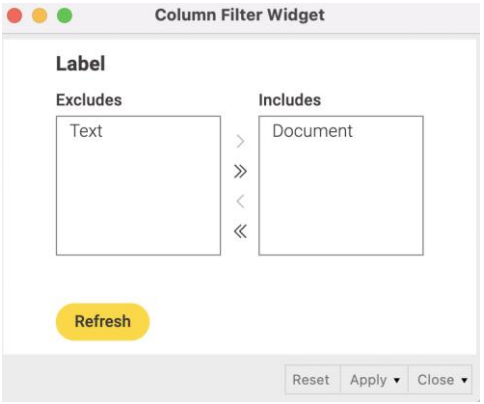


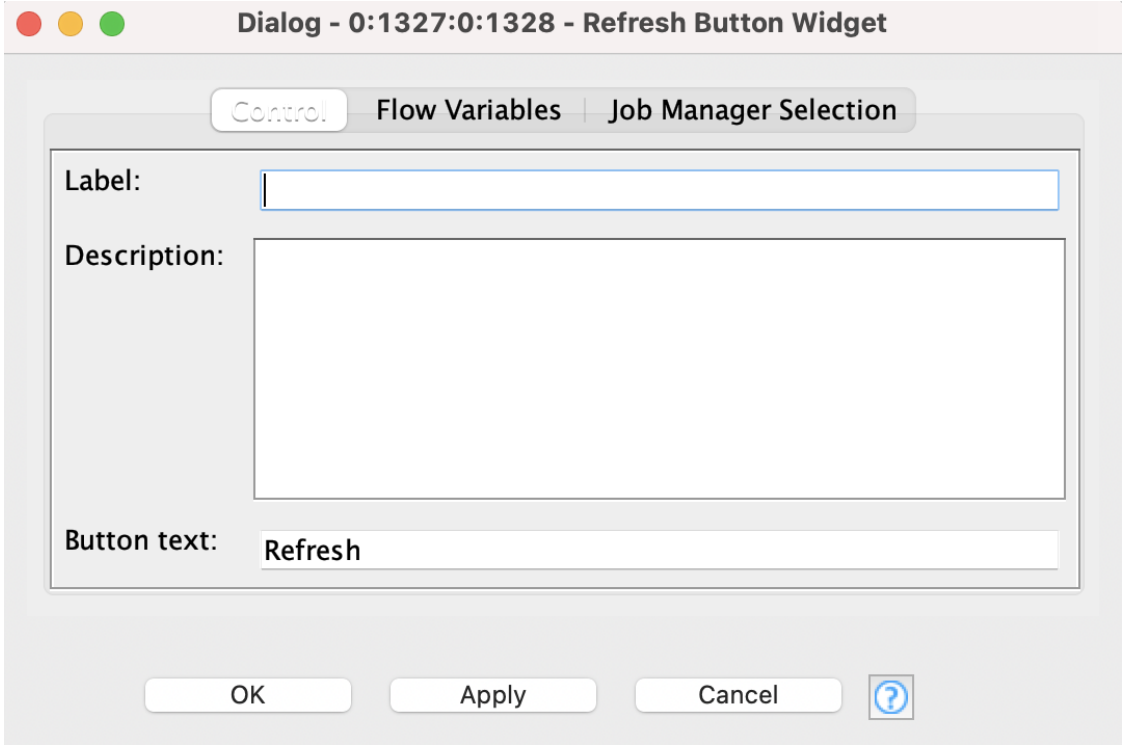


Re-execution widget node: Refresh Button Widget

- Re-execute all downstream nodes.







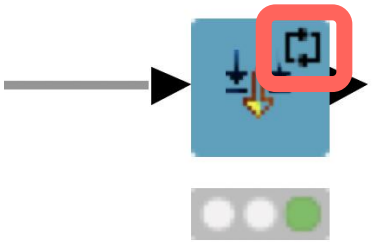
KNIME knowledge check 2

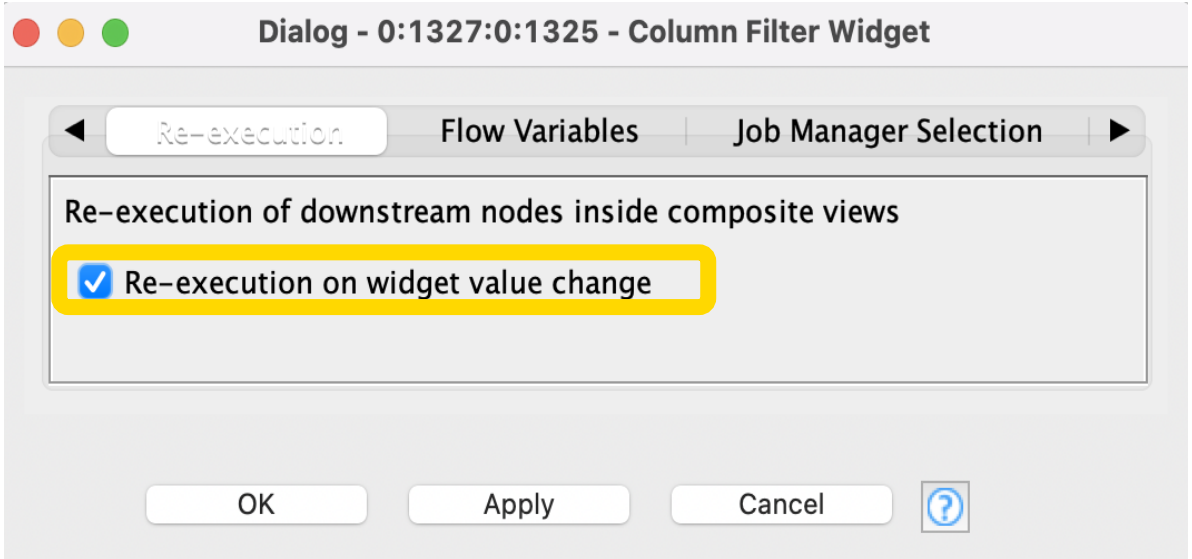
- Which widget allows a user to type text:
  - Text Output Widget
  - String Widget
  - Refresh Button Widget

Re-execution using dialog options

- **Automatically** re-execute all downstream nodes when a selection is made.

Column Filter Widget

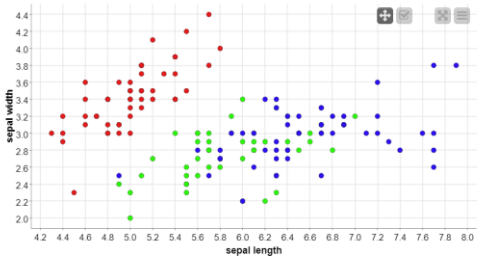


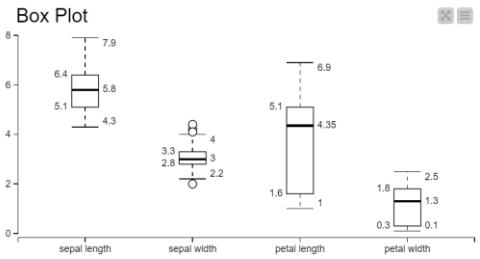


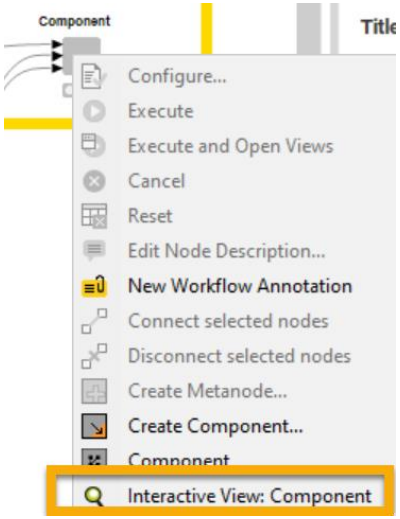
Note: Not all widgets have the Re-execution tab

View your creation

- To View:
  - Right click the component
  - Select Interactive View: Component







Show 10 entries

Search:

RowID	sepal length	sepal width	petal length	petal width	class
Row0	5.1	3.5	1.4	0.2	Iris-setosa
Row1	4.9	3	1.4	0.2	Iris-setosa
Row2	4.7	3.2	1.3	0.2	Iris-setosa
Row3	4.6	3.1	1.5	0.2	Iris-setosa
Row4	5	3.6	1.4	0.2	Iris-setosa
Row5	5.4	3.9	1.7	0.4	Iris-setosa
Row6	4.6	3.4	1.4	0.3	Iris-setosa
Row7	5	3.4	1.5	0.2	Iris-setosa
Row8	4.4	2.9	1.4	0.2	Iris-setosa
Row9	4.9	3.1	1.5	0.1	Iris-setosa

Showing 1 to 10 of 150 entries

Previous 1 2 3 4 5 ... 15 Next

Excludes

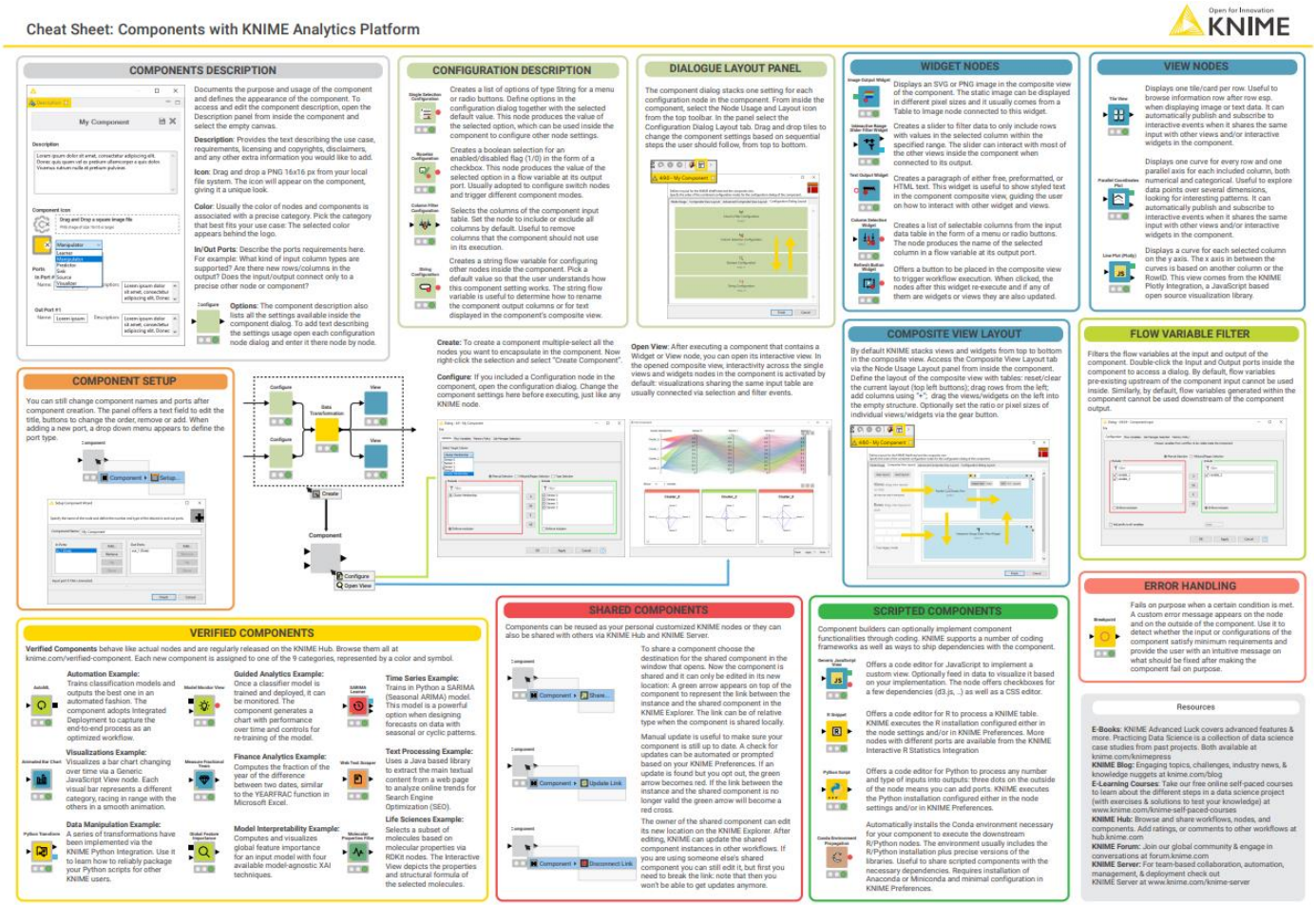
Includes

Iris-setosa
Iris-versicolor
Iris-virginica

4.30 7.90

Reset Apply Close

Cheat Sheet for components and widgets



Open for Innovation KNIME

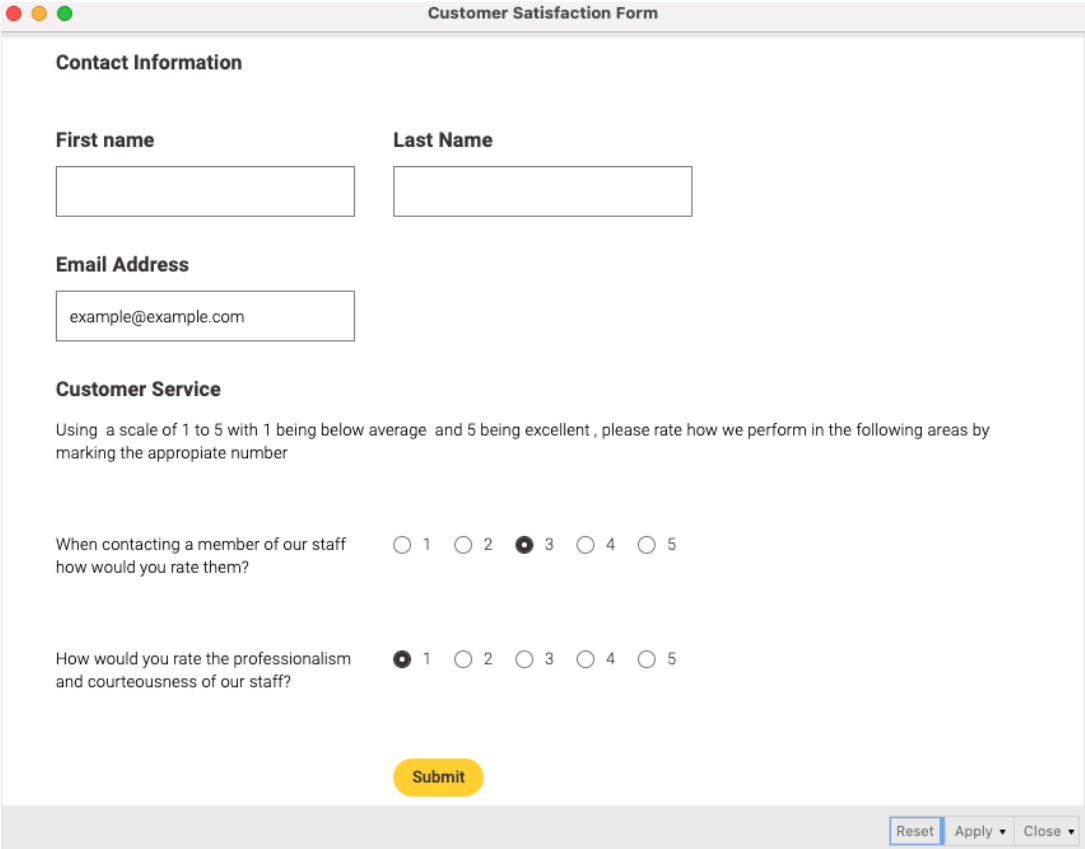
Data Collection

At the end of this section, you will be able to:

1. Recognize data access nodes.
2. Perform webpage retrieval.
3. Differentiate between widgets.
4. Build a data collection tool.

03: Customer survey demo

- **Goal:**
  - Create an interactive survey
- **Method:**
  - Add four unique widgets
  - Build component
  - Define layout

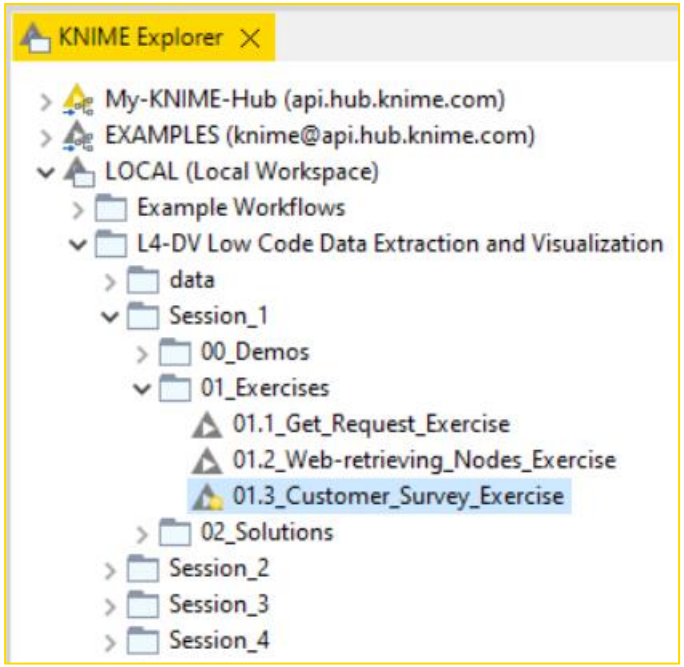


Exercises: Section 1

1. Get Request Exercises
   Given a table with retrieve information with the Get Request node.

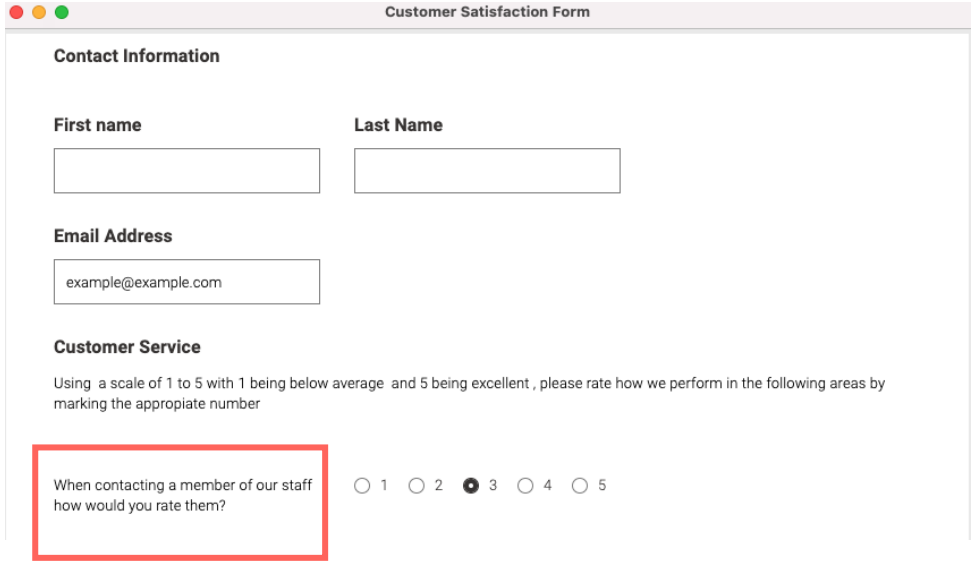
2. From Links to Data Exercise
   Given a list of URLs, acquire the text from each link.

3. Customer Survey Exercise
   Create your own customer feedback form.



KNIME knowledge check 3

- Which widget was used to create the question in the red box?
  - String Widget
  - Text Output Widget
  - Column Filter Widget
  - Single Selection Widget



Summary of section 1

Now you should be able to:

1. Recognize data access nodes.
2. Perform webpage retrieval.
3. Differentiate between widgets.
4. Build a data collection tool.

Open for Innovation
KNIME

[L4-DV] Low Code Data Extraction and Visualization

Section 2

Data Visualization

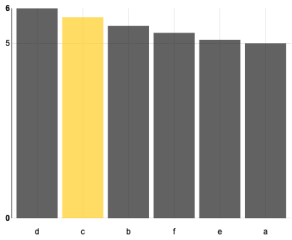
At the end of this section, you will be able to:

1. Match correct visualization to a specific task.

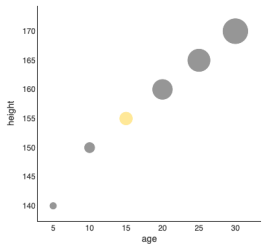
2. Apply visualizations to common tasks.

Match correct visualization for a task

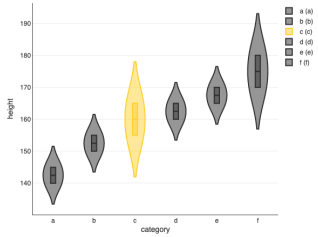
Comparison



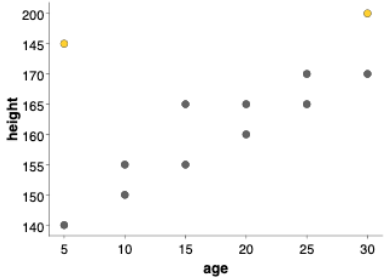
Correlation



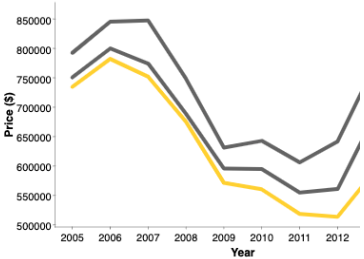
Distribution



Outliers

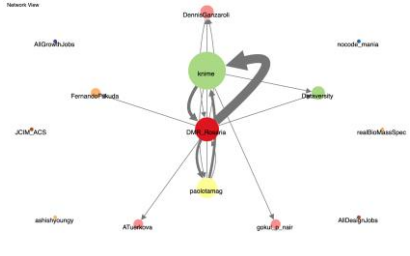


Time

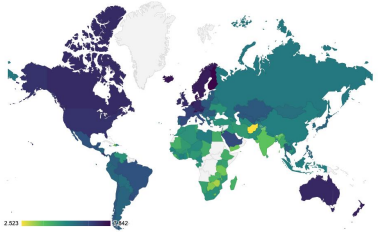


Text

Networks



Geography

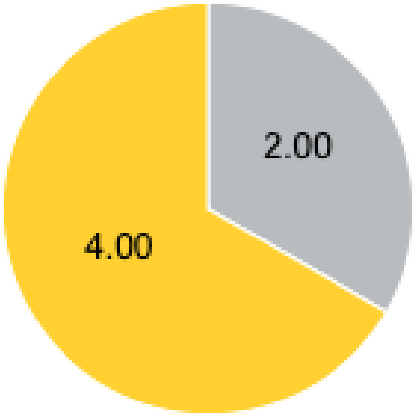


The visualizations discussed today

- **Comparison**
  - Bar chart, pie chart
- **Correlation**
  - Scatter plot, bubble chart
- **Distribution**
  - Histogram, violin plot
- **Outliers**
  - Box plot, scatter plot
- **Time**
  - Line plot
- **Text**
  - Tag cloud
- **Networks**
  - Network viewer
- **Geography**
  - Choropleth map

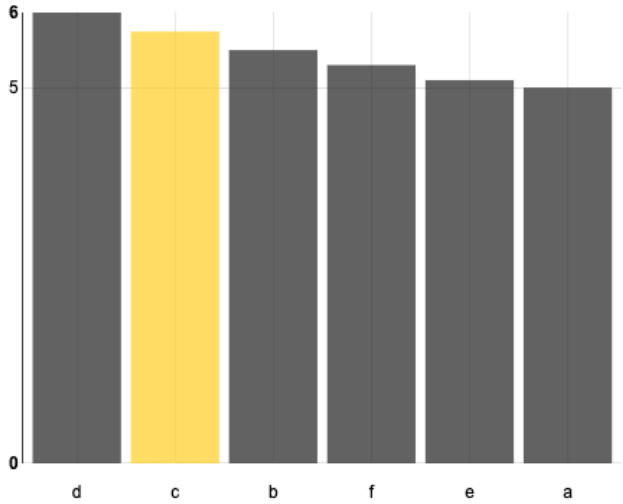
Comparison: Pie chart vs bar chart

- Pie chart
  - differences between **2 categories**



Example: compare revenue

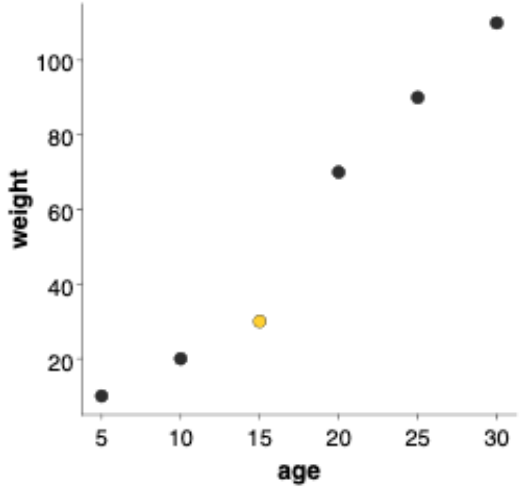
- Bar chart
  - Differences between **2 or more categories**



Example: compare growth

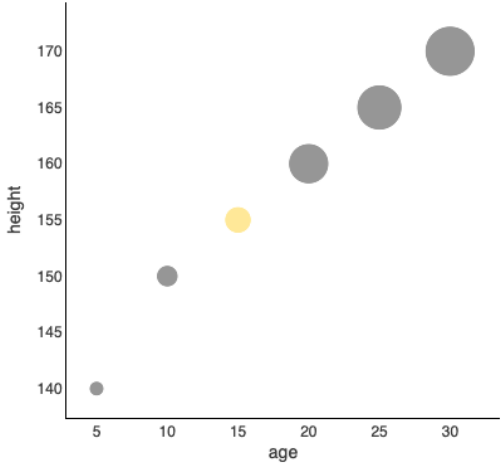
Correlation: Scatter plot vs bubble chart

- Scatter plot
  - correlation among **2 numeric columns**



Example: comparing age vs weight

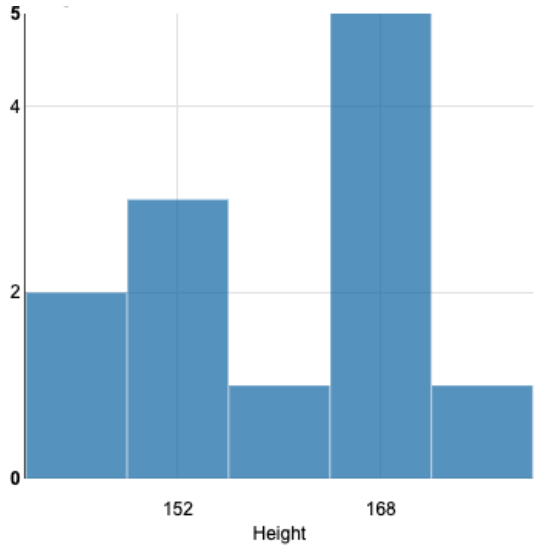
- Bubble chart
  - correlation among **3 numeric columns**



Example: comparing height vs age vs weight

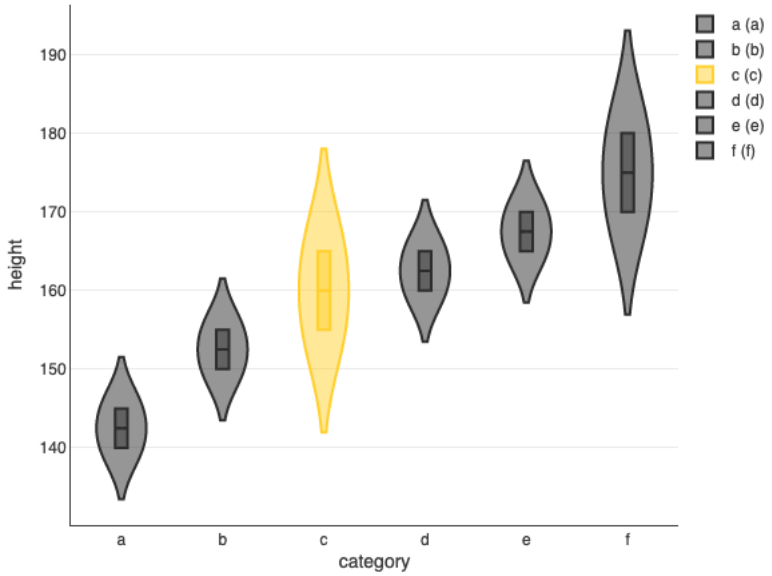
Distribution: Histogram vs violin plot

- Histogram
  - the distribution of a **single numeric column**



Example: distribution of height

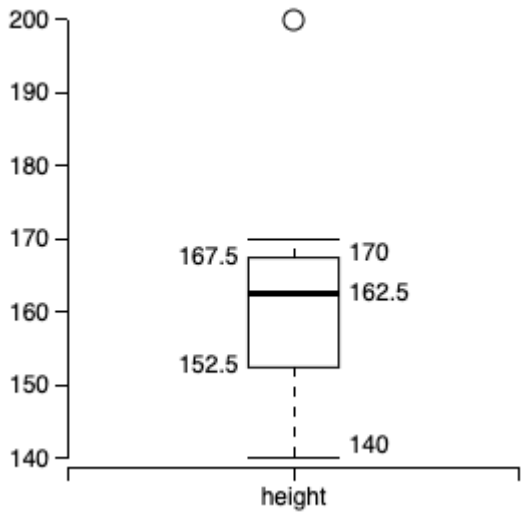
- Violin plot
  - the distribution of a **numeric column per category**



Example: distribution of height per category

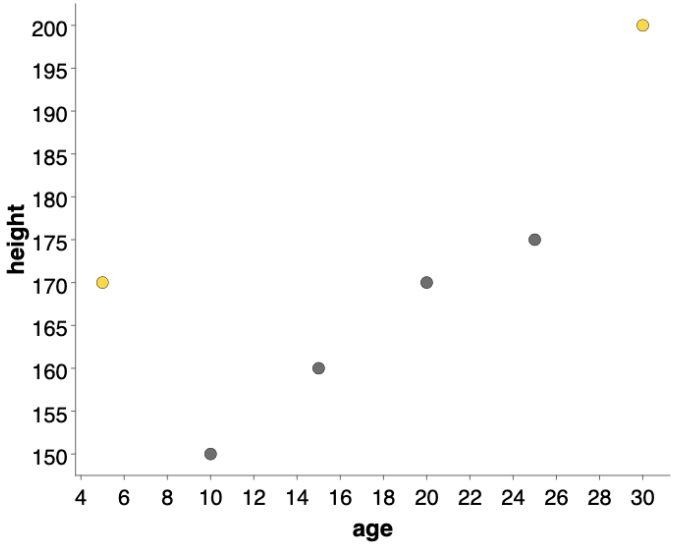
Outliers: Box plot vs scatter plot

- Box plot
  - spot outliers for **1 numeric column**



Example: spot unusually tall or short people

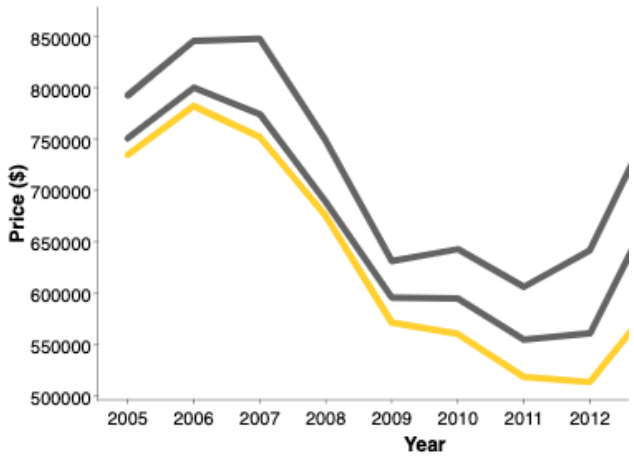
- Scatter plot
  - spot outliers among **2 numeric columns**



Example: spot unusually tall or short people for their age

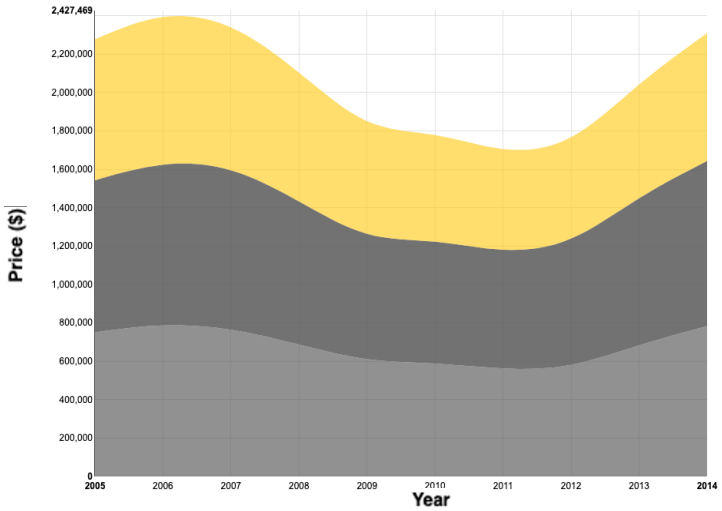
Time: Line plot vs stacked area chart

- Line plot
  - 1 or more **numeric columns** with a **time dependency**



Example: prices of American houses in the early 2000s

- Stacked area chart
  - 2 or more **aggregated numeric columns** with a **time dependency**



Example: prices of American houses in the early 2000s

KNIME Knowledge Check 01

- For each visualization node below, state the type of data visualized such as categorical or numerical data and the number of columns required.
  - For instance,
    - the Bar Chart node would be categorical with one or two columns required.
- Nodes to label: Line Plot, Box Plot, Histogram, Bubble Chart, Violin Plot

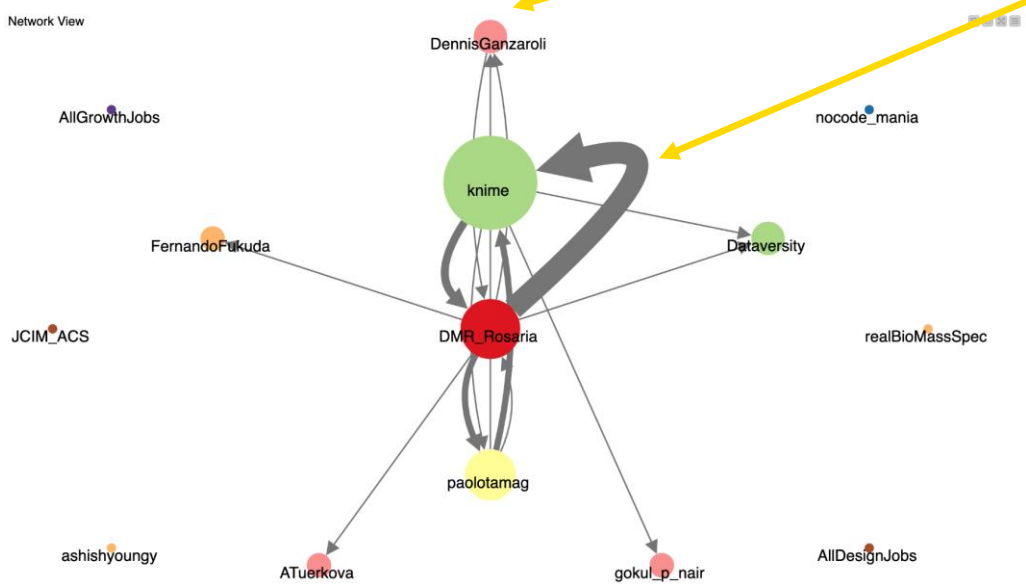
Text: Tag cloud

- Tag cloud
  - analyze word usage and frequency in data

- Key features:
  - Size of word corresponds to frequency
- Use cases:
  - Tweet analysis, product review analysis, customer topic analysis

Networks: Network viewer

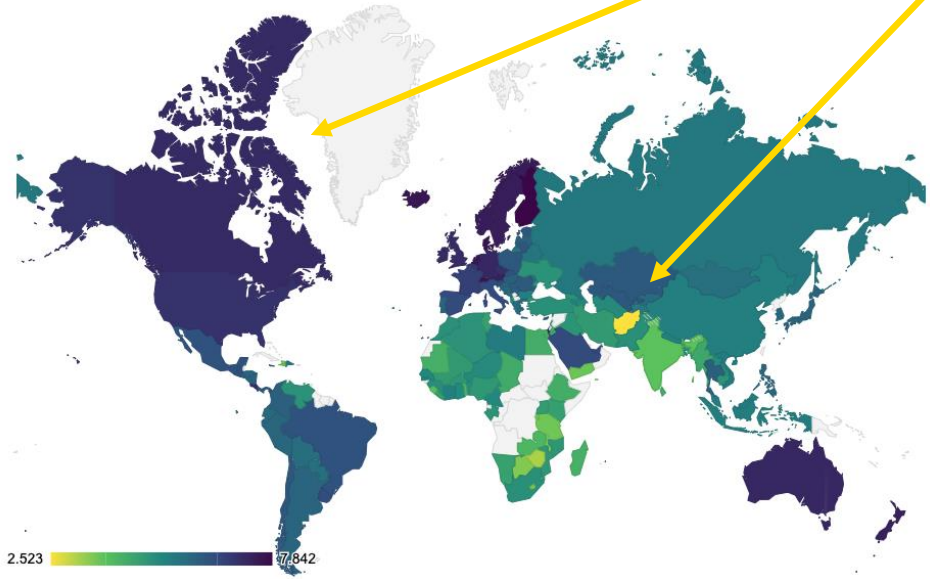
- Networks
  - analyze relationship strength among entities



- Key features:
  - Bubble size represents one column
  - Arrow size indicates strength of relationship
  - Pointing arrows indicate a one-way relationship
- Use case:
  - Locating trend setters on Twitter, marketing groups for TikTok

Geography: Choropleth map

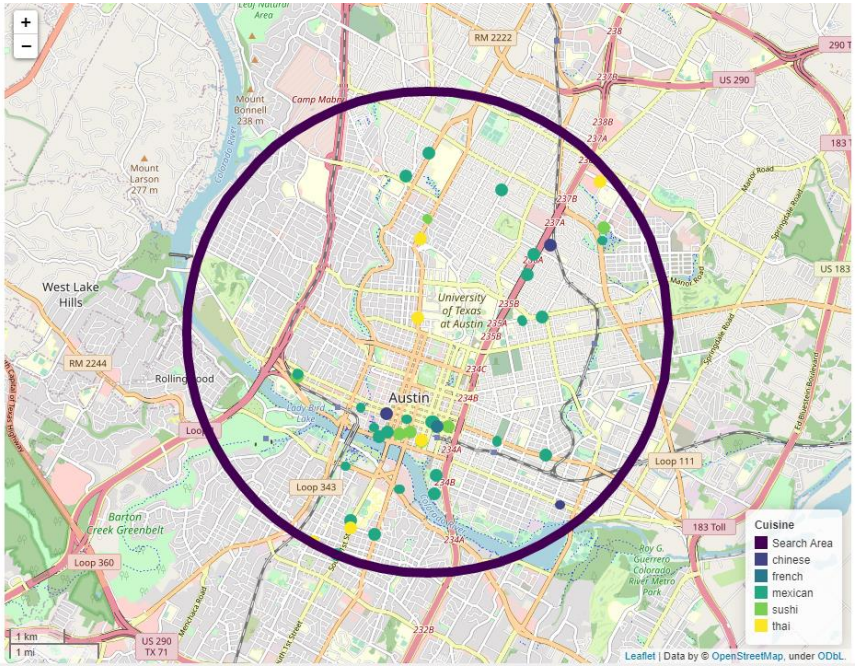
- Choropleth map
  - analyze numeric metrics by regions



- Key features:
  - Gradient colors indicate the progression of the numeric values
- Use cases:
  - Regional sales report, customer satisfaction, hot-spot tracking

Geography: Geospatial view

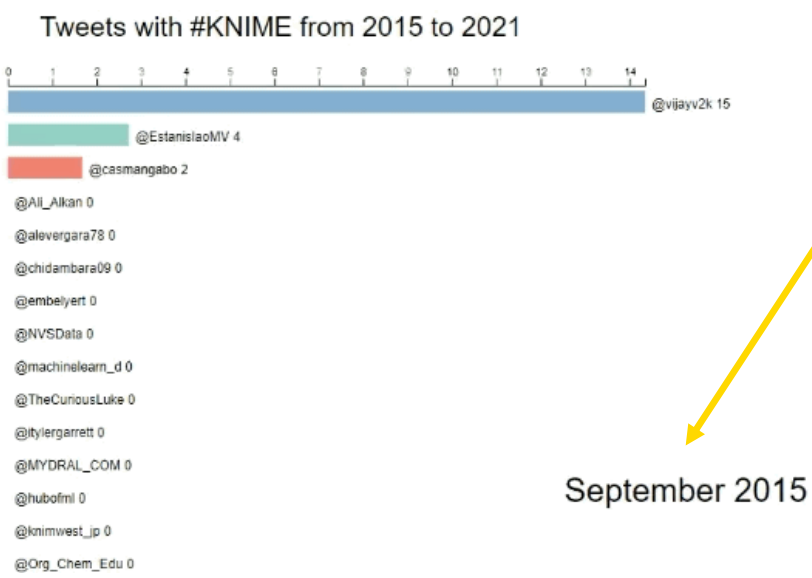
- Analyze data with location information (e.g., latitude and longitude) in an interactive map



- Key features
  - A part of the [Geospatial Analytics Extension for KNIME](#) by the Center for Geographic Analysis from Harvard and KNIME
- Use cases
  - Impact analysis
  - Points of interest search
  - Location optimization

Bonus: Animated bar chart

- Animated bar chart
  - visualize different entities competing or changing over time

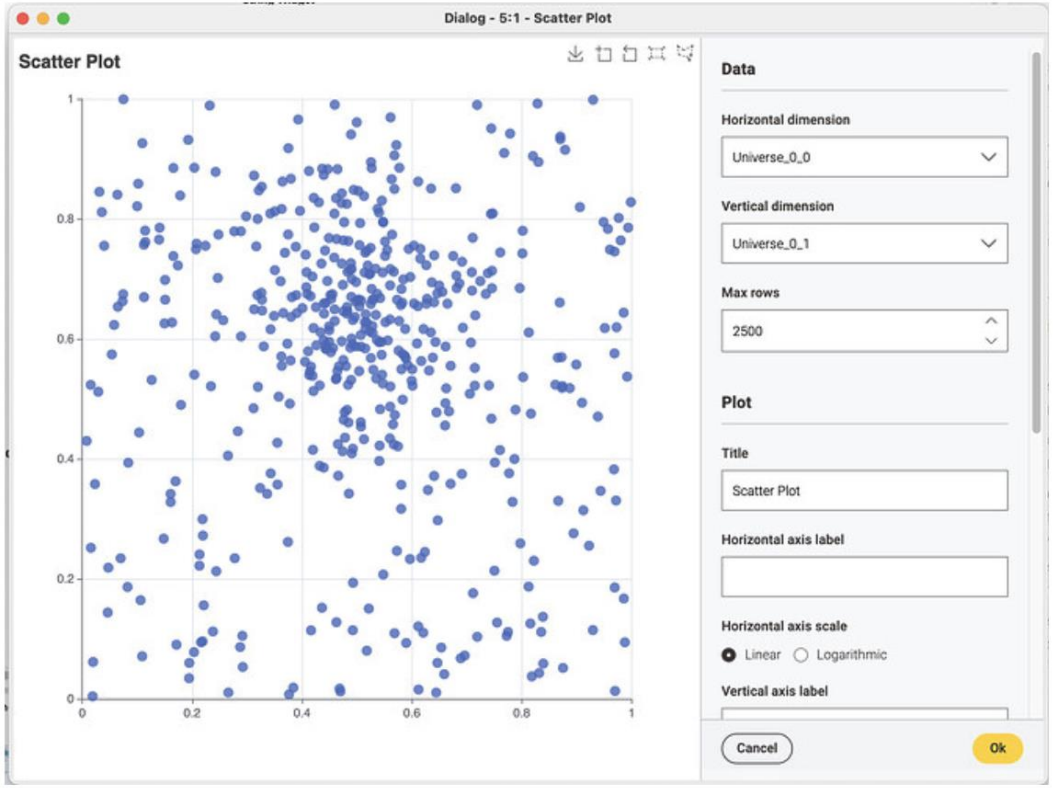


September 2015

- Key features:
  - Time change
- Use case:
  - Attention-grabbing, trend development analysis

What’s next for KNIME visualizations?

- To preview the next generation of KNIME visualizations, check out the labs extension called “KNIME Views (Labs)”



KNIME Knowledge Check 02

- Which visualization is best for examining relationships among entities?
  - Tag Cloud
  - Choropleth map
  - Histogram
  - Network

Data Visualization

At the end of this section, you will be able to:

1. Match correct visualization for a task.

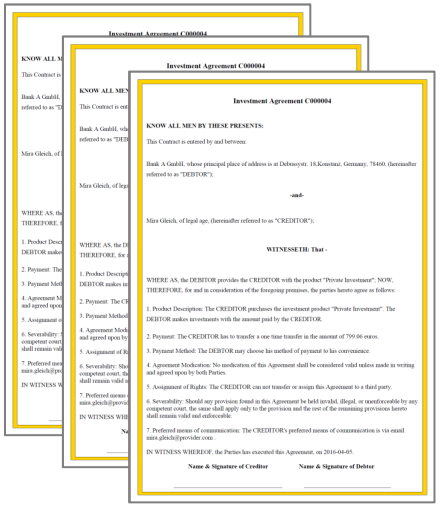
2. Apply visualizations to common tasks.

Common Tasks

1. Visualizing text topics.
2. Visualizing relationship strength.
3. Add multiple lines on a line plot.
4. Color bars on a 2-D bar chart.
5. Adding filters on plots.
6. Accessing and visualizing geospatial data.

Visualizing text topics

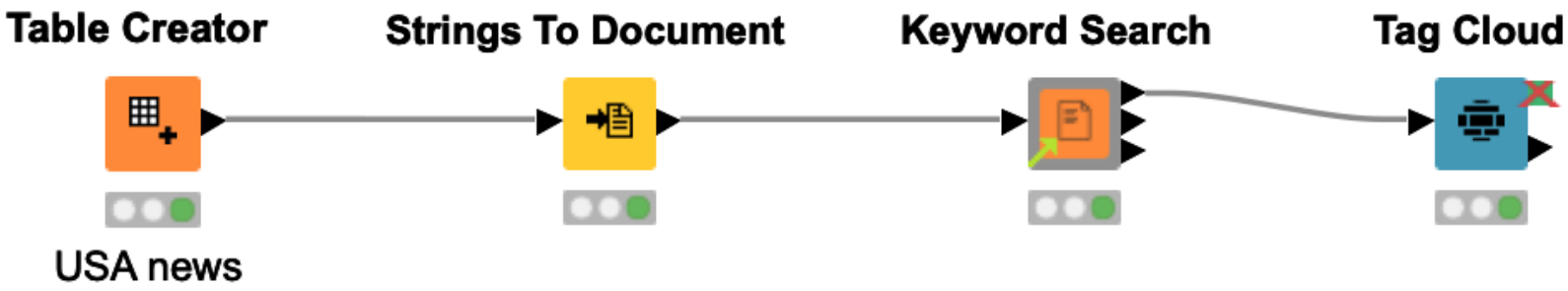
- A 3-stage process: acquire text, rank text, and visualize



Word	Frequency
KNIME	100
Data	75

Visualizing text topics in KNIME

- Making tag clouds is a simple process with 4 nodes.



Strings to Document node

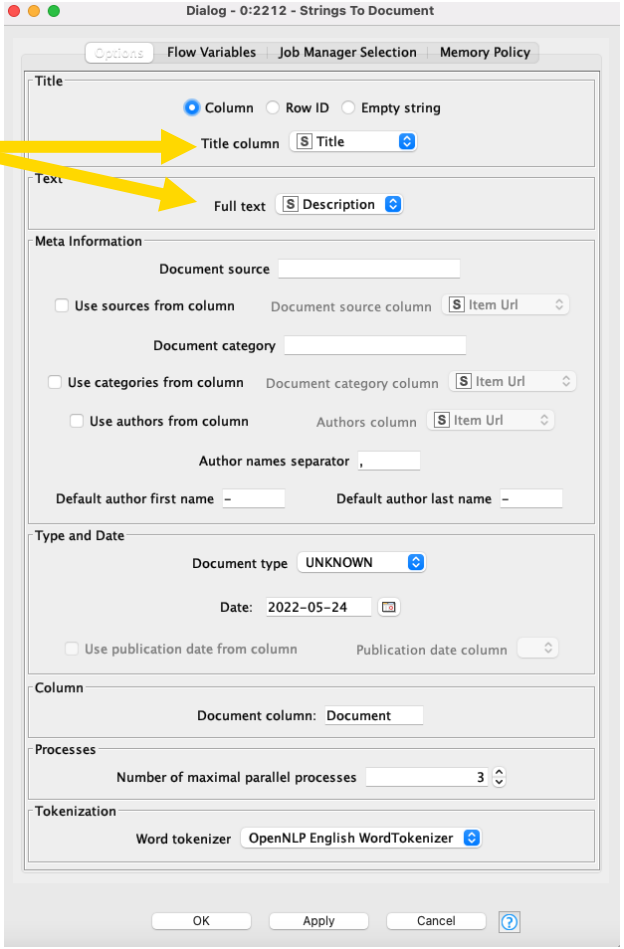
- Converts “String” to “Document” type for text processing tasks.

Strings To Document

Document
"Great food , interesting service"
"Excellent Lunch Destination"
"Hidden treasure near KaDaWe"
"Excellent Food Very Reasonable !"
"Good food , great prices !"
"Nice food at a reasonable price"

Must set these

Warning: The Keyword Search will use words from both the Title column and Full Text.



What’s a Document type?

- A composite/aggregate data type for textual content
- Fields include:
  - Title
  - Text
  - Source
  - Category
  - Author(s)
  - Date, …
  - Generic Meta Data

Document
"Great food , interesting service"
"Excellent Lunch Destination"
"Hidden treasure near KaDaWe"
"Excellent Food Very Reasonable !"
"Good food , great prices !"
"Nice food at a reasonable price"

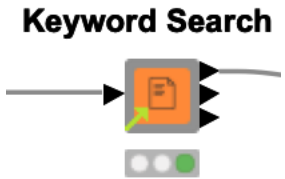
Warning: The text seen in a KNIME table for a document type only displays the title.

Keyword Search component

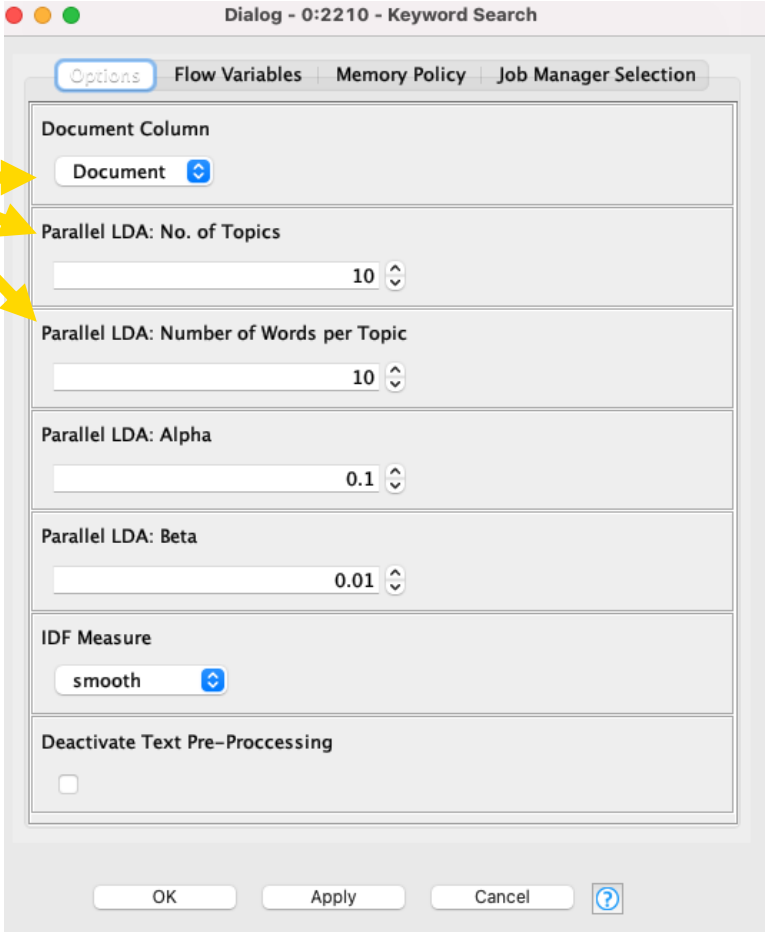
- Extracts keywords from a document for tag clouds, networks, etc.

Must set these

Warning: The number of terms found will be: Topics x Words



S Term	D Weight
women	2
winter	1
winners	1
weather	1
visit	2



Keyword Search outputs 1 and 2

- Outputs useful for tag clouds and networks

Output 1:
Word and
importance
(for tag clouds)

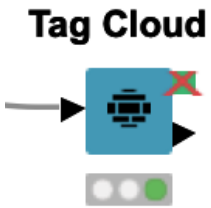
S Term	D Weight
women	2
winter	1
winners	1
weather	1
visit	2

Output 2:
Word pair
frequency
(for networks)

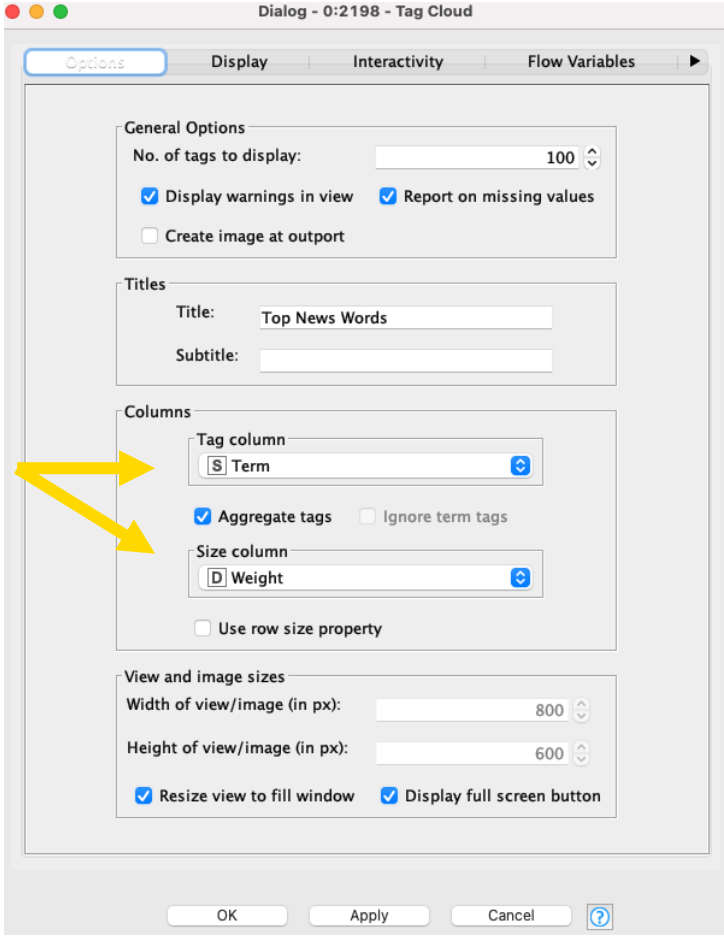
S UniqueID	S Term1	S Term2	I Document cooccurrence
york-university	university	york	1
york-bestowed	bestowed	york	1
york-renowned	renowned	york	1
york-pop	pop	york	1
york-singer	singer	york	1

Tag Cloud node

- Generates a tag cloud showing words and their frequencies/importance

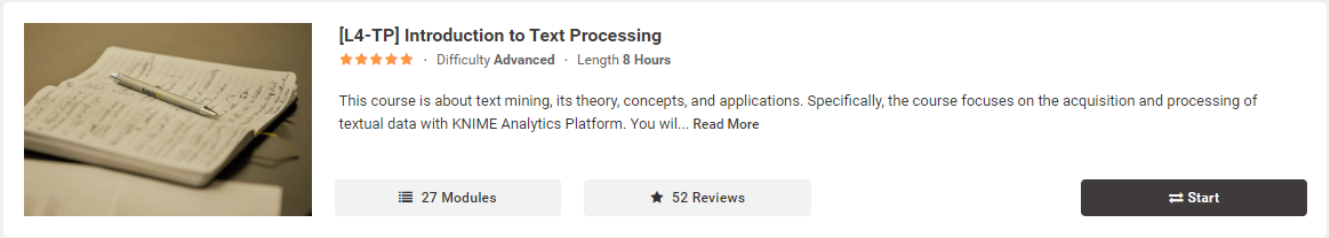


Must set these



A note on text / natural language processing

- The Keyword Search component performs document preprocessing automatically.
- The document preprocessing techniques are introduced in our free, self-paced text processing course on LearnUpon:



[Click here for KNIME Text Processing Educational Material]

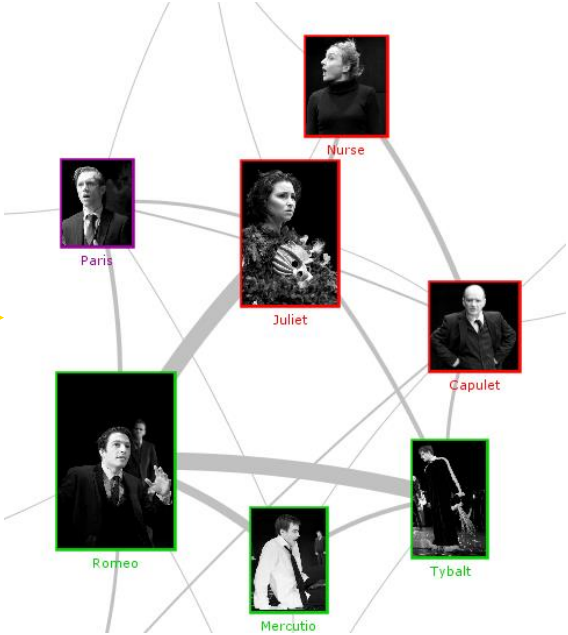
Visualizing relationship strength

- A 3-stage process: acquire text, build a network, and visualize the graph

Romeo, art thou mad?

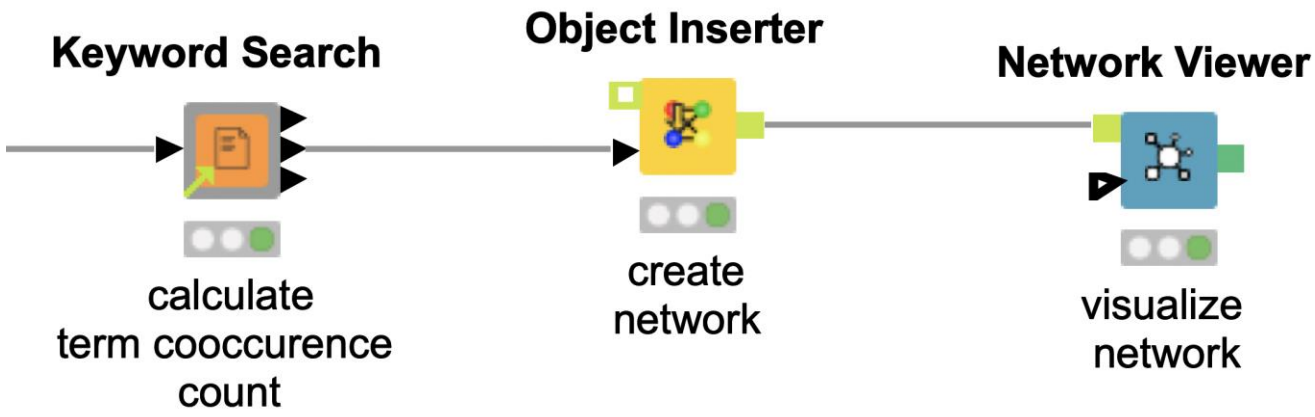
Network id: objectInserter **uri:** objectInserter
Directed: false **weighted:** true

No of nodes: 14 **No of edges:** 30



Visualizing relationship strength in KNIME

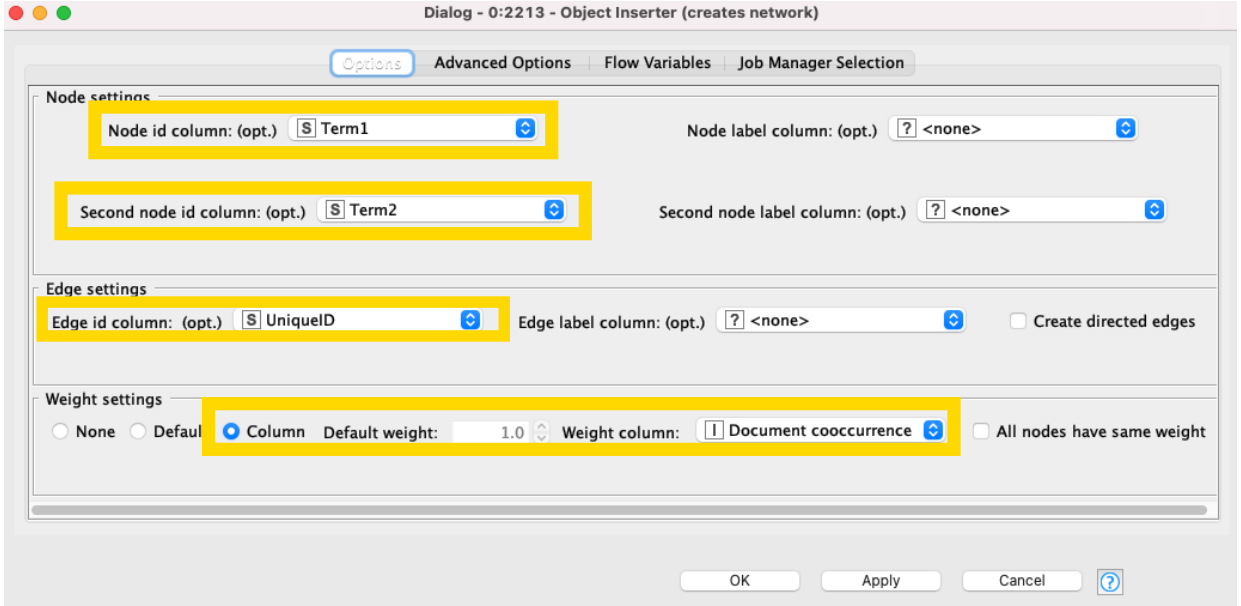
- Example workflow for making a network graph (with 2 *more* nodes)

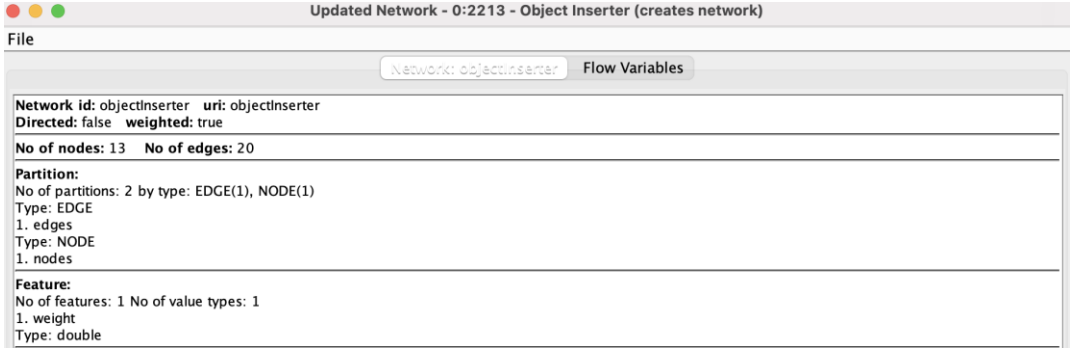


Object Inserter node

- Generates a network from a table.
- Provides input for the Network Viewer node.

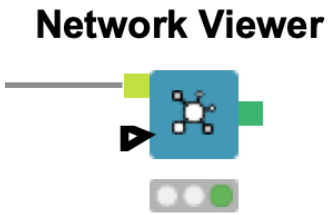
Object Inserter

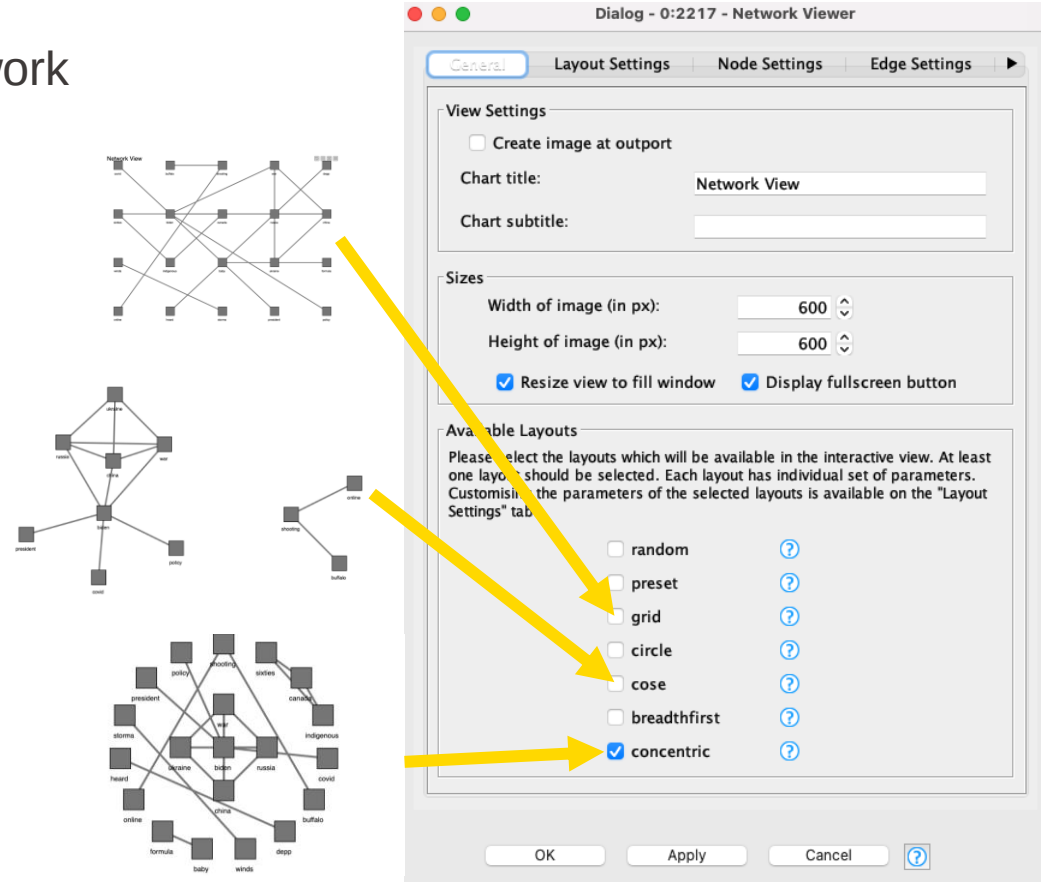




Network Viewer node

- Allows us to view a network



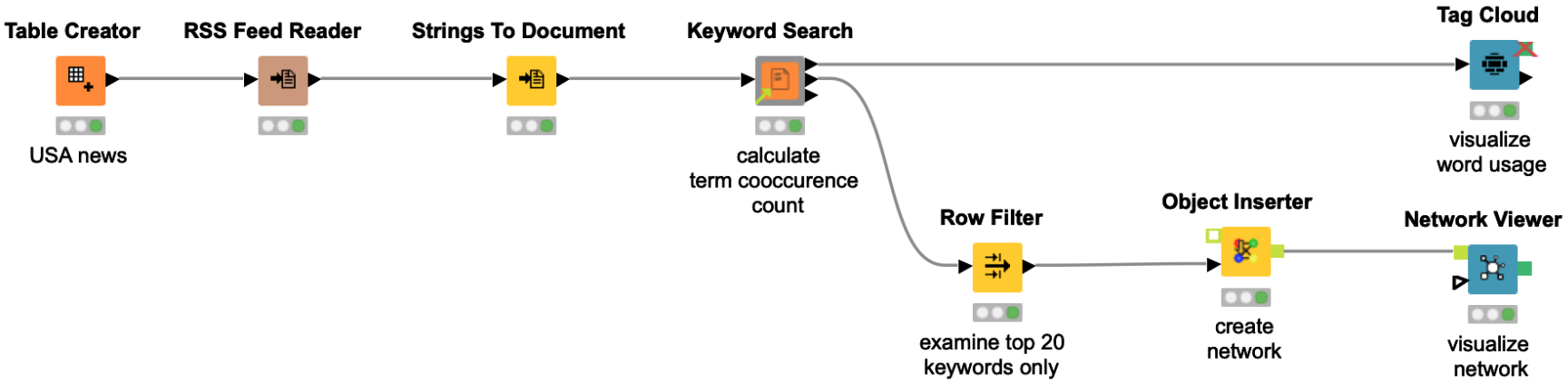


[Click here to learn more about network analysis]

KNIME Knowledge Check 03

- Word-Pair Frequency is useful for which visualization?
  - Tag Clouds
  - Networks
  - Choropleth maps

01: Simple tag cloud and network demo

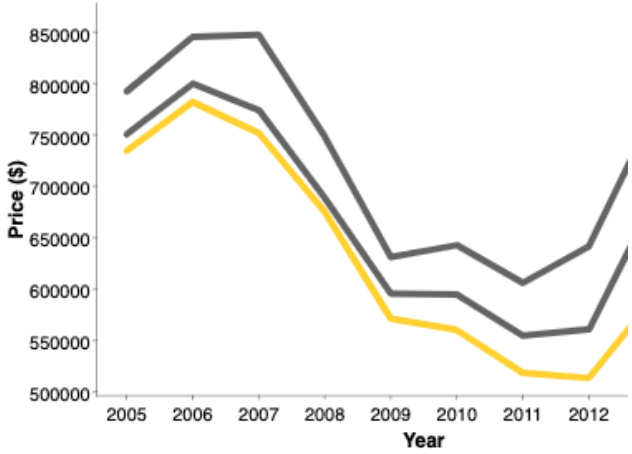


Add multiple lines on a line plot

- How do we visualize each region on a line plot?

S RegionName	I Year	D Mean(ColumnValues)
San Francisco, CA	2004	627,953.417
San Francisco, CA	2005	750,609
San Francisco, CA	2014	782,904.583
San Jose, CA	2003	601,880.917
San Jose, CA	2006	845,468.25
San Jose, CA	2015	958,072.667
Santa Cruz, CA	2004	614,160.167
Santa Cruz, CA	2008	675,264.333
Santa Cruz, CA	2021	1,078,152.833

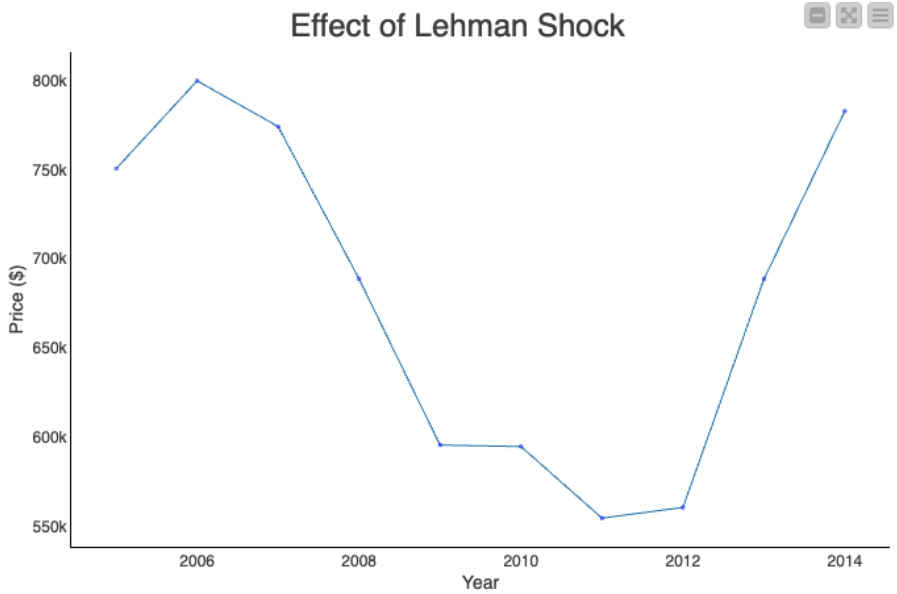
?



Background

- Typically, we plot one line using data in this format:

I Year	D San Francisco.
2005	750,609
2006	799,768.75
2007	774,172.25
2008	688,893.25
2009	595,844
2010	595,020.833
2011	554,982.25
2012	560,916.167
2013	688,696.25
2014	782,904.583



Problem

- We cannot plot multiple lines from a column with mixed data
- We'll need to reformat

Distinct region data only

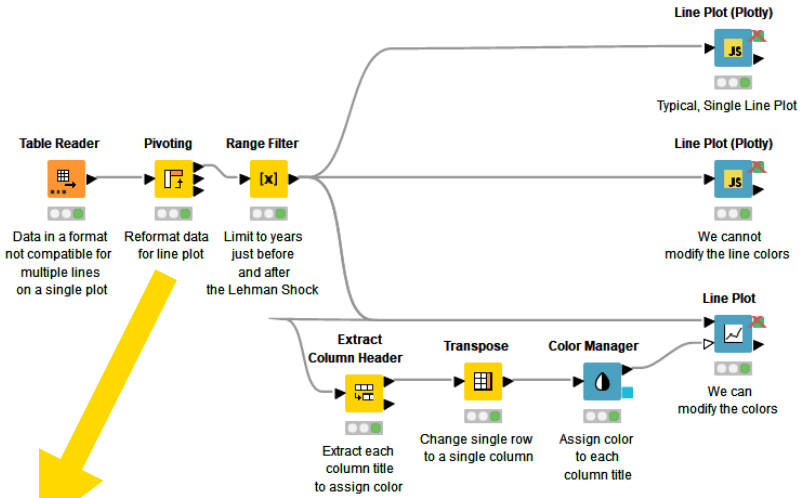
I Year	D San Francisco
2005	750,609
2006	799,768.75
2007	774,172.25
2008	688,893.25
2009	595,844
2010	595,020.833
2011	554,982.25
2012	560,916.167
2013	688,696.25
2014	782,904.583

Mixed region data

S RegionName	I Year	D Mean(ColumnValues)
San Francisco, CA	2004	627,953.417
San Francisco, CA	2005	750,609
San Francisco, CA	2014	782,904.583
San Jose, CA	2003	601,880.917
San Jose, CA	2006	845,468.25
San Jose, CA	2015	958,072.667
Santa Cruz, CA	2004	614,160.167
Santa Cruz, CA	2008	675,264.333
Santa Cruz, CA	2021	1,078,152.833

Solution

- Key Points
  1. For each line we want, we need a **distinct column**
  2. Therefore, we **pivot** the data
  3. Adding colors to the columns requires **exact column name** matching



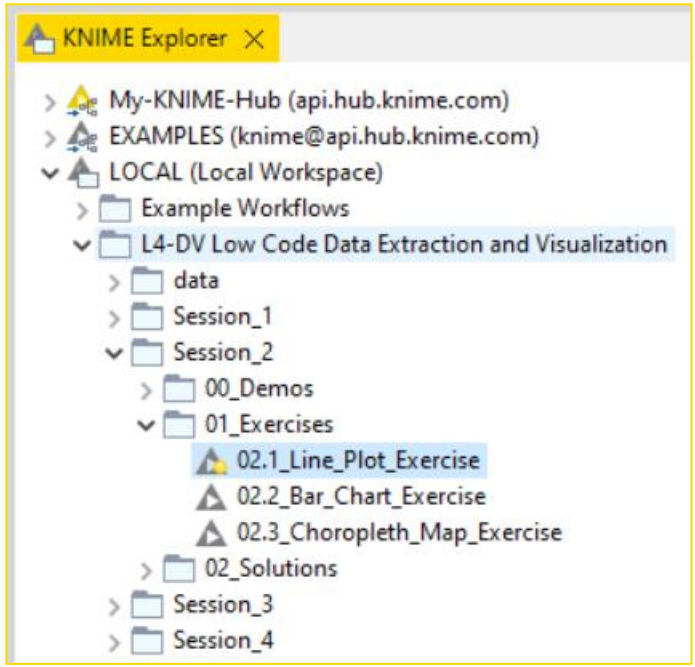
I Year	D San Francisco, CA+First*(Mean(ColumnValues))	D San Jose, C...	D Santa Cruz, ...
2000	394,841.833	489,902.75	382,738.25
2001	468,226.417	590,062.417	475,892
2002	500,029.75	585,127.25	505,906.917
2003	545,562.75	601,880.917	546,822.75
2004	627,953.417	663,949.417	614,160.167
2005	750,609	792,311	734,655
2006	799,768.75	845,468.25	782,231.917
2007	774,172.25	847,380.417	751,948.75
2008	688,893.25	748,603.583	675,264.333

Exercises: Section 2

1. Line Plot
   Using the data provided, transform it so that you can display multiple lines at once.

2. Bar Chart
   Using the data provided, transform it to display the data using custom colors.

3. Choropleth Map
   Using the data provided, transform it to create a map highlighting key assets.

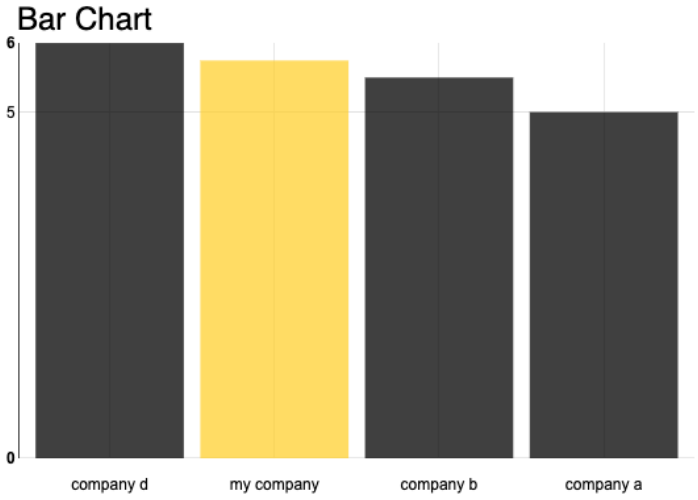


Color bars on a 2-D bar chart

- How do we go from two columns (categorical and numeric data) to a colored bar chart?

Row ID	S category	D value
Row0	a	5.001
Row1	b	5.5
Row2	c	5.747
Row3	d	5.999
Row4	e	5.1
Row5	f	5.3

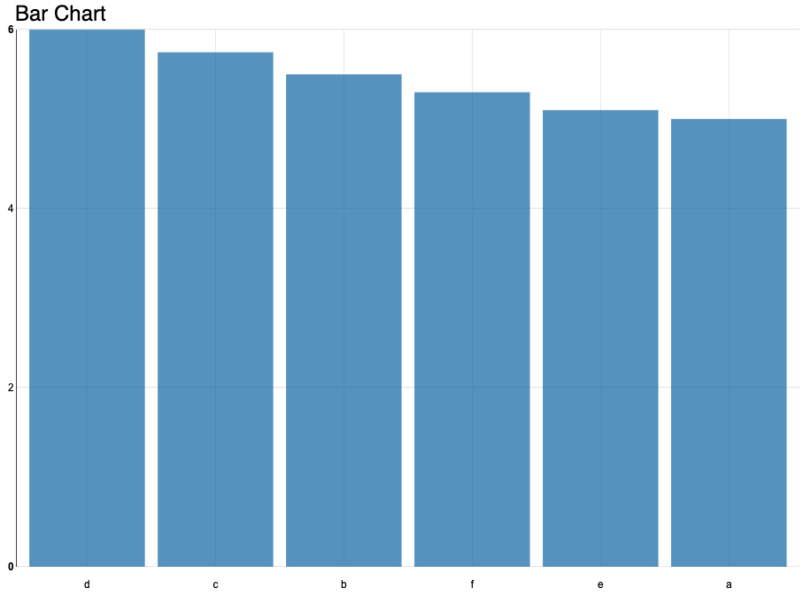
?



Background

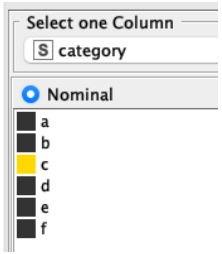
- Typically, we plot multiple bars using data in this format:

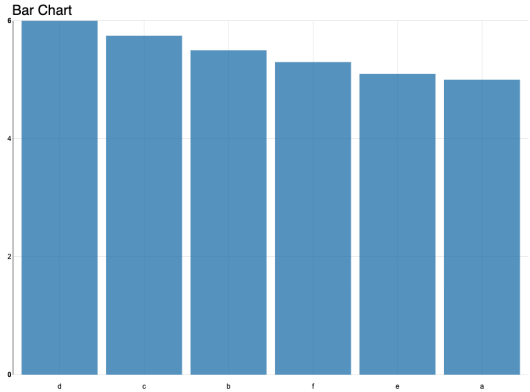
S category	D value
d	5.999
c	5.747
b	5.5
f	5.3
e	5.1
a	5.001



Problem

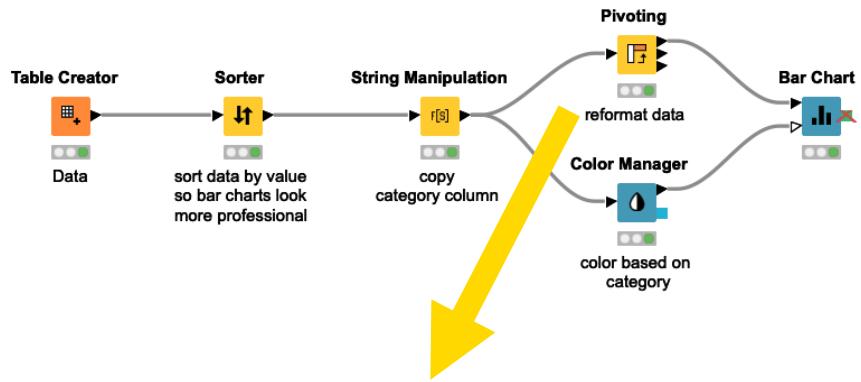
- The Color Manager node doesn’t seem to affect the bar colors.





Solution

- Key Points
  1. We need to have **one column per value.**
  2. Therefore, we make a **copy** of our column we want to color.
  3. Then we **pivot** the data.

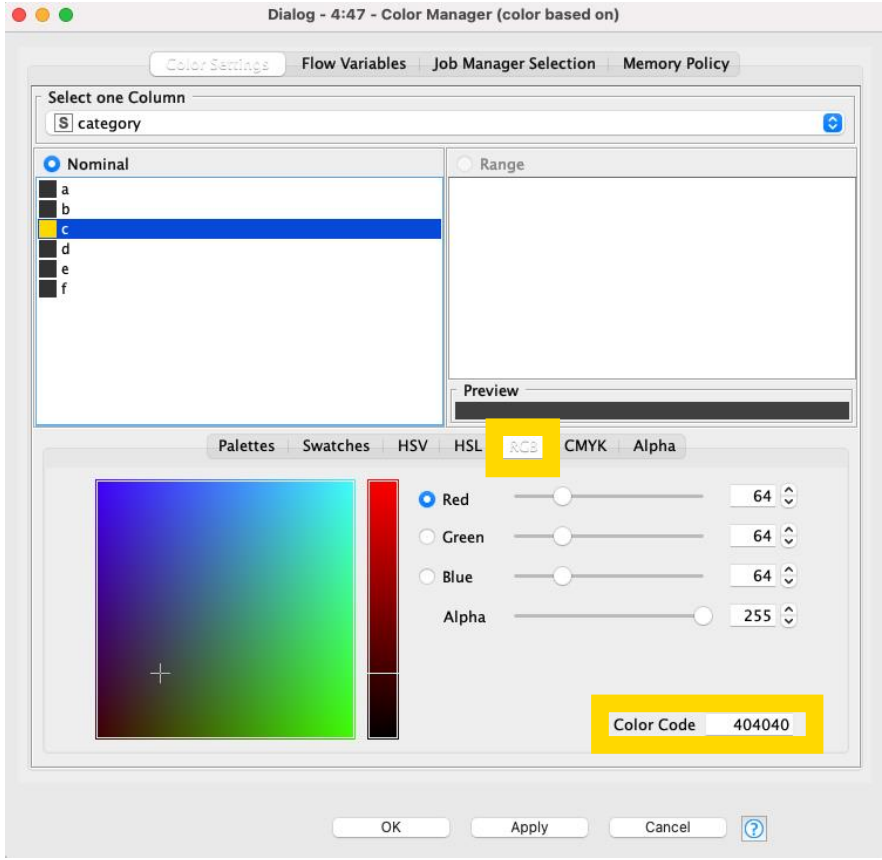


Row ID	S category	D f	D e	D d	D c	D b	D a
Row0	d	?	?	5.999	?	?	?
Row1	c	?	?	?	5.747	?	?
Row2	b	?	?	?	?	5.5	?
Row3	f	5.3	?	?	?	?	?
Row4	e	?	5.1	?	?	?	?
Row5	a	?	?	?	?	?	5.001

Adding distinct color palettes

- Use custom colors instead of the default colors or given palettes.
- For distinct colors use color codes.
  1. In the Color Manager node, select the RGB tab
  2. Then input the color code you would like

Note: Find color palettes and/or hex codes (color codes) on the web.

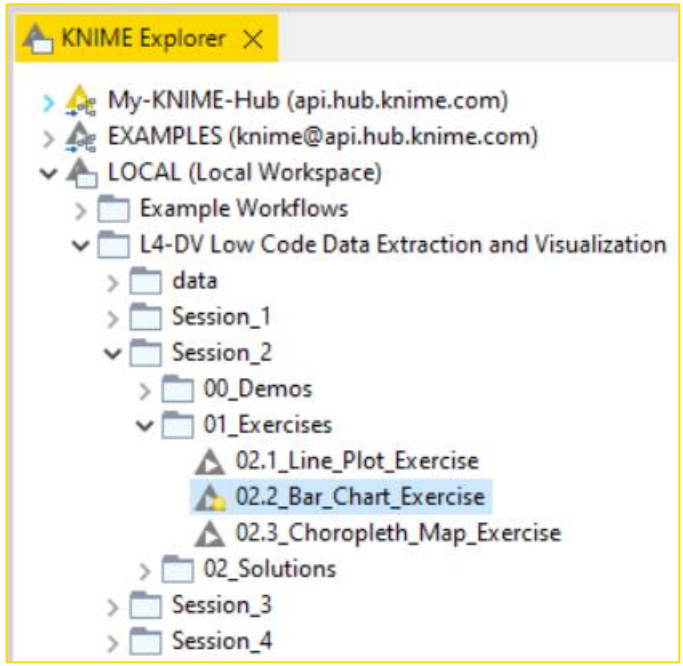


Exercises: Section 2

1. Line Plot
   Using the data provided, transform it so that you can display multiple lines at once.

2. Bar Chart
   Using the data provided, transform it to display the data using custom colors.

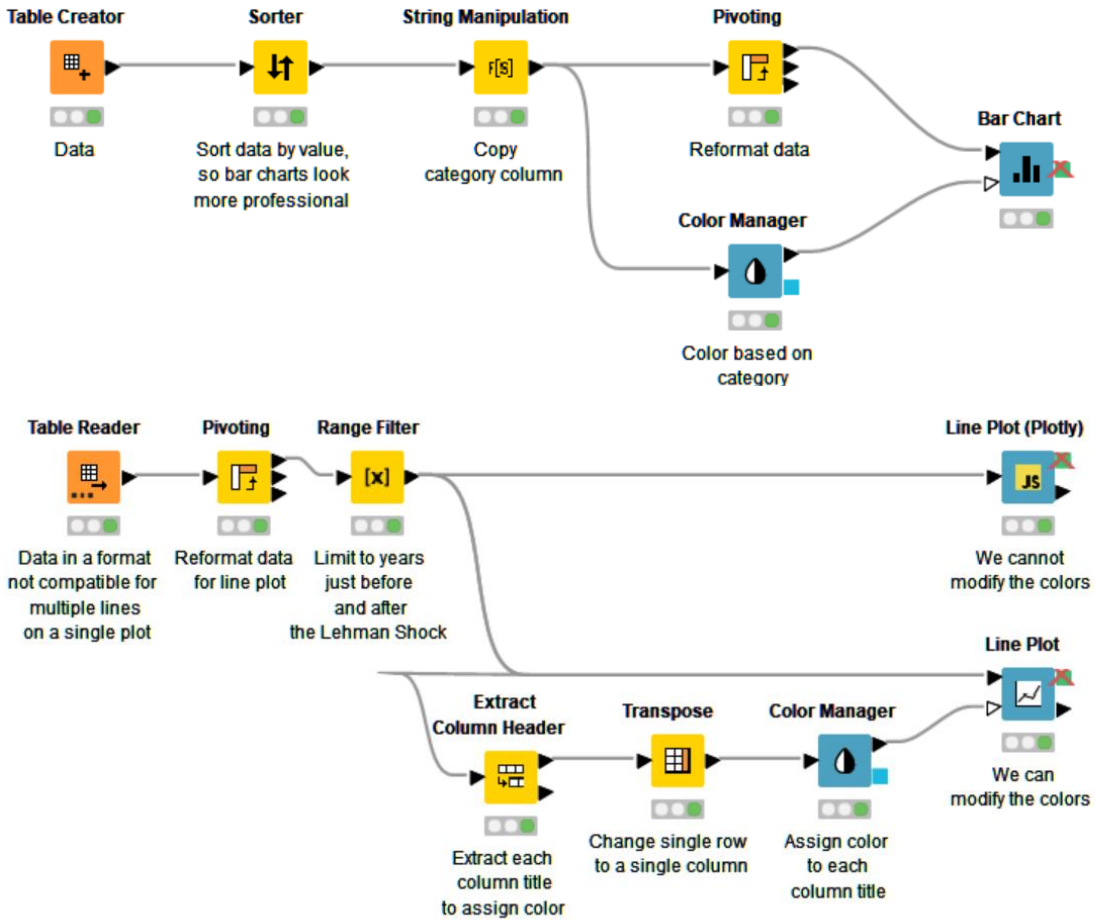
3. Choropleth Map
   Using the data provided, transform it to create a map highlighting key assets.



KNIME Knowledge Check 04

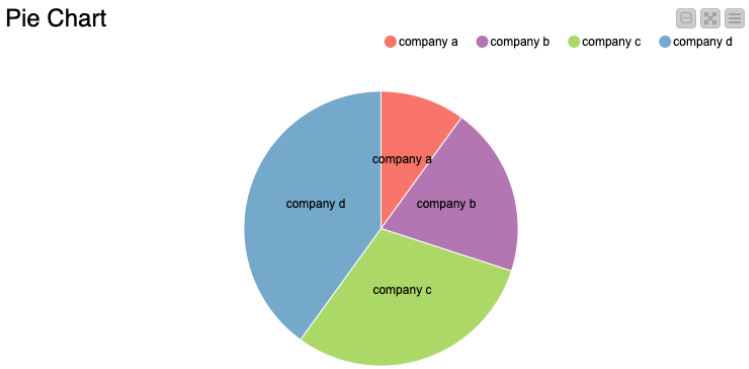
- Coloring bars on a 2D bar chart and adding lines to a line plot had a similar issue. What was it?
  - They required distinct rows with categorical data
  - They required distinct columns with numerical data
  - They required more rows with numerical data
  - They required more columns with categorical data

02: Coloring lines and bars on a chart demo

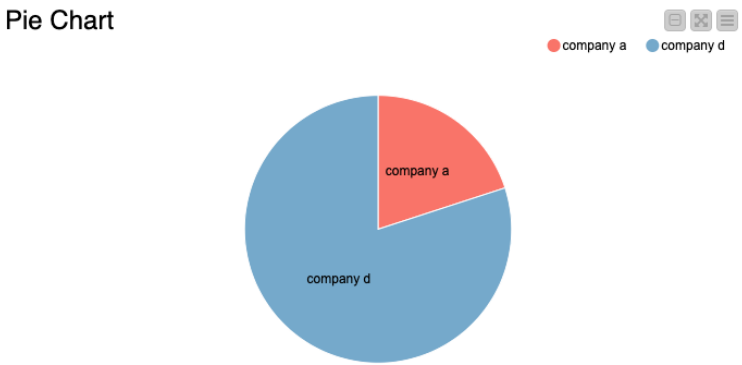


Adding filters on plots

- How do you add filtering in the interactive view to update the visuals?



?

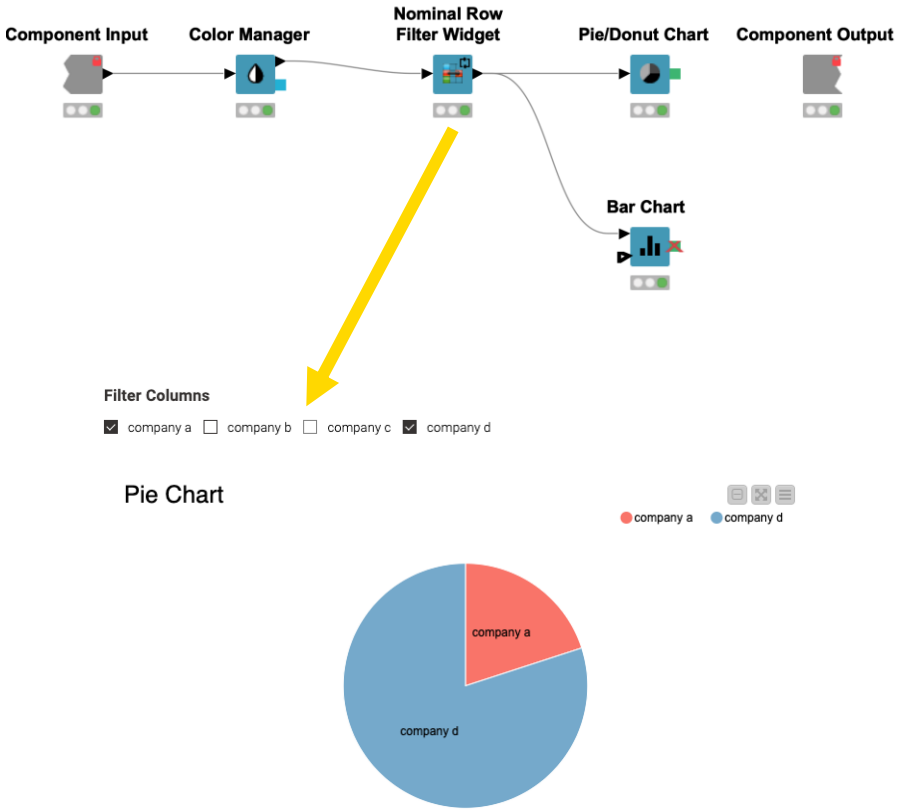


Filtering plots

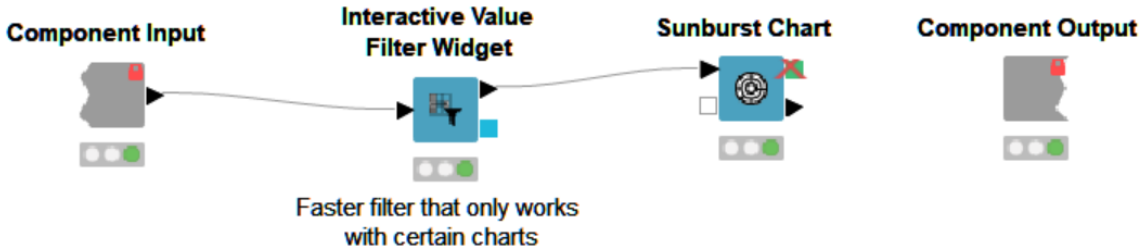
- Approaches:
  1. Try to use the **Interactive Value Filter Widget node**.

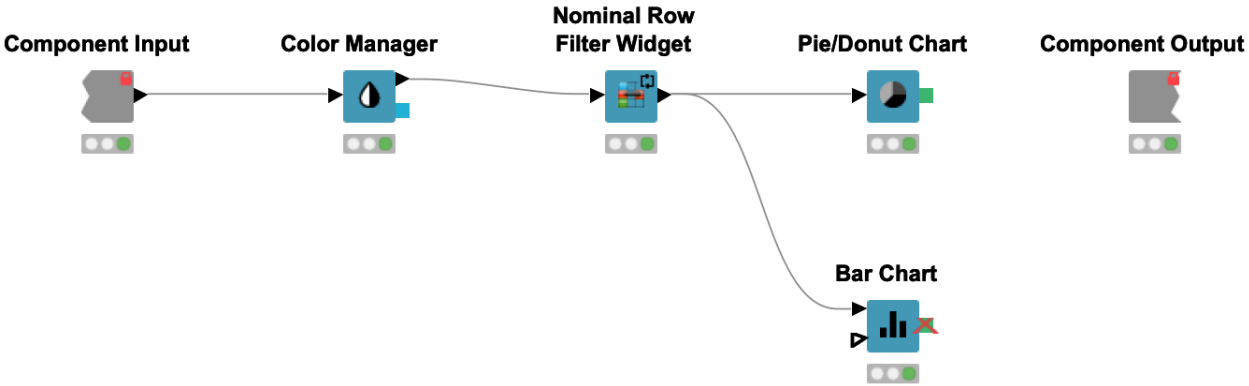
Not all visualizations in KNIME can make use of the Interactive Value Filter Widget node, so…

  2. To filter for bar charts, pie charts, etc., use the **Nominal Row Filter Widget or Column Filter Widget with Re-execution enabled**.



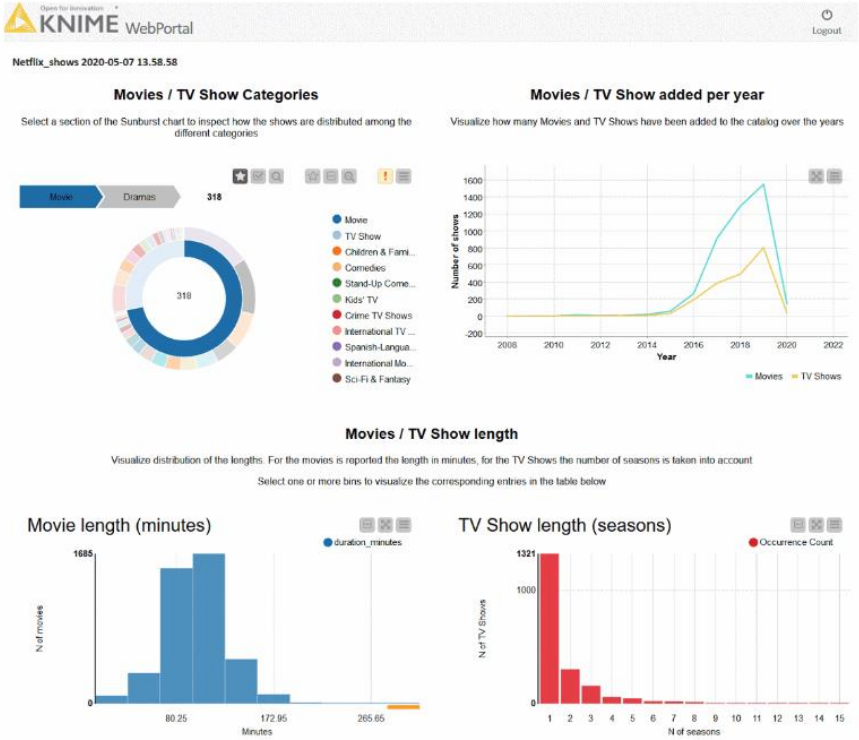
03: Adding interactive filtering demo





More resources for interactive dashboards

- https://www.knime.com/blog/how-to-create-an-interactive-dashboard-in-three-steps-with-knime

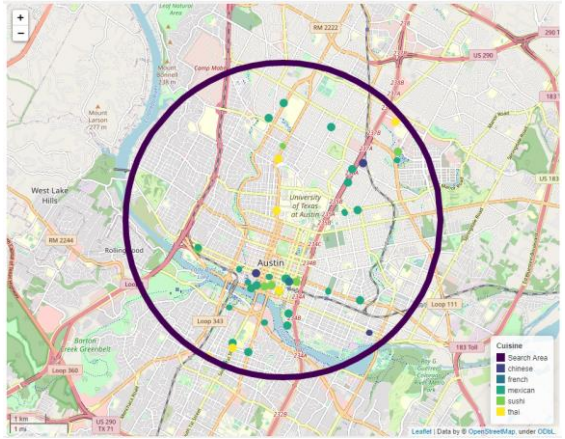


Accessing and visualizing geospatial data

- Enter and access location information and visualize it in an interactive map

Row ID	D Lat	D Long
Row0	30.282	-97.74

parking
bench
parking_space
place_of_worship
restaurant
school
waste_basket
bicycle_parking
fuel
cafe
fast_food
shelter
bank
toilets
recycling
pharmacy
post_box
kindergarten
drinking_water

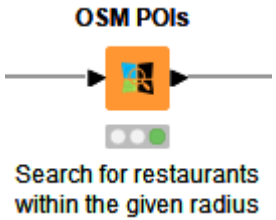


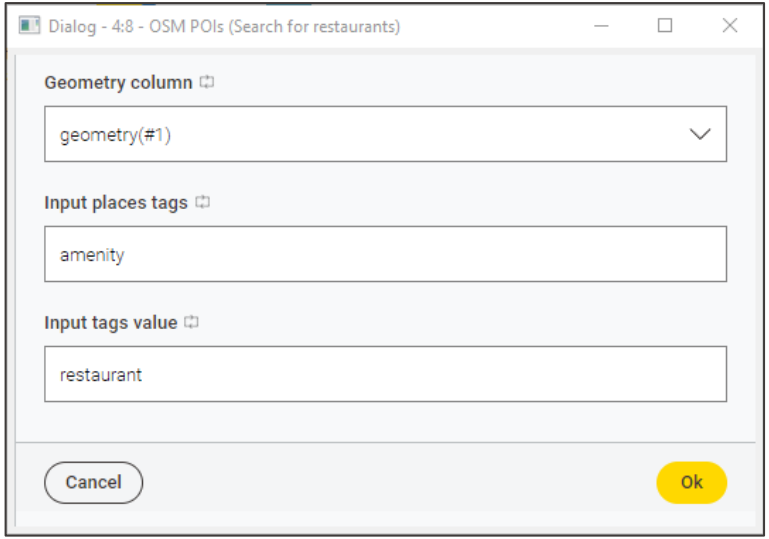
Geospatial Analytics Extension for KNIME

- KNIME Community Extension (Trusted) jointly developed by the Center for Geographic Analysis from Harvard and KNIME for
  - Reading and writing geospatial files (e.g., shapefiles, GeoJSON)
  - Performing spatial calculations (e.g., computing distances, joining)
  - Viewing data on an interactive map
- Access a collection of example workflows on the KNIME Community Hub
- View the KNIME Geospatial Analytics playlist on YouTube

OSM POIs node

- Returns the geolocations of selected entities (e.g., restaurant, parking) within a geometric boundary

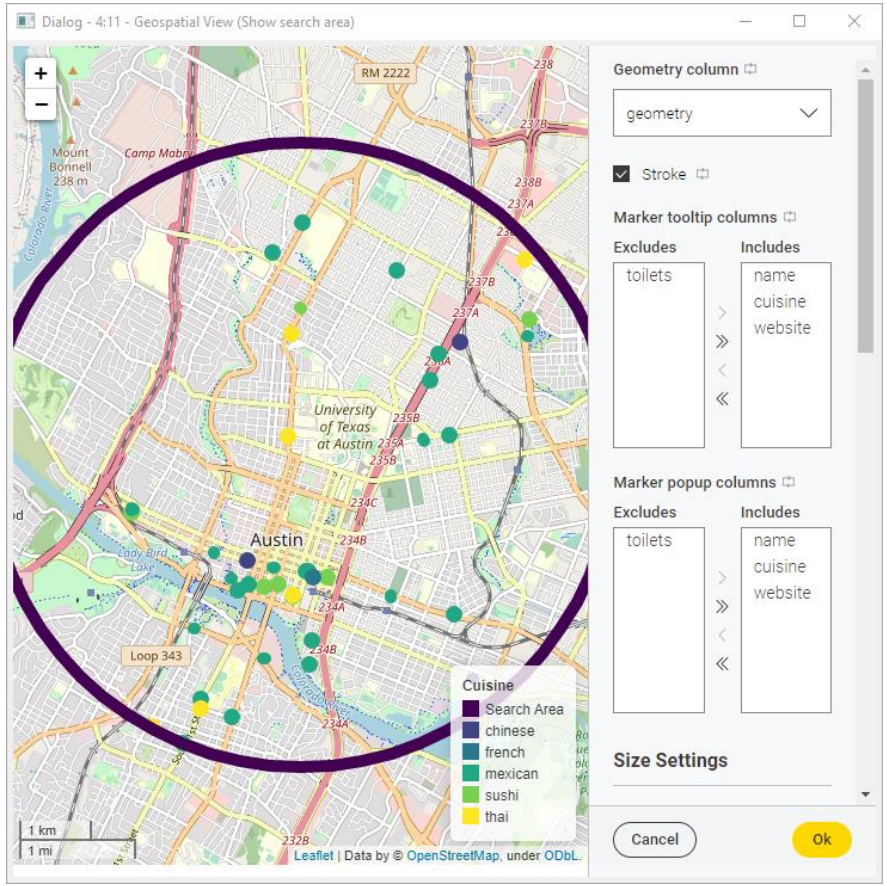




amenity	name	geometry	addr:city	addr:h...	addr:po...	addr:str...	cuisine
restaurant	Mangia Pizza	POINT - EPSG:4326	?	?	?	?	?
restaurant	Maudie's Te...	POINT - EPSG:4326	Austin	2608	78703	West 7th S...	tex-mex
restaurant	Galaxy Cafe	POINT - EPSG:4326	?	1000	78703	West Lynn St	american
restaurant	Blue Star Ca...	POINT - EPSG:4326	Austin	?	78756	?	?
restaurant	Gusto Italia...	POINT - EPSG:4326	Austin	?	78756	?	italian

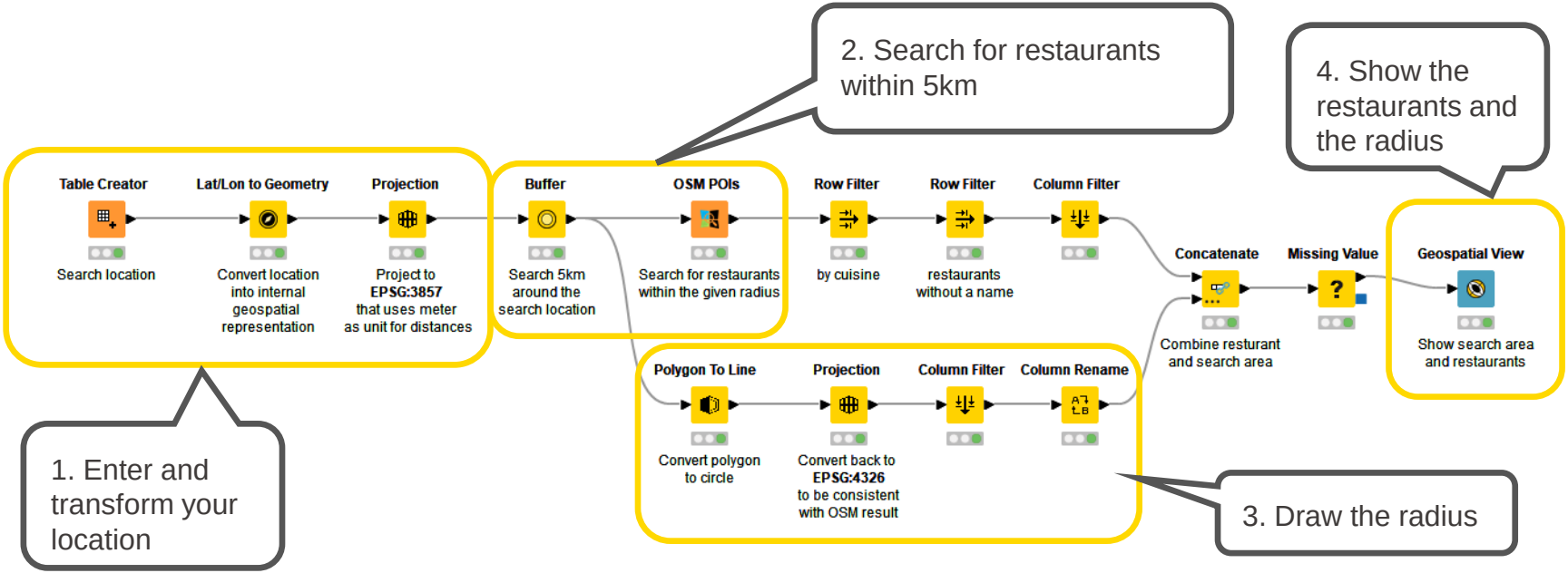
Geospatial View node

- Visualize geometric elements (e.g., points, line strings and polygons) in an interactive map and display additional information via tooltips, and shape/size/color of the markers

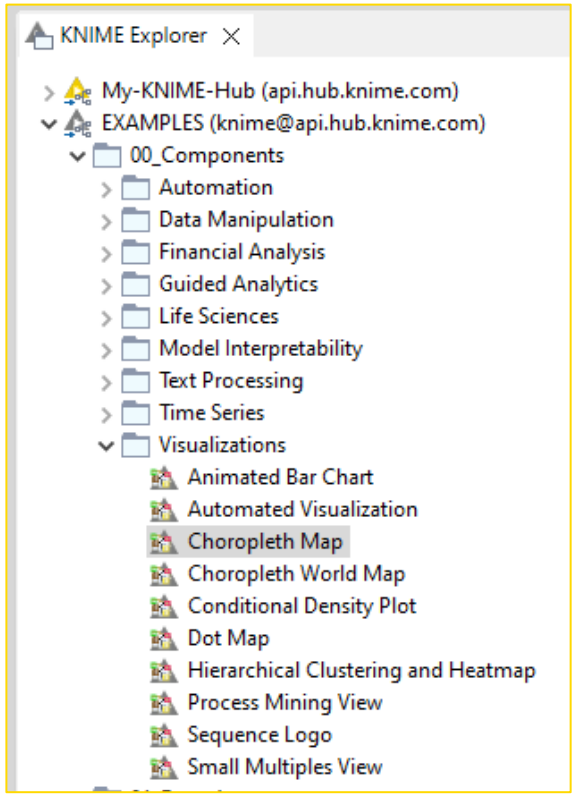


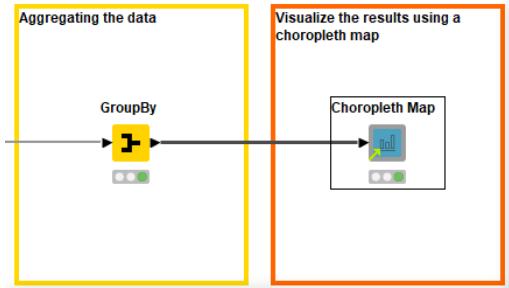
Finding restaurants nearby

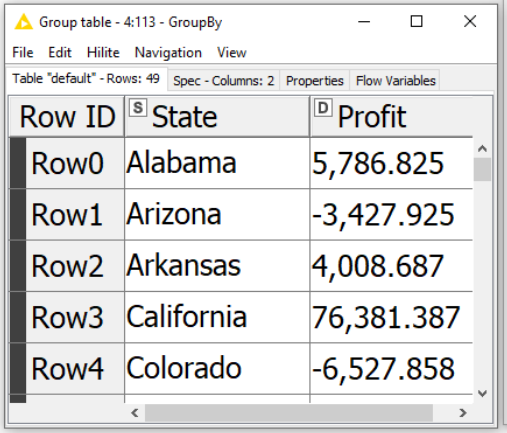
- Example workflow for preprocessing geospatial data, retrieving restaurants nearby, and visualizing their proximity in an interactive map
- View and download the [POI Search](#) workflow from the KNIME Community Hub



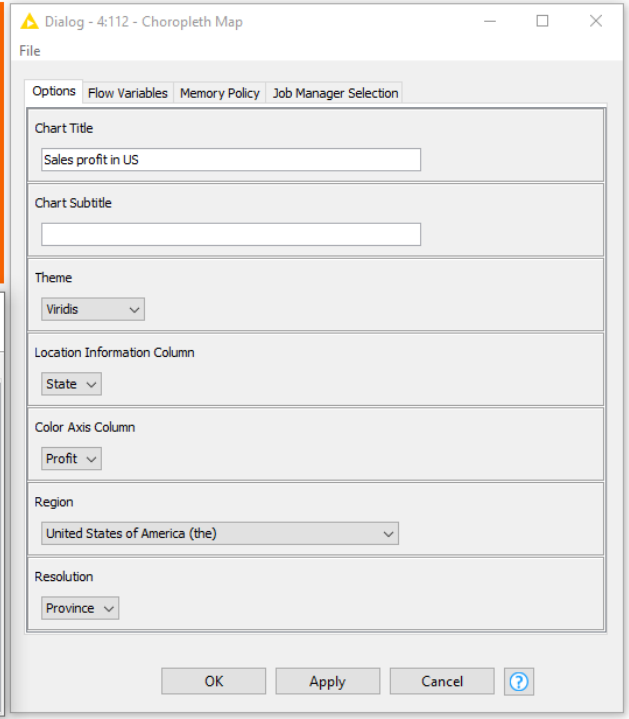
Verified Component: Choropleth Map







Row ID	State	Profit
Row0	Alabama	5,786.825
Row1	Arizona	-3,427.925
Row2	Arkansas	4,008.687
Row3	California	76,381.387
Row4	Colorado	-6,527.858

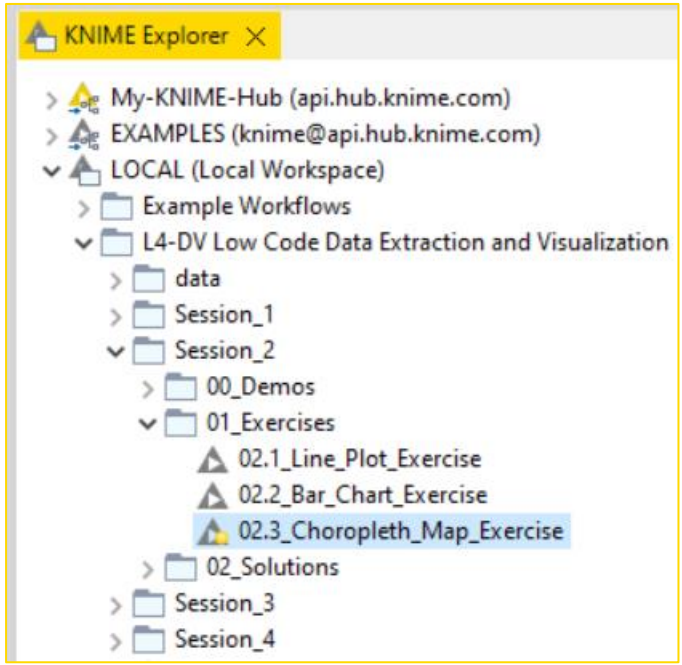


Exercises: Section 2

1. Line Plot
   Using the data provided, transform it so that you can display multiple lines at once.

2. Bar Chart
   Using the data provided, transform it to display the data using custom colors.

3. Choropleth Map
   Using the data provided, transform it to create a map highlighting key assets.



Summary of Section 2

Now you should be able to:

1. Match correct visualization for a task.

2. Apply visualizations to common tasks.

Open for Innovation
KNIME

[L4-DV] Low Code Data Extraction and Visualization

Section 3

Data Extraction

At the end of this section, you will be able to:

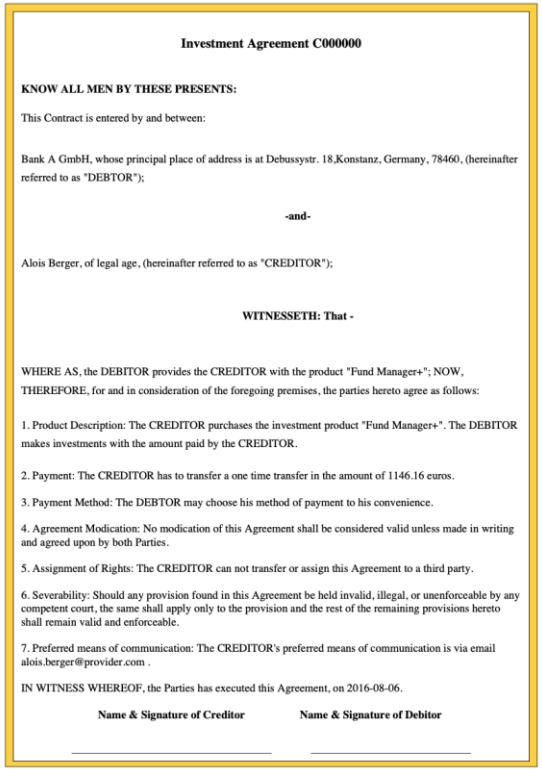
1. Parse PDFs.

2. Recognize and reproduce Regex.

3. Manipulate a data extraction tool.

What’s your type (of PDF)?

Text-based
(Can click on text in pdf)



Investment Agreement C000000

KNOW ALL MEN BY THESE PRESENTS:

This Contract is entered by and between:

Bank A GmbH, whose principal place of address is at Debussystr. 18,Konstanz, Germany, 78460, (hereinafter referred to as "DEBTOR");

-and-

Alois Berger, of legal age, (hereinafter referred to as "CREDITOR");

WITNESSETH: That -

WHERE AS, the DEBITOR provides the CREDITOR with the product "Fund Manager+"; NOW, THEREFORE, for and in consideration of the foregoing premises, the parties hereto agree as follows:

1. Product Description: The CREDITOR purchases the investment product "Fund Manager+". The DEBITOR makes investments with the amount paid by the CREDITOR.

2. Payment: The CREDITOR has to transfer a one time transfer in the amount of 1146.16 euros.

3. Payment Method: The DEBTOR may choose his method of payment to his convenience.

4. Agreement Modication: No modication of this Agreement shall be considered valid unless made in writing and agreed upon by both Parties.

5. Assignment of Rights: The CREDITOR can not transfer or assign this Agreement to a third party.

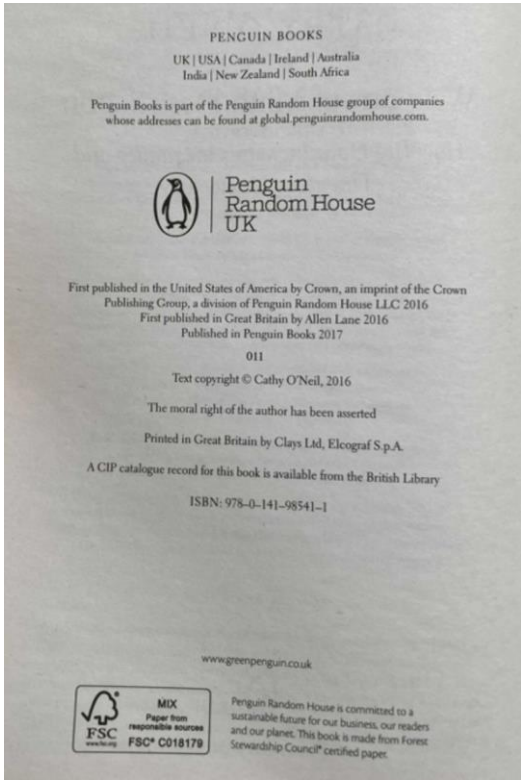
6. Severability: Should any provision found in this Agreement be held invalid, illegal, or unenforceable by any competent court, the same shall apply only to the provision and the rest of the remaining provisions hereto shall remain valid and enforceable.

7. Preferred means of communication: The CREDITOR's preferred means of communication is via email alois.berger@provider.com .

IN WITNESS WHEREOF, the Parties has executed this Agreement, on 2016-08-06.

Name & Signature of Creditor **Name & Signature of Debitor**

Image-based
(Cannot click on text in pdf)



PENGUIN BOOKS

UK | USA | Canada | Ireland | Australia
India | New Zealand | South Africa

Penguin Books is part of the Penguin Random House group of companies whose addresses can be found at global.penguinrandomhouse.com.

Penguin Random House UK

First published in the United States of America by Crown, an imprint of the Crown Publishing Group, a division of Penguin Random House LLC 2016
First published in Great Britain by Allen Lane 2016
Published in Penguin Books 2017

011

Text copyright © Cathy O'Neil, 2016

The moral right of the author has been asserted

Printed in Great Britain by Clays Ltd, Elcograf S.p.A.

A CIP catalogue record for this book is available from the British Library

ISBN: 978-0-141-98541-1

www.greenpenguin.co.uk

MIX
Paper from responsible sources
FSC* C018179

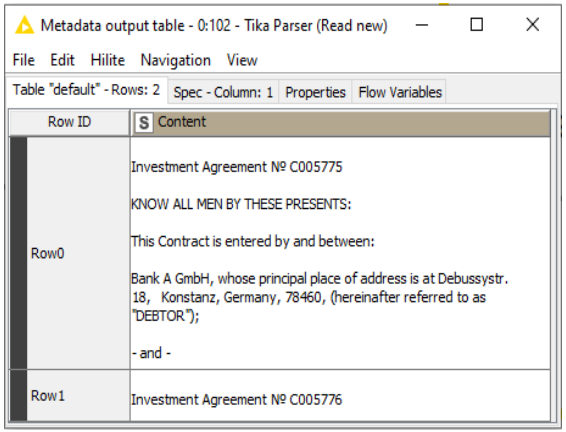
Penguin Random House is committed to a sustainable future for our business, our readers and our planet. This book is made from Forest Stewardship Council* certified paper.

Open for Innovation KNIME

Text-based PDF extraction: Tika Parser

- Tika Parser extracts text **with formatting** from text-based **sources**

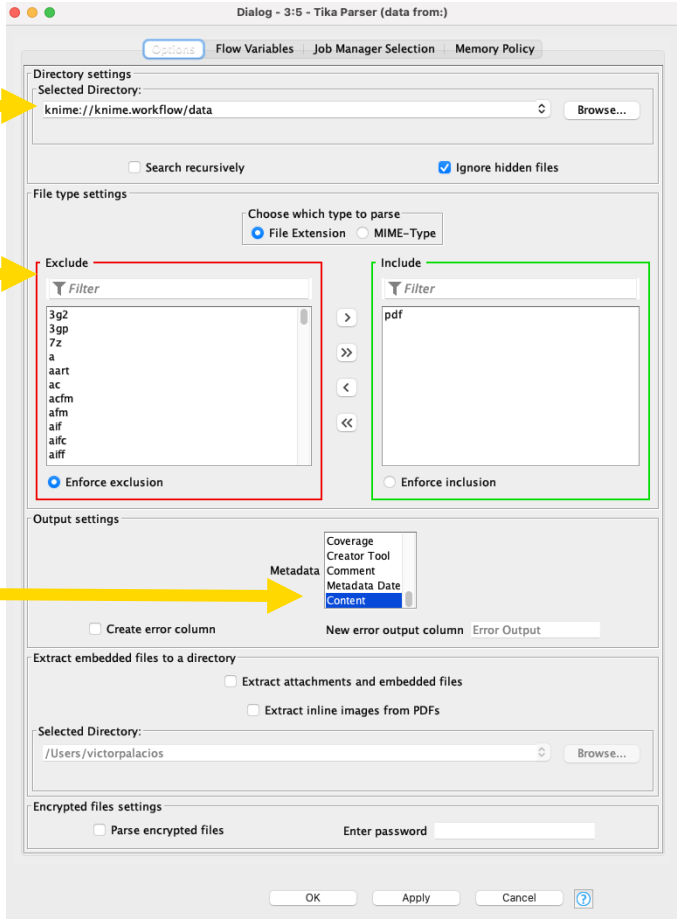
Tika Parser



Data Location

Data Type

Metadata Desired

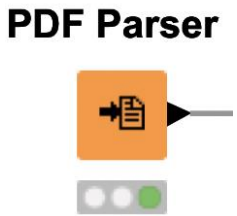


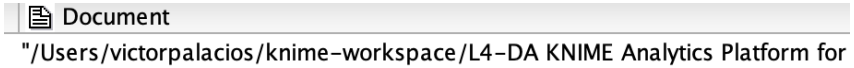
Example of how to use the [Tika Parser on the KNIME Hub]

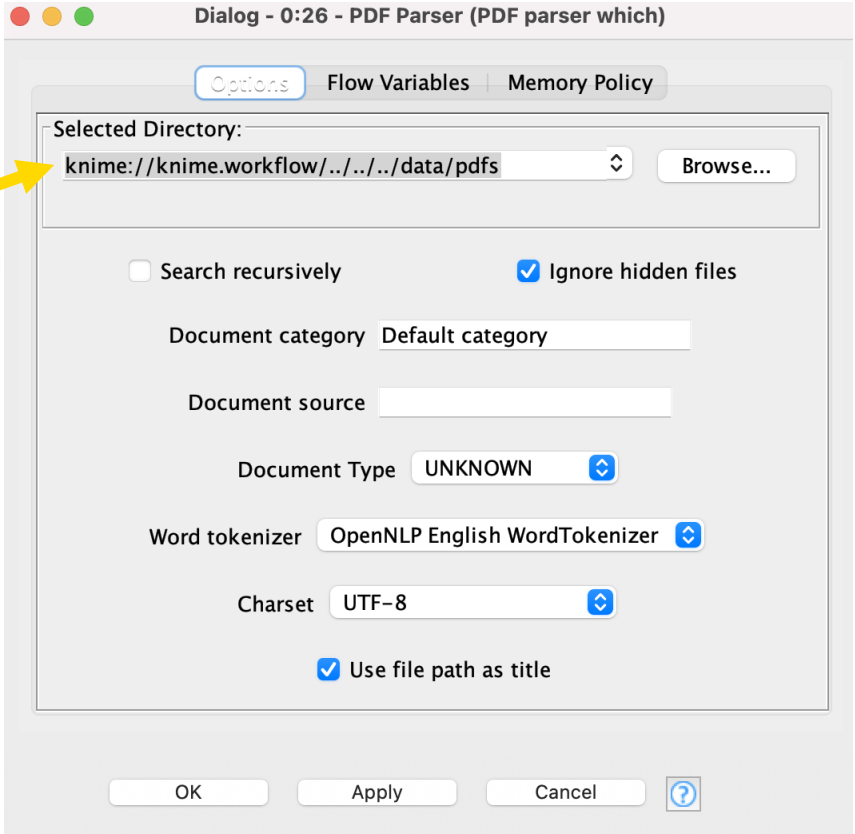
Text-based PDF extraction: PDF Parser

- PDF Parser extracts text **without formatting** from text-based **PDFs** as a **document type**

Data Location

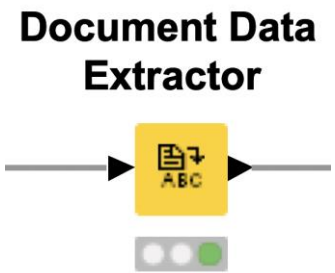




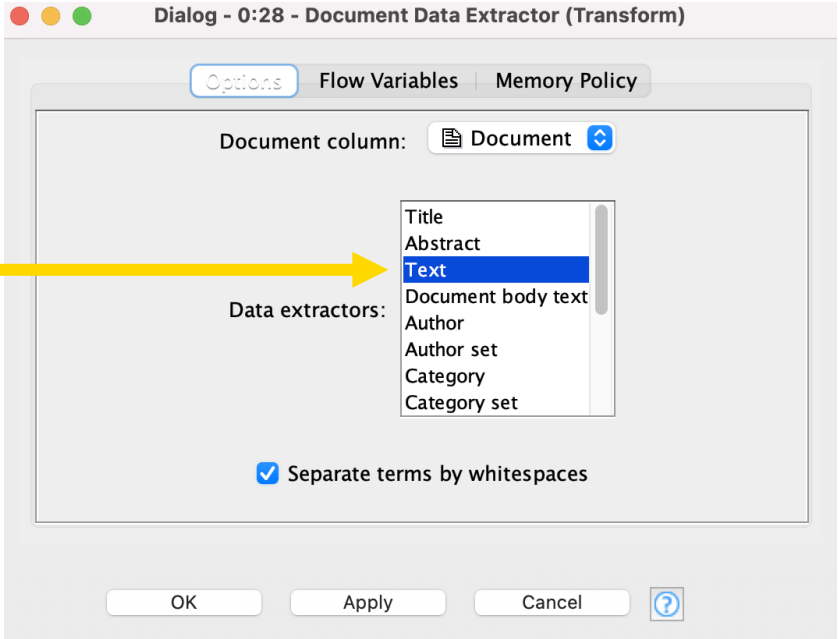


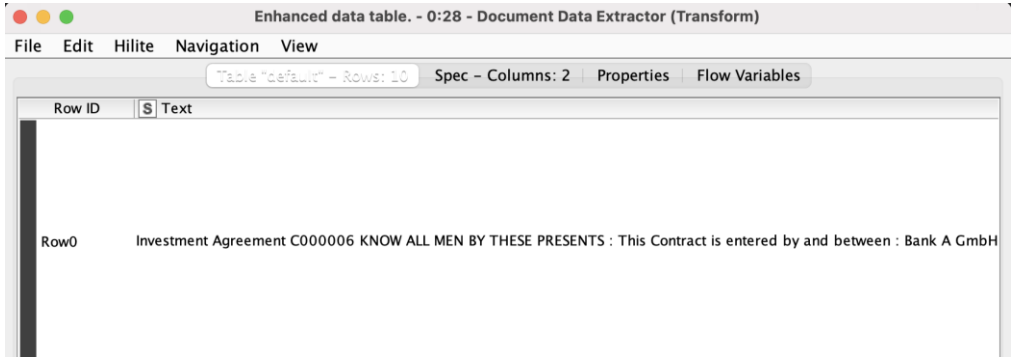
Text-based PDF extraction: Document Data Extractor

- Document Data Extractor extracts text from a document type.



Metadata Desired





Tika Parser vs PDF Parser

	Tika Parser	PDF Parser
Versatility	reads many file types	reads only PDFs
Output Type	string	document
Output Text	text only	text and path
Formatting	kept	removed

Tika Parser or PDF Parser?

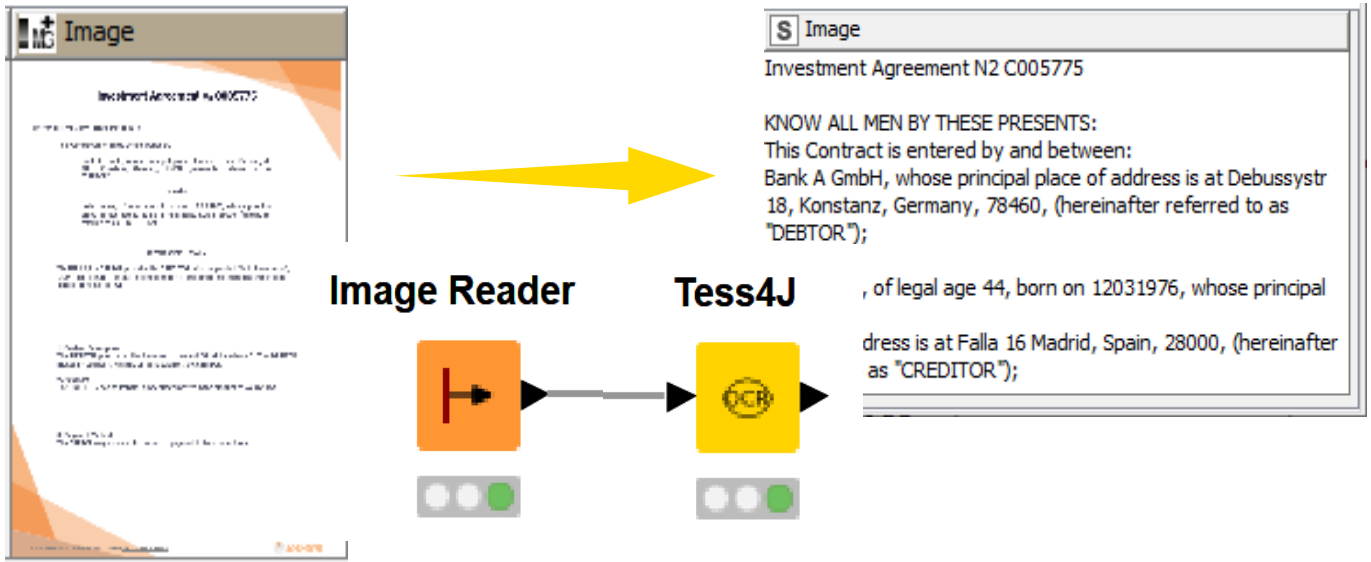
- Use the Tika Parser node when…
  - You need the formatting preserved
  - You need to extract data from multiple sources including PDFs

- Use the PDF Parser node when…
  - You want the formatting removed
  - You are doing Natural Language Processing or Text Processing tasks

Image-based PDF extraction

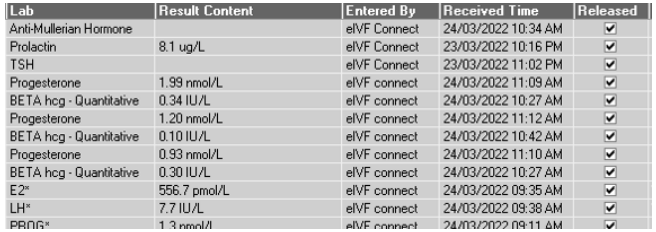
Tess4J node

- Tess4J for OCR (Optical Character Recognition) only on Windows



Example of how to use the Tess4J node on the KNIME Hub

Image quality and structure determine output

	High Quality Image maximum differentiation between background and target text (i.e., white and black color scheme)	Low Quality Image poor differentiation between background and target text and tabular (i.e., various shades of grey)
Image Input	Besides tweaking the configurations of the Tess4J node to the use case at hand, it is a good practice to preprocess input images thoroughly, if needed. In particular, Tesseract works best when images are sufficiently scaled up such that the pixel count of the x-height of characters is at least 20 pixels; images are correctly aligned and have a sufficiently high resolution; and any dark borders is removed, or they might be misinterpreted as characters [3]. The KNINE Image Processing extension includes several nodes for image cleaning, manipulation and transformation, and many example workflows can be found on the KNIME Hub.	
KNIME Output	Besides tweaking the configurations of the Tess4J node to the use case at hand, it is a good practice to preprocess input images thoroughly, if needed. In particular, Tesseract works best when images are sufficiently scaled up such that the pixel count of the x-height of characters is at least 20 pixels; images are correctly aligned and have a sufficiently high resolution; and any dark borders is removed, or they might be misinterpreted as characters [3]. The KNINE Image Processing extension includes several nodes for image cleaning, manipulation and transformation, and many mple workflows can be found on the KNIME Hub.	Arm MuHenan Hum Pm'achn TSH Manama: am he: - mam-v. Pvagexlsvuns BETA hcg mama-v: ngeslemne am hcg mam»: E2" M" van:- 2' ug/L ' as nmuVL u 93 mum/L u an m/L 555 7 max/L eWF Emma-2A 2un3/2u22 m 3o AM 2 1

Lab	Result Content	Entered By	Received Time	Released
Anti-Mullerian Hormone		eIVF Connect	24/03/2022 10:34 AM	☑
Prolactin	8.1 ug/L	eIVF Connect	23/03/2022 10:16 PM	☑
TSH		eIVF Connect	23/03/2022 11:02 PM	☑
Progesterone	1.99 nmol/L	eIVF connect	24/03/2022 11:09 AM	☑
BETA hcg - Quantitative	0.34 IU/L	eIVF connect	24/03/2022 10:27 AM	☑
Progesterone	1.20 nmol/L	eIVF connect	24/03/2022 11:12 AM	☑
BETA hcg - Quantitative	0.10 IU/L	eIVF connect	24/03/2022 10:42 AM	☑
Progesterone	0.93 nmol/L	eIVF connect	24/03/2022 11:10 AM	☑
BETA hcg - Quantitative	0.30 IU/L	eIVF connect	24/03/2022 10:27 AM	☑
E2"	556.7 pmol/L	eIVF connect	24/03/2022 09:35 AM	☑
LH"	7.7 IU/L	eIVF connect	24/03/2022 09:38 AM	☑
PROG"	1.3 nmol/L	eIVF connect	24/03/2022 09:11 AM	☑

See KNIME Forum discussion

After successful extraction… What's next?

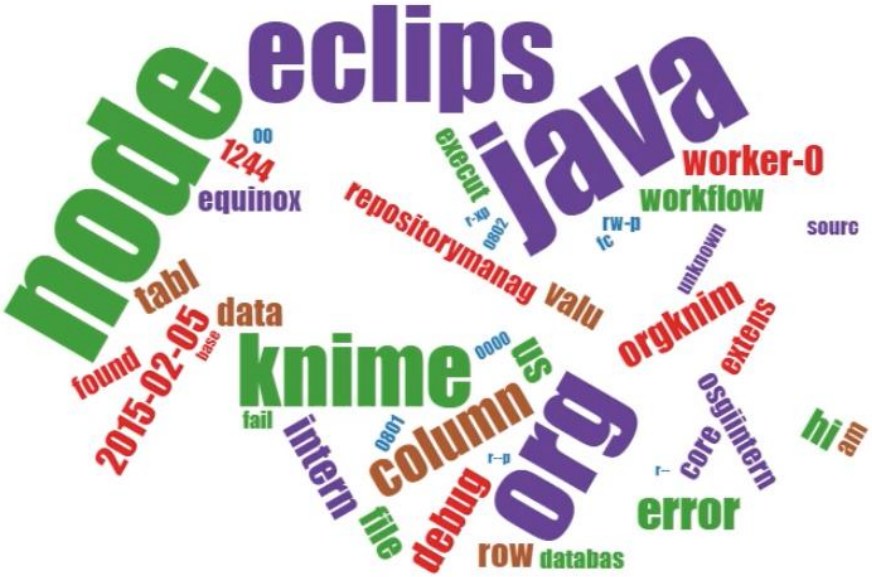
Common tasks after extraction:

1. Table Creation
2. Topics Labeling or Modeling
3. Target Text Analysis

WS FORM F-6

PRELIMINARY LOCAL CLIMATOLOGICAL DATA

LATITUDE 44 DEGREES 16 MINUTES NORTH						LONGITUDE	
TEMPERATURE (°F)							
						DEGREE DAYS	
DAY	MAX	MIN	AVG	NORM	DEPART	HEAT	COOL
1	32	27	30	37	-7	35	0
2	42	31	37	37	0	28	0
3	45	41	43	37	6	22	0
4	48	41	45	36	9	20	0
5	50	43	47	36	11	18	0
6	52	41	47	35	12	18	0
7	53	41	47	35	12	18	0
8	52	38	45	34	11	20	0
9	51	38	45	34	11	20	0
10	48	39	44	34	10	21	0



IN WITNESS WHEREOF, the Parties has executed this Agreement, o 2016-08-06.

Name & Signature of Creditor **Name & Signature of Debitor**

Table creation

- Analyze weather data from a PDF table

Text-based PDF (with table)

WS FORM F-6

PRELIMINARY LOCAL CLIMATOLOGICAL DATA

LATITUDE						LONGITUDE	
44 DEGREES 16 MINUTES NORTH							
TEMPERATURE (°F)							
						DEGREE DAYS	
DAY	MAX	MIN	AVG	NORM	DEPART	HEAT	COOL
1	32	27	30	37	-7	35	0
2	42	31	37	37	0	28	0
3	45	41	43	37	6	22	0
4	48	41	45	36	9	20	0
5	50	43	47	36	11	18	0
6	52	41	47	35	12	18	0
7	53	41	47	35	12	18	0
8	52	38	45	34	11	20	0
9	51	38	45	34	11	20	0
10	48	39	44	34	10	21	0

Tika Parser Output (1 row)

Metadata output table - 3:5 - Tika Parser (data from:)

File Edit Hilite Navigation View

Table "default" – Rows: 1 | Spec – Column: 1 | Properties | Flow Variables

Row ID	S Content
Row0	WS FORM F-6 STATION MOUNT WASHINGTON OBSERVATORY PRELIMINARY LOCAL CLIMATOLOGICAL DATA MONTH YEAR OCTOBER 2021 LATITUDE LONGITUDE GROUND ELEVATION (H) STANDARD TIME 44 DEGREES 16 MINUTES NORTH 71 DEGREES 18 MINUTES WEST 6280 FT EASTERI TEMPERATURE (°F) PRECIPITATION (IN.) WIND (MPH) SUNSHINE SKY DEGREE DAYS TOTAL SNOW & SNOW/ICE ON AVG FASTEST MILE (MINUTES) COVER DAY MAX MIN AVG NORM DEPART HEAT COOL (EQUIV) ICE GROUND-7AM SPEED S 1 32 27 30 37 -7 35 0 0.04 T 1 49.2 77 310 (NW) 0 0 10 1246 2 42 31 37 37 0 28 0 0.85 T 1 37.1 55 300 (NW) 0 0 10 1246 3 45 41 43 37 6 22 0 0.98 0.0 0 23.6 44 280 (W) 0 0 10 12 4 48 41 45 36 9 20 0 0.02 0.0 0 6.8 14 210 (SW) 235 33 9 12 5 50 43 47 36 11 18 0 0.01 0.0 0 5.0 14 310 (NW) 60 9 10 12 6 52 41 47 35 12 18 0 0.00 0.0 0 6.6 15 330 (NW) 697 99 3 12 7 53 41 47 35 12 18 0 0.00 0.0 0 14.3 38 290 (W) 699 100 3 12 8 52 38 45 34 11 20 0 0.00 0.0 0 8.8 27 340 (N) 680 98 5 12 9 51 38 45 34 11 20 0 0.00 0.0 0 9.4 33 230 (SW) 654 95 8 10 48 39 44 34 10 21 0 0.00 0.0 0 17.4 34 290 (W) 285 41 10 12

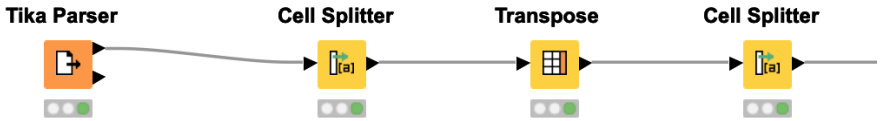
Workflow Output (1 table)

Filtered table - 3:1366 - Column Filter (remove)

File Edit Hilite Navigation View

Table "default" – Rows: 31 | Spec – Columns: 8 | Properties | Flow Variables

Row ID	S Day	S Max	S Min	S Avg	S Norm	S Depart	S Heat	S Cool
Row14	1	32	27	30	37	-7	35	0
Row15	2	42	31	37	37	0	28	0
Row16	3	45	41	43	37	6	22	0
Row17	4	48	41	45	36	9	20	0
Row18	5	50	43	47	36	11	18	0
Row19	6	52	41	47	35	12	18	0
Row20	7	53	41	47	35	12	18	0
Row21	8	52	38	45	34	11	20	0
Row22	9	51	38	45	34	11	20	0
Row23	10	48	39	44	34	10	21	0



data from: https://www.mountwashington.org/uploads/forms/2021/10.pdf

separate all new lines

columns to rows

split by space

See KNIME Forum discussion

Topic labeling

- Categorize each row with one or more **known** labels
  - Example: Look for “crack, material, vibration, missing, and coating” in the text:

Input with renamed columns - 4:1358 - Column Rename (Regex) (rename all splits)

File Edit Hilite Navigation View

Table "default" – Rows: 37 | Spec – Columns: 3 | Properties | Flow Variables

Row ID	S Finding	S Category 1	S Category 2
Row0	part inspekted, foud cracks	crack	?
Row1	part inspected, found crack	crack	?
Row2	past final inspection after applicatiion of new material. for further ...	material	?
Row3	test failed due to vibrations	vibration	?
Row4	missing coating detected. part should be transferd to shop 67-2	missing	coating

See KNIME Forum discussion

Topic modeling

- Discover **unknown** topics in your text
  - Example: Parse news feeds and try to automatically detect topics.

Document	Assi...
"york mets apologize mascot met fan finger"	topic_0
"stanley cup final penguin 2-0 series lead l...	topic_0
"forget story pushball game giant bewitch b...	topic_0

Probably "sports" is the topic

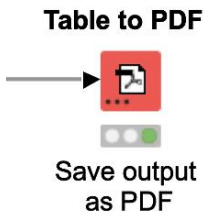
Document	Assi...
"brine turkey mad hunky meat poultry bri...	topic_3
"motor city bbq sauces"	topic_3
"smoked aubergine recipe eggplant miso ...	topic_3

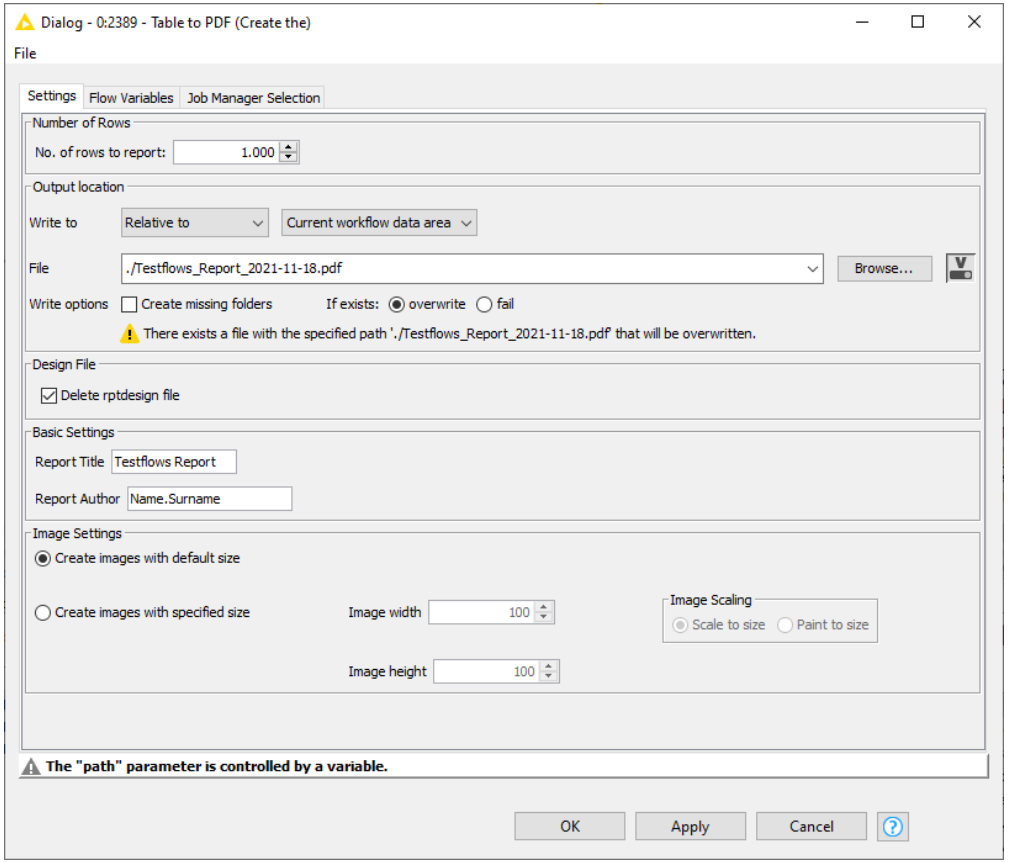
Probably "food" is the topic

See KNIME Hub example

Bonus: Table to PDF node

- Create very basic PDF report
  - With report title, author, & date





Data Extraction

At the end of this section, you will be able to:

1. Parse PDFs.
2. Recognize and reproduce Regex.
3. Manipulate a data extraction tool.

Regex (Regular Expressions)

- What is it?
  - A special set of characters that can help with…
    - Searching through text
    - Replacing text
    - Filtering / Splitting columns

- Why use it?
  - You know there is a certain **pattern** in your documents (such as date) but it is time-consuming to find it manually.

- Let’s see some pattern examples.

Regex possible targets

Dates with various formats

2022-2-22

7/9/2010

12.25.2010

Regex example

2022-2-22

[0-9]{4}-[0-9]{1,2}-[0-9]{1,2}

Regex example

2022-2-22

Any Number

[0-9]{4}-[0-9]{1,2}-[0-9]{1,2}

Occurring 4 times

Occurring 1 or 2 times

Mini regex cheat sheet

Category	Expression	Will find	Example	Example extraction:
Wildcard	.	any character	F.r	For, Fur, Far, etc.
(Unknown) Quantity	+ or *	repeated pattern	[0-9]+	77, 9, 123, etc.
Letters	[A-Z]	any uppercase letter	[a-zA-Z]	A, a, B, b, etc.
Or	|	Pattern on left or pattern on right of |	[0-9]{4}|0-9]{2}	4 digit sequence or 2 digit sequence
Capture Group	()	a group to capture	(.*)([0-9]{4})(.*)	Captures 3 groups: 1. (any sequence) 2. (4 digits) 3. (any sequence)

See full regex documentation

KNIME Knowledge Check 01

- What regex can capture the dates or IDs in these sentences:

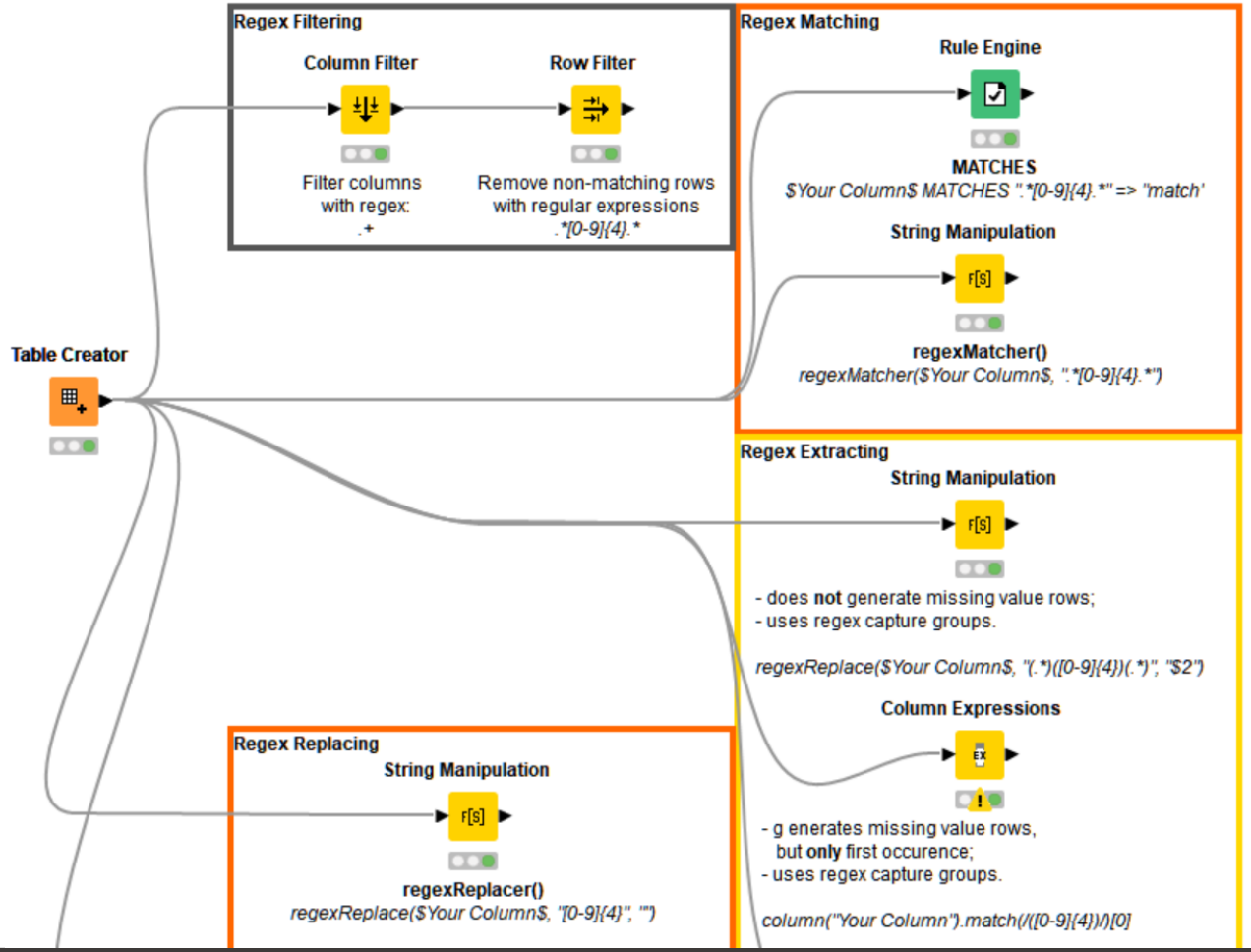
Date	IDs
12-4-2002 is the date	The id is D0000
Today's date is: 12/4/2002	D001 is the ID

Hint: You can use a combination of [0-9], period wildcard, and the repeated pattern expression +

KNIME Knowledge Check 01: Solution

Date	IDs
12-4-2002 is the date	The id is D0000
Today's date is: 12/4/2002	D001 is the ID
[0-9]{2}.[0-9]{1}.[0-9]{4} or **[0-9]+.[0-9]+.[0-9]+**	**D[0-9]{3,4}** or **D[0-9]+**

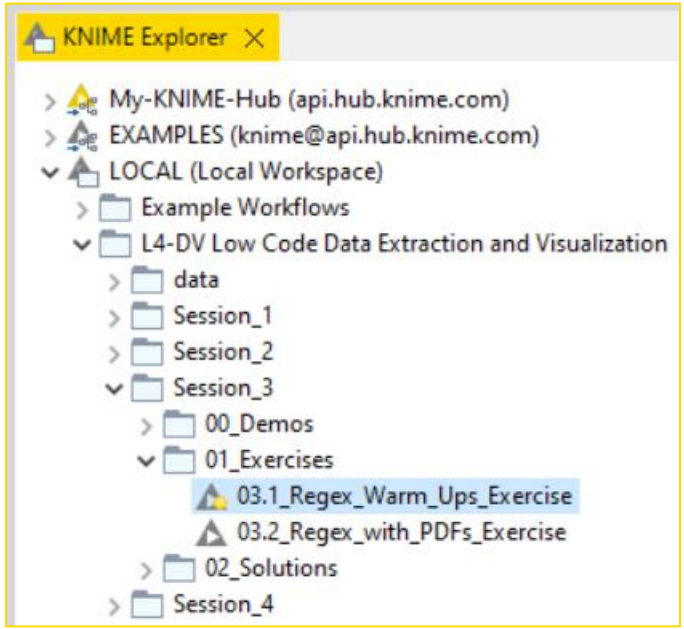
01: Various Examples of Regex Demo



Exercises: Section 3

1. Regex Warm Ups
   Using the data provided, extract information using regex.

2. Regex with PDFs
   Using the PDFs provided, parse each PDF and then extract the relevant information using regex.



KNIME Knowledge Check 02

- Write regex that can capture the date-like and true date entities in these sentences:

Date-Like	**Date**
12-4-2002 is the date.	12-4-2002 is the date.
Today's month and year is: 12-2005.	Today's date is: December-14-2004

Hint: Use the "or" expression |

KNIME Knowledge Check 02: Solution

Date-Like

12-4-2002 is the date.

Today’s month and year is: 12-2005.

Date

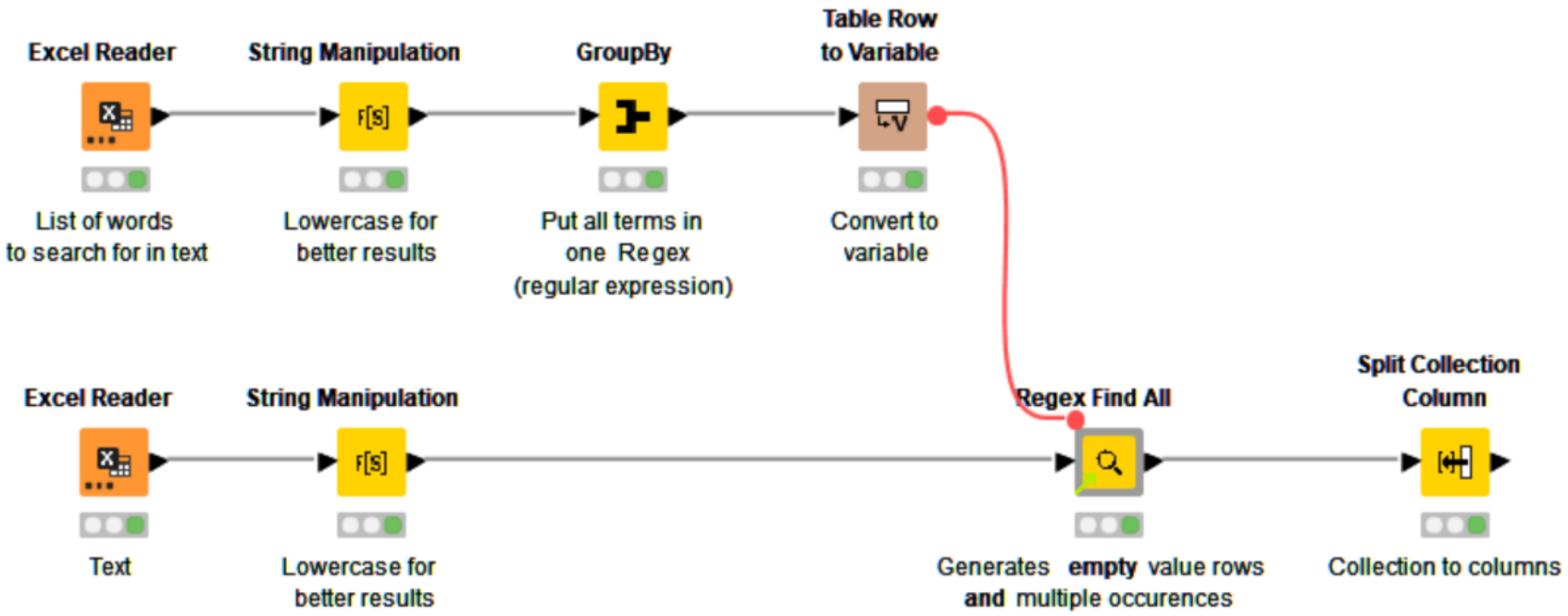
12-4-2002 is the date.

Today’s date is: December-14-2004

[0-9]+.[0-9]+.[0-9]+**|[0-9]+-[0-9]{4}**

[a-zA-Z0-9]+.[0-9]+.[0-9]+

02: Topic Labeling Regex demo



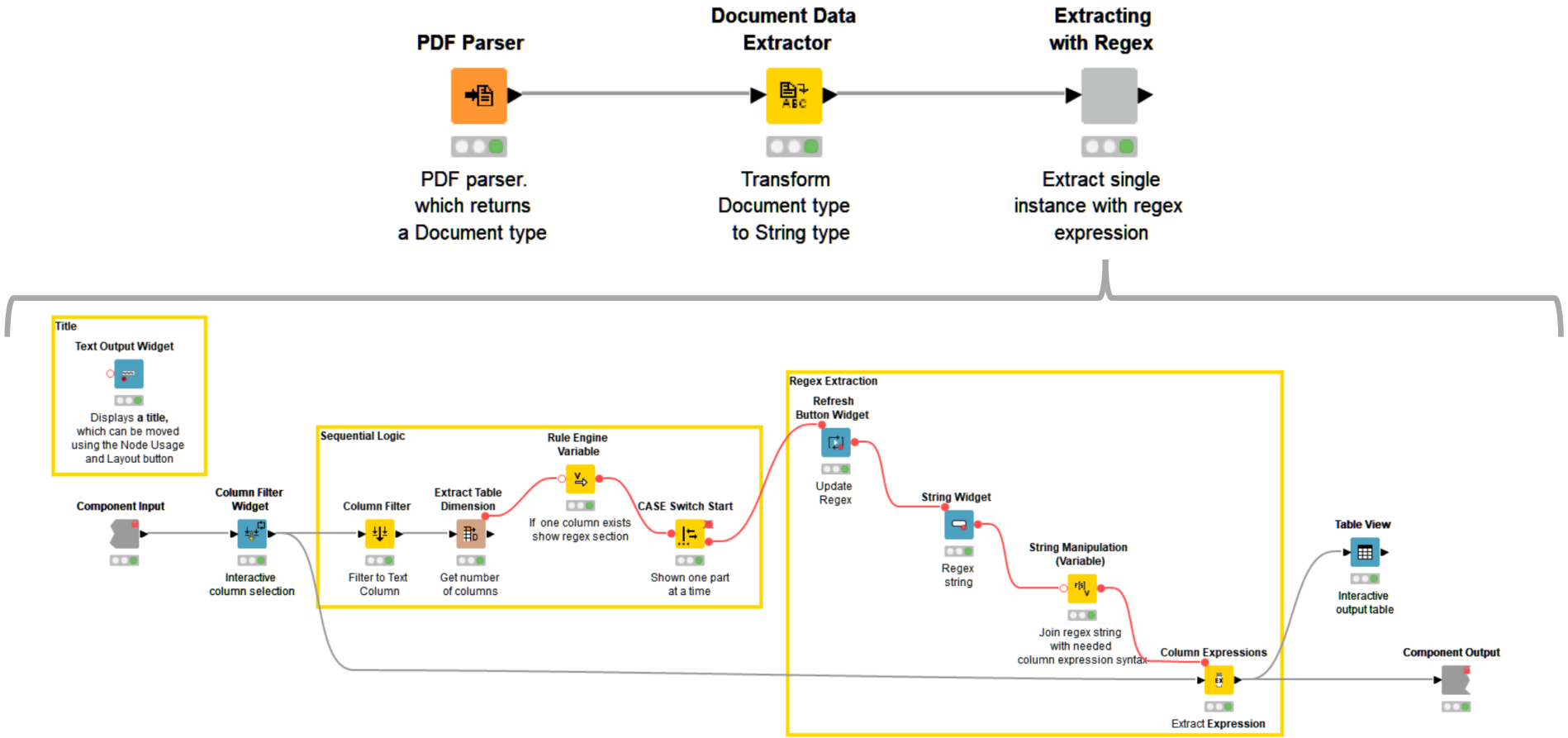
Check out solutions from the KNIME community as well

Data Extraction

At the end of this section, you will be able to:

1. Parse PDFs.
2. Recognize and reproduce Regex.
3. Manipulate a data extraction tool.

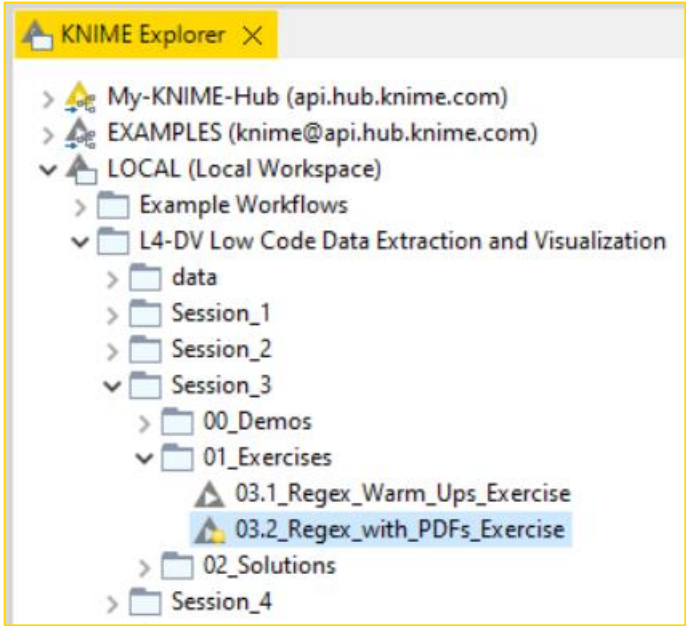
03: Parsing PDFs demo



Exercises: Section 3

1. Regex Warm Ups
   Using the data provided, extract information using regex.

2. Regex with PDFs
   Using the PDFs provided, parse each PDF and then extract the relevant information using regex.



Why create a parsing tool?

- Convenience
  - Consolidate PDF extraction and common Regex techniques in 1 place
- Hide complex steps such as:
  - Extracting data from the PDF
  - Using the Column Expression node without the extra syntax
  - Making the process reusable
- Go from PDF input to target text analysis quickly and efficiently

Summary of Section 3

Now you should be able to:

1. Parse PDFs.

2. Recognize and reproduce Regex.

3. Manipulate a data extraction tool.

Open for Innovation
KNIME

[L4-DV] Low Code Data Extraction and Visualization

Section 4

Data Quality & Visualization Best Practices

At the end of this section, you will be able to:

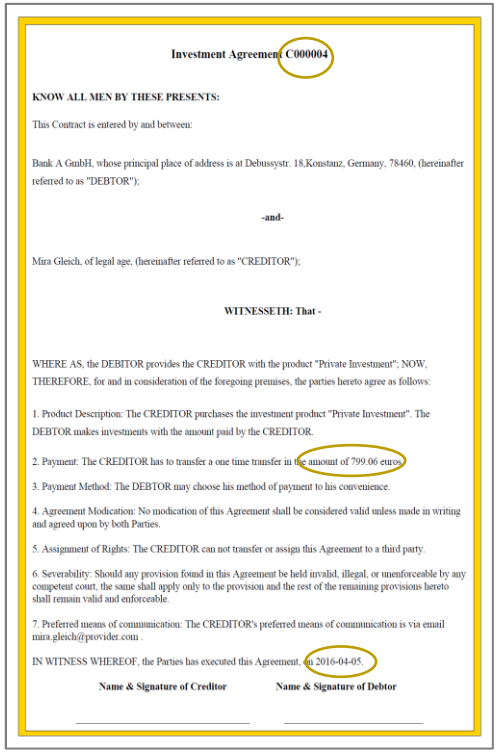
1. Assess data quality via outlier detection

2. Apply best practices for data visualization

Data quality via outlier detection

Detecting wrong (fake?) contracts

- Detect exceptional IDs, prices, or dates in large databases



Investment Agreement C000004

KNOW ALL MEN BY THESE PRESENTS:

This Contract is entered by and between:

Bank A GmbH, whose principal place of address is at Debussystr. 18,Konstanz, Germany, 78460, (hereinafter referred to as "DEBTOR");

-and-

Mira Gleich, of legal age, (hereinafter referred to as "CREDITOR");

WITNESSETH: That -

WHERE AS, the DEBITOR provides the CREDITOR with the product "Private Investment"; NOW, THEREFORE, for and in consideration of the foregoing premises, the parties hereto agree as follows:

1. Product Description: The CREDITOR purchases the investment product "Private Investment". The DEBTOR makes investments with the amount paid by the CREDITOR.

2. Payment: The CREDITOR has to transfer a one time transfer in the amount of 799.06 euros.

3. Payment Method: The DEBTOR may choose his method of payment to his convenience.

4. Agreement Modication: No modication of this Agreement shall be considered valid unless made in writing and agreed upon by both Parties.

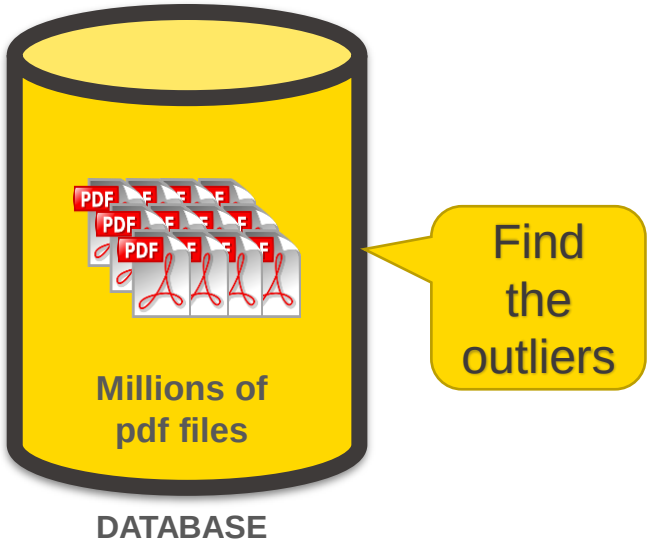
5. Assignment of Rights: The CREDITOR can not transfer or assign this Agreement to a third party.

6. Severability: Should any provision found in this Agreement be held invalid, illegal, or unenforceable by any competent court, the same shall apply only to the provision and the rest of the remaining provisions hereto shall remain valid and enforceable.

7. Preferred means of communication: The CREDITOR's preferred means of communication is via email mira.gleich@provider.com .

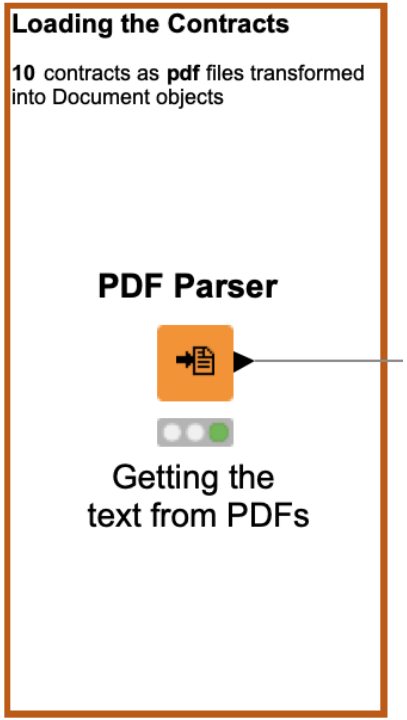
IN WITNESS WHEREOF, the Parties has executed this Agreement, on 2016-04-05.

Name & Signature of Creditor **Name & Signature of Debtor**



The Data

- From contracts extract all relevant data into a table



Row ID	S contract_id	31 date	D payment
Row1	C000002	2016-01-15	860.4
Row9	C000006	2016-01-20	1,647.7
Row8	C000001	2016-02-11	1,996.55
Row0	C000003	2016-02-26	1,184.51
Row3	C000005	2016-03-23	1,180.31
Row4	C000004	2016-04-05	799.06
Row7	C000000	2016-08-06	1,146.16
Row2	C000008	2016-10-03	912.72
Row5	C000007	2016-11-11	690.21
Row6	C000009	2016-12-03	475.94

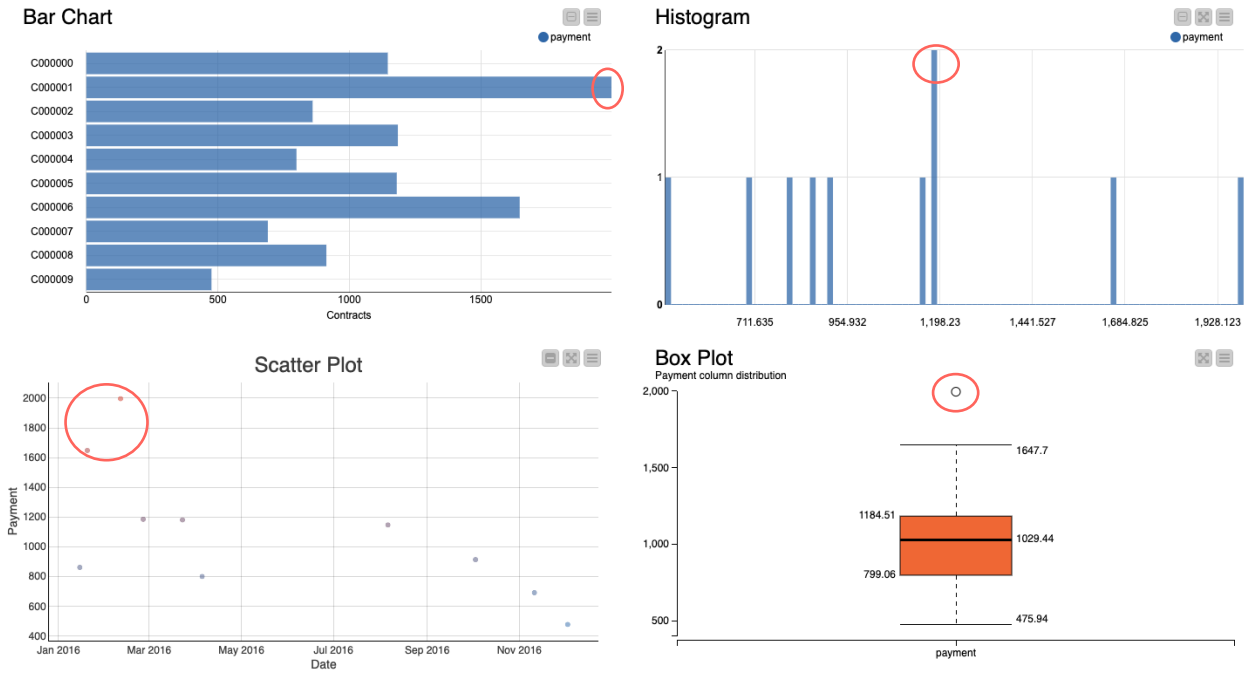
3 Techniques for Outlier Detection

- Visualizations
- Statistics
- Machine Learning

Visual outlier detection - comparative dashboard

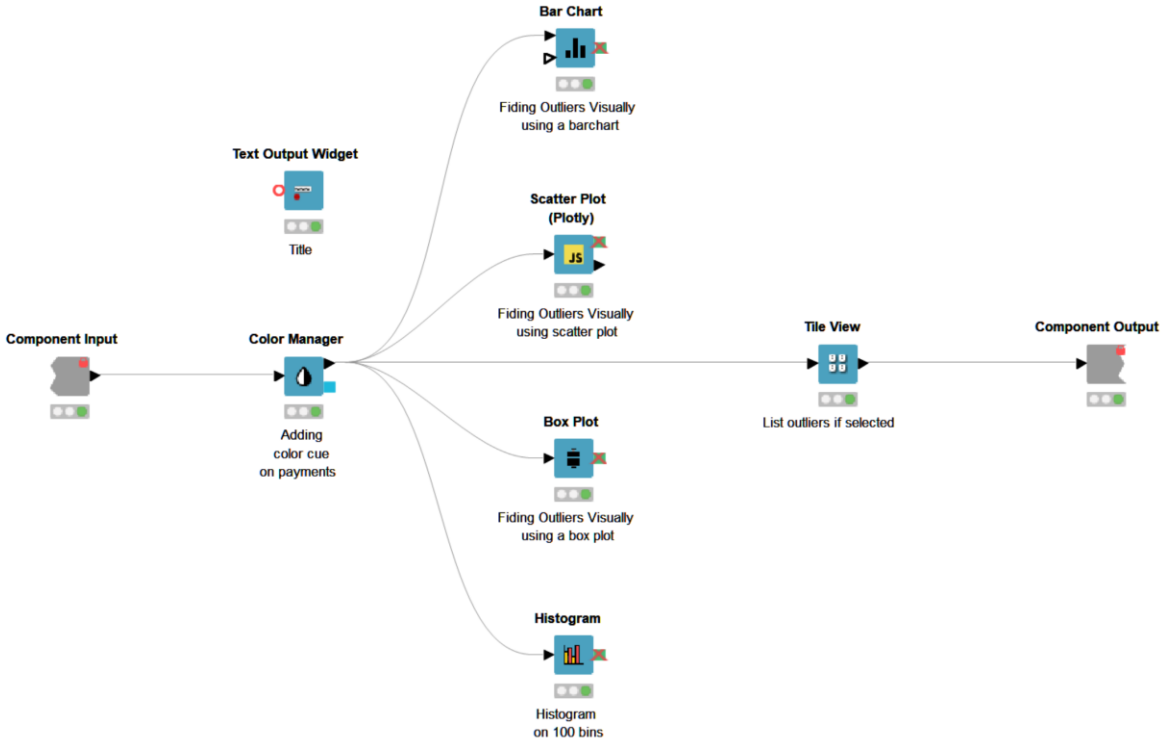
Bar chart, histogram, scatter plot, & box plot

Outlier Detection via Visual Exploration



Creating the visual dashboard

- Simply placing view nodes into a component lets us visualize them as a dashboard



Should you use visualization techniques?

Pros:

- Aesthetically pleasing
- Outliers are easy to spot and interpret

Cons:

- Can only investigate 1~2 dimensions at a time
- Manual checking
- Manual selection for further exploration

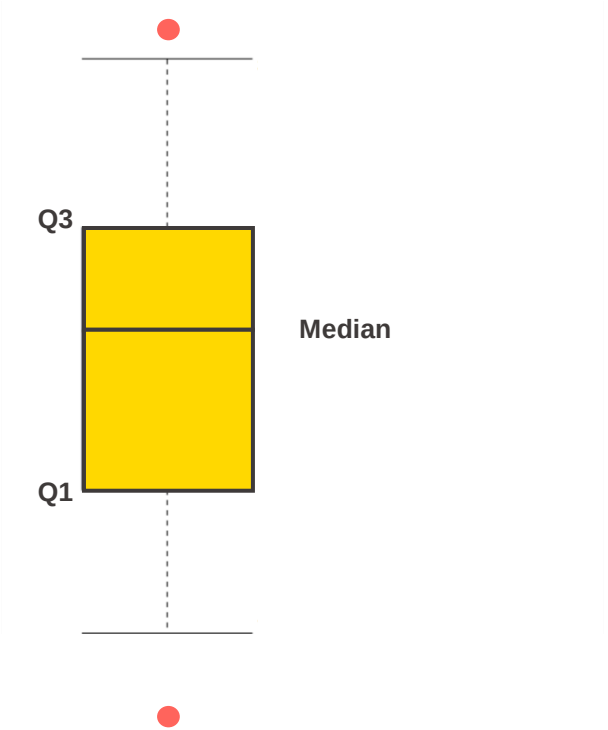
Is it possible to automate this process?

3 Techniques for Outlier Detection

- Visualizations
- Statistics
- Machine Learning

Finding Outliers: Statistics

IQR Technique



1. Find median of data to split data into a lower and upper half.

2. Find median of upper half of data (Q3)

3. Find median of lower half of data (Q1)

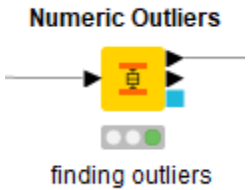
4. Calculate IQR and choose k.
   IQR = Q3 − Q1
   k = 1.5 (usually)

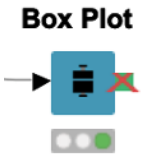
5. Point is an outlier if:
   Point > Q3 + k * IQR
   or
   Point < Q1 − k * IQR

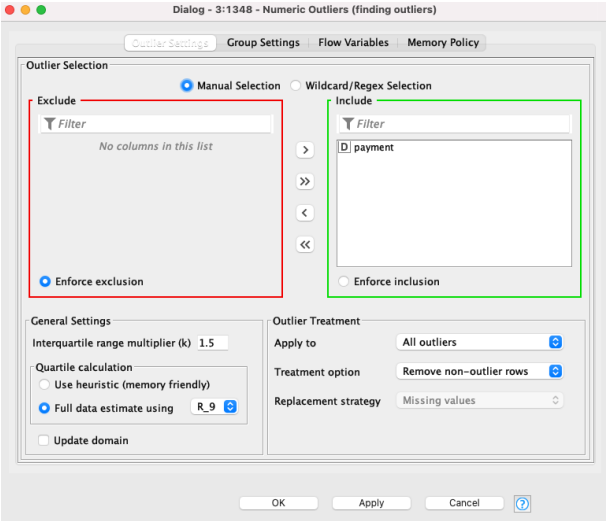
Finding Outliers: Statistics

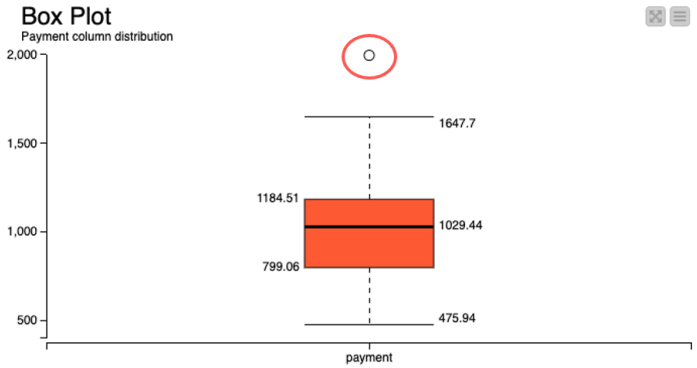
IQR Technique

The Numeric Outliers node implements (and the Box Plot node visualizes) the IQR technique.



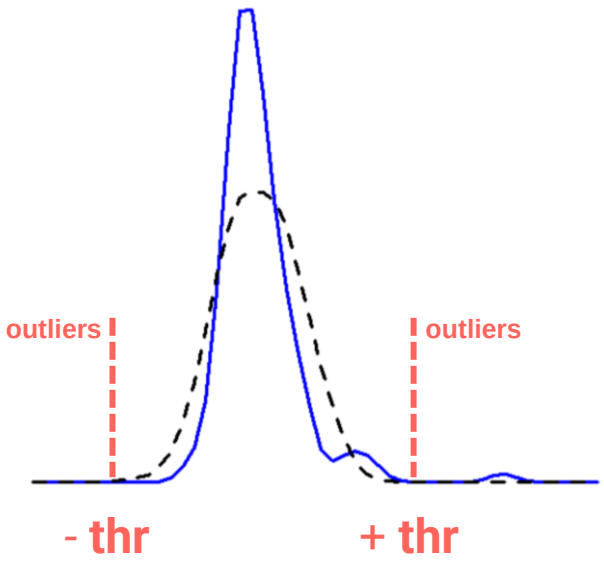






Finding Outliers: Statistics

Z-score Technique



1. Convert points to their z-scores:

$$z = \frac{x - \mu}{\sigma}$$

2. Define a threshold (**thr**).
   Often thr = 2.5 or greater

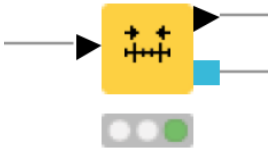
3. z is an outlier if:
   |z| > **thr**

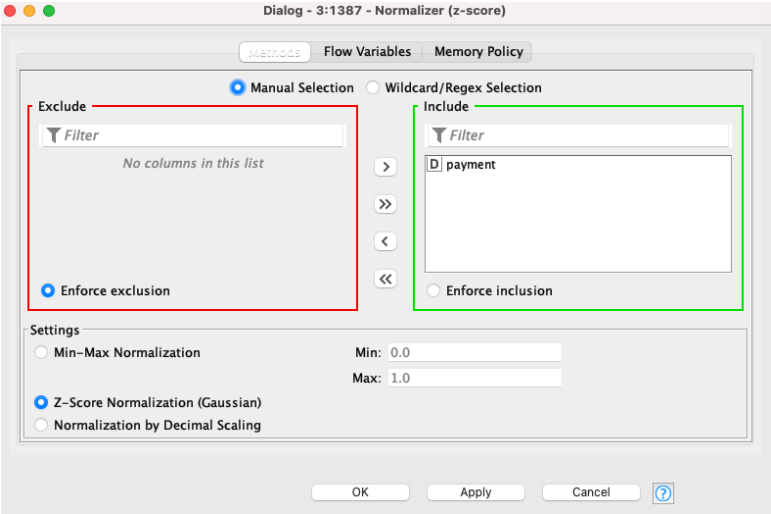
Finding Outliers: Statistics

Z-score Technique

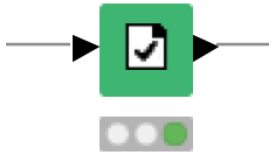
Z-score normalization and Rule Engine can implement the z-score technique

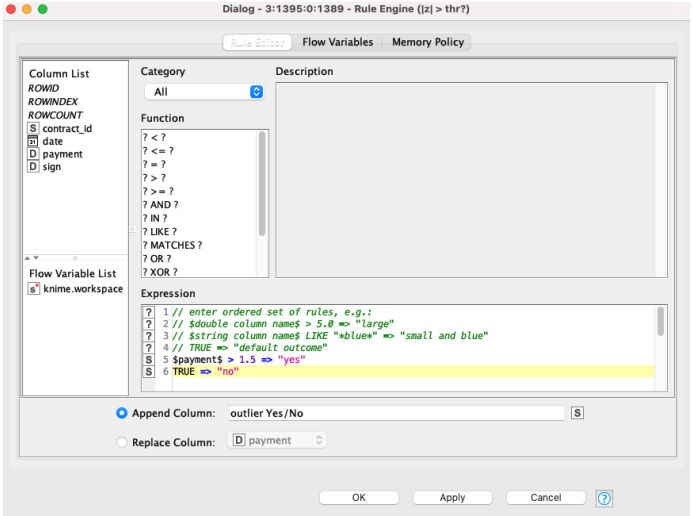
Normalizer





Rule Engine





KNIME Knowledge Check 01

- True or False.
  For the statistics methods we showed, it is possible to find outliers across 2 dimensions.

  For example, a data point that is an outlier because of a height and age combination.

Should you use statistical techniques?

Pros:

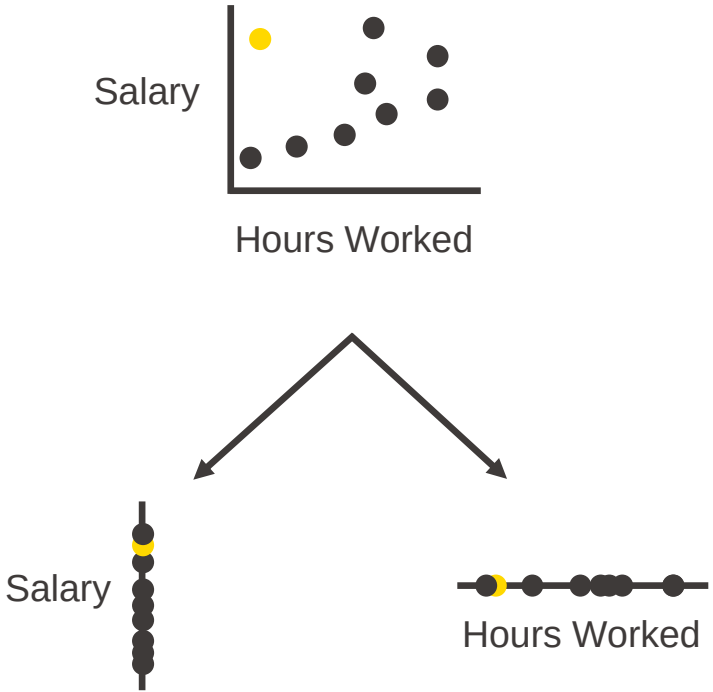
- Extraction automation
- Easy to understand
- Just 1 or 2 nodes

Cons:

- You might need to know some statistics
- Unidimensional analysis
- Z-score technique: the data is forced into a Gaussian distribution

What about multidimensional analysis?

One feature or many dimensions?



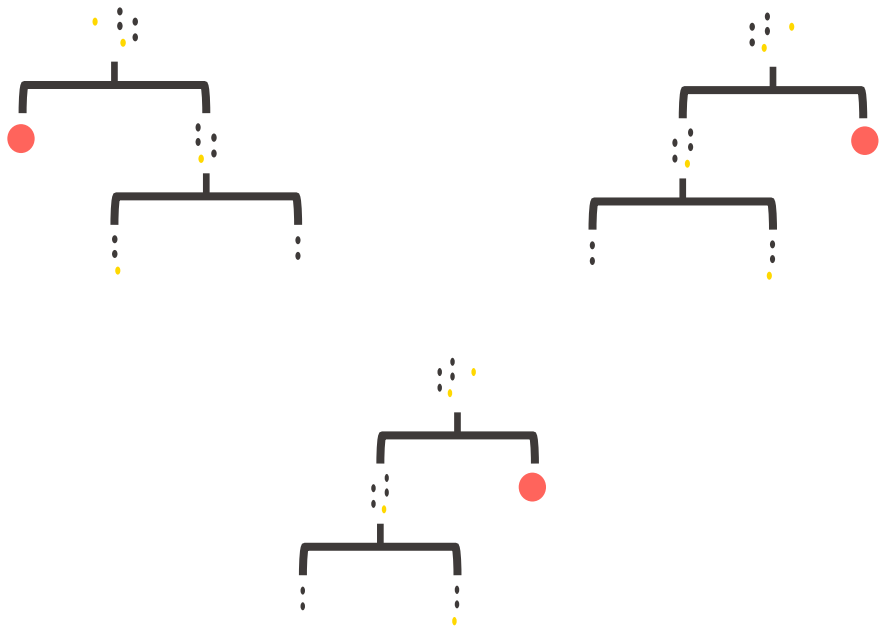
- In unidimensional space, spotting outliers may not be possible.
- We recommend beginning with multidimensional outlier detection first, and then run outlier detection procedures for each dimension.

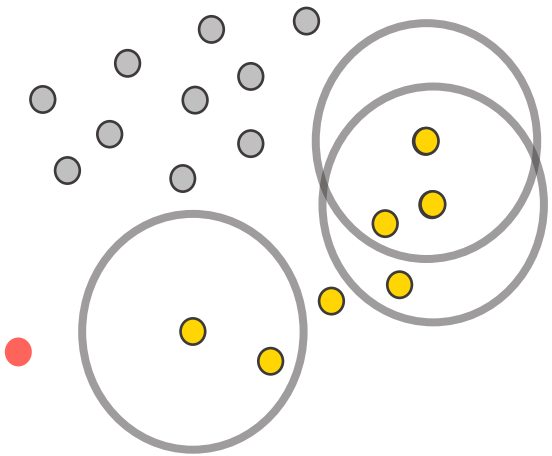
3 Techniques for Anomaly Detection

- Visualizations
- Statistics
- Machine Learning

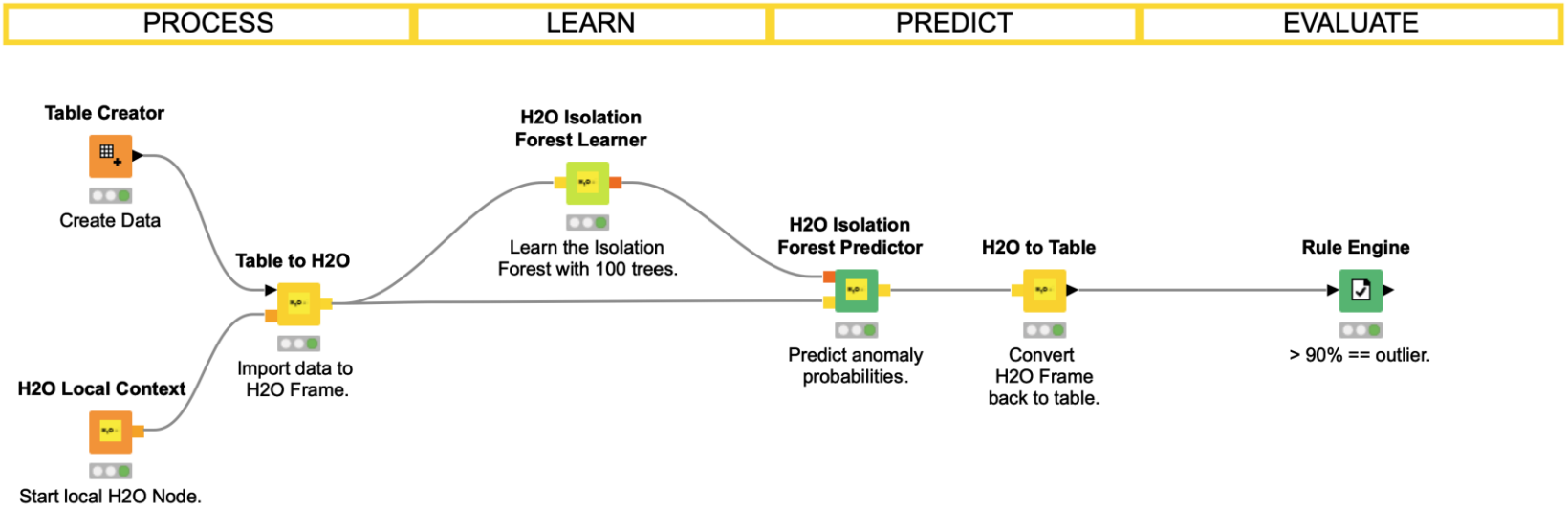
Two Machine Learning Algorithms for Outlier Detection

- Isolation Forest – a tree-based algorithm
- DBSCAN – a clustering-based algorithm



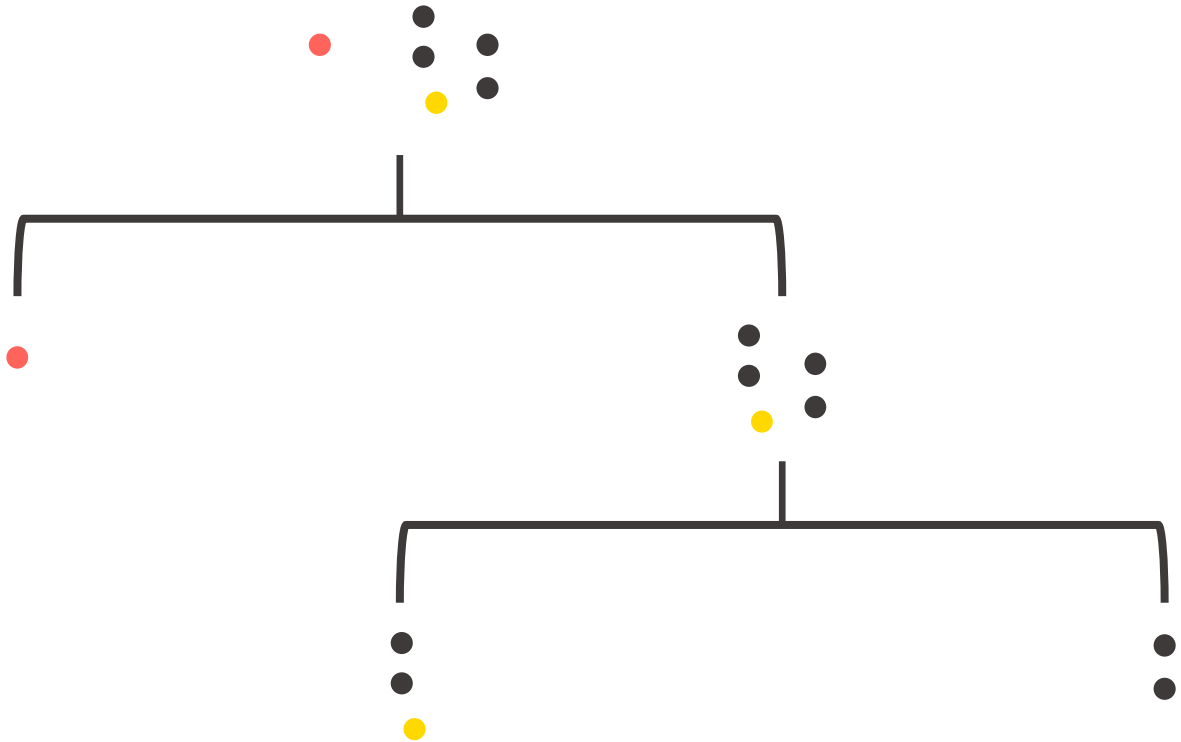


Isolation Forest in KNIME



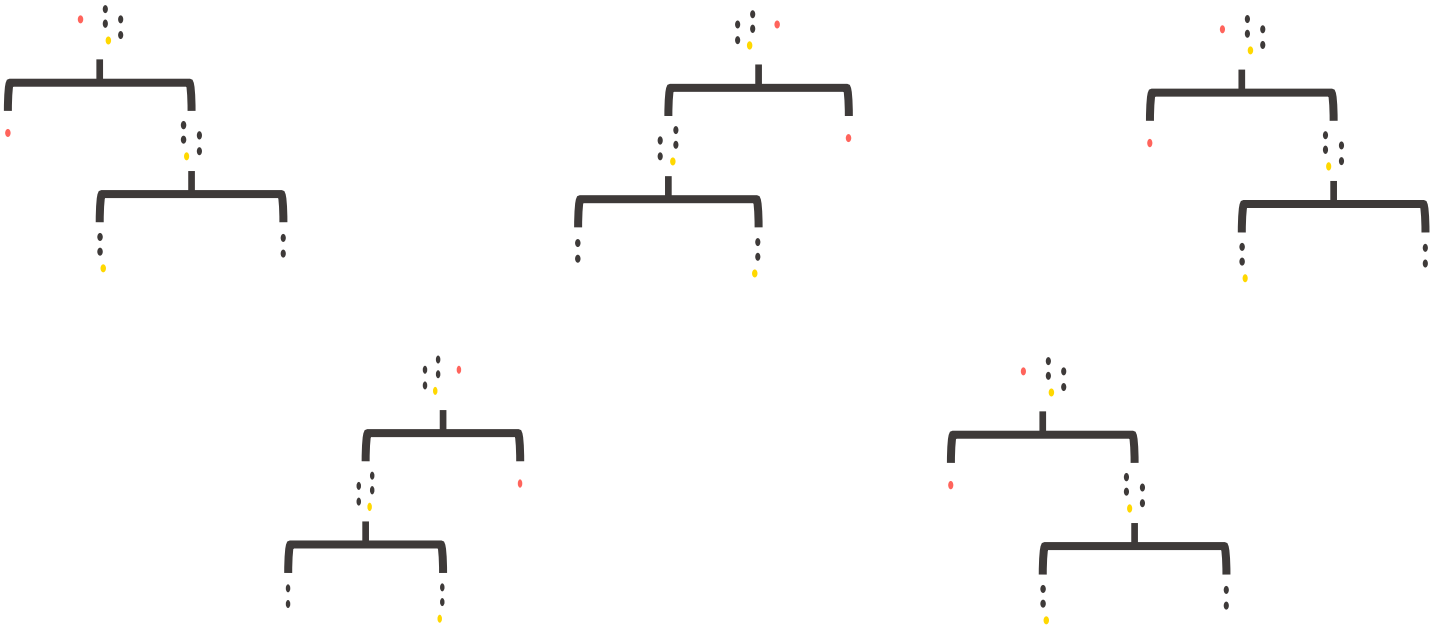
Isolation Tree

How many splits did it take for a point to be isolated?
Fewer implies outlier.



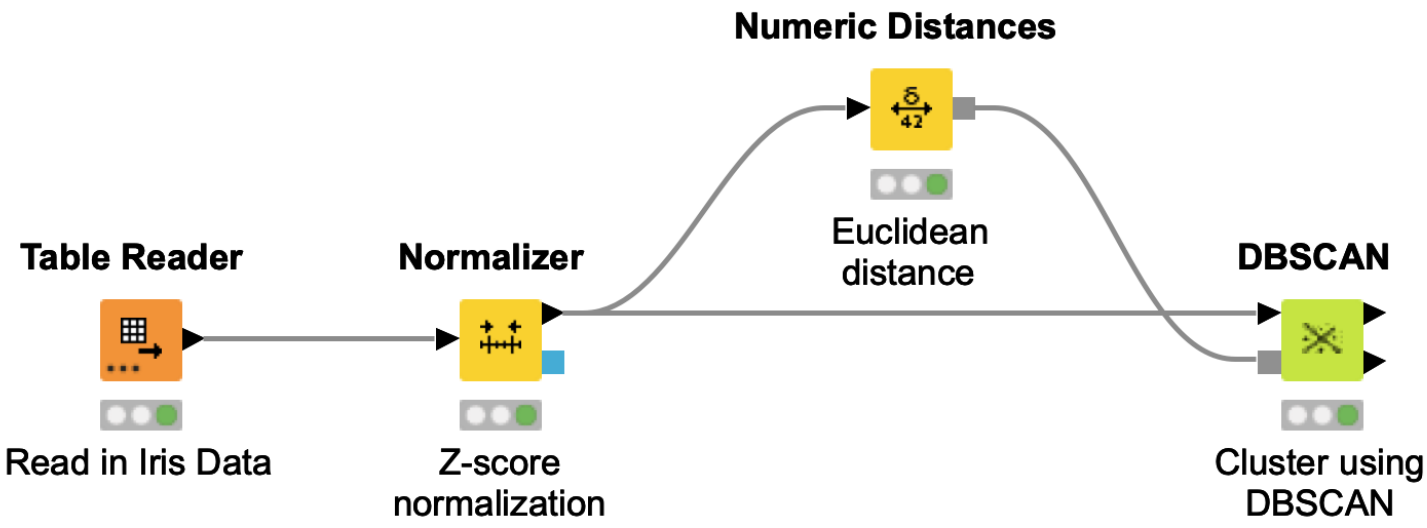
Isolation Forest

In a forest, use many trees and a voting mechanism to detect outliers.



DBSCAN in KNIME

- DBSCAN is a 4-node process in KNIME.

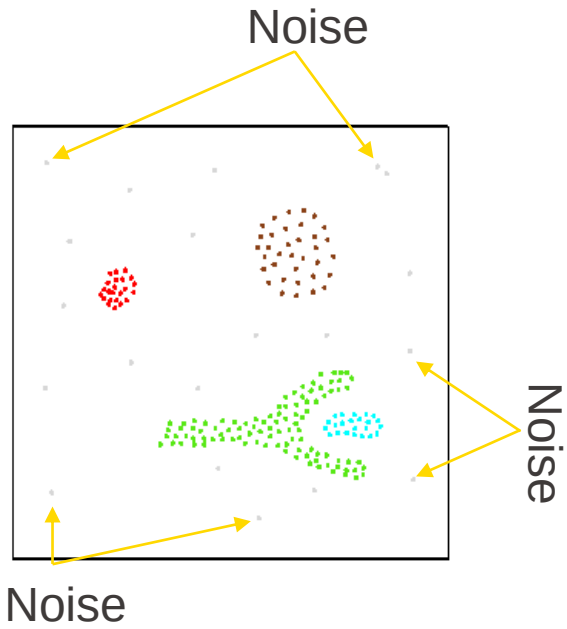


DBSCAN

DBSCAN - defines 5 types of points in a dataset but we'll talk about them as 2.

- **Cluster Points** within a specified distance (ε).
- **Noise Points** farther than ε from other points

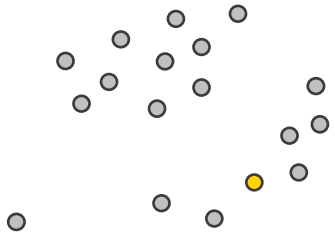
Clusters are built by joining points to one another.



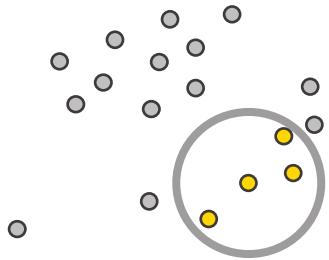
DBSCAN

- For each point is it in an ε-environment? If not, **then it's noise.**

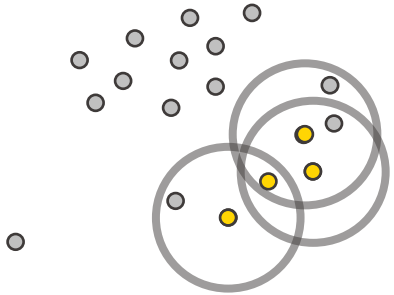
1 Pick random point.



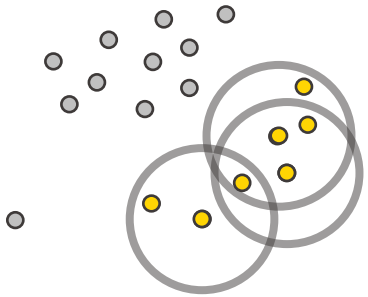
2 Emit ε-environment.



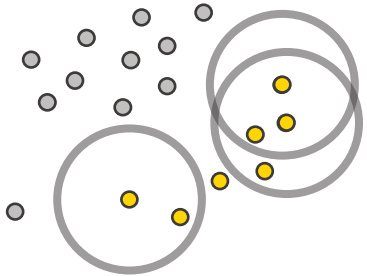
3 Emit ε-environment.



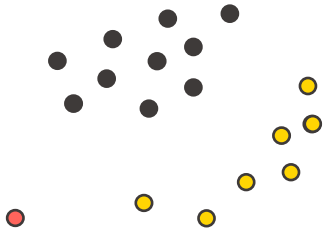
4 Continue…



5 …until no neighbors.

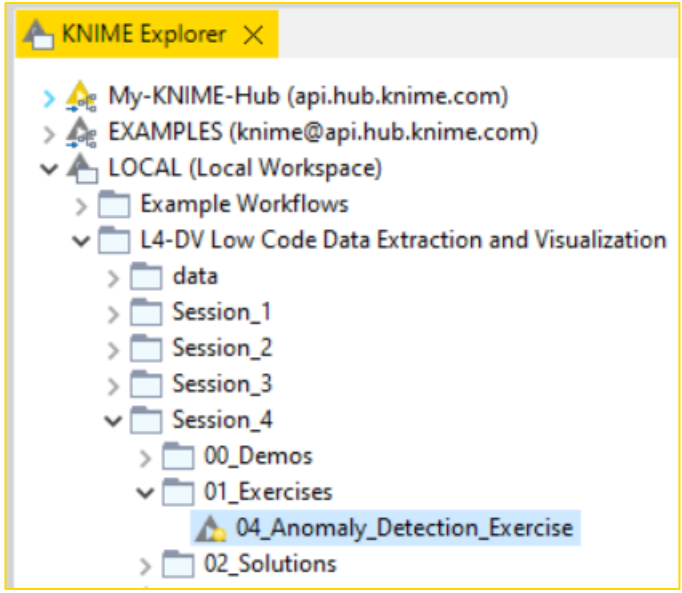


6 Rinse and repeat.



Exercises: Section 4

1. Anomaly Detection
   Using the data provided, use various methods to detect outliers and compare the methods as well as output.



KNIME Knowledge Check 02

- True or False.
  For the machine learning methods we showed, it is possible to find outliers across 2 or more dimensions.

  For example, a data point that is an outlier because of a height and age and weight combination.

Should you use machine learning?

Pros:

- Extraction of outliers is automated
- Multidimensional analysis possible
- Isolation Forest: Non-parametric (no distribution required) & Fast
- DBSCAN: Easy to implement (4 nodes)

Cons:

- DBSCAN: Can be slow
- Hyperparameter setting required

Data Quality & Visualization Best Practices

At the end of this section, you will be able to:

1. Assess data quality via outlier detection

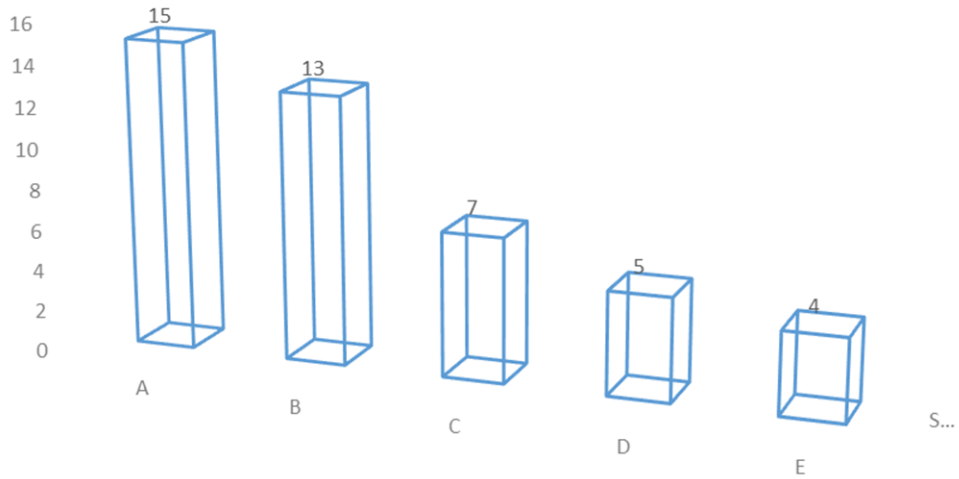
2. Apply best practices for data visualization

Visualization Principles

1. Simplification
2. Color
3. Purpose

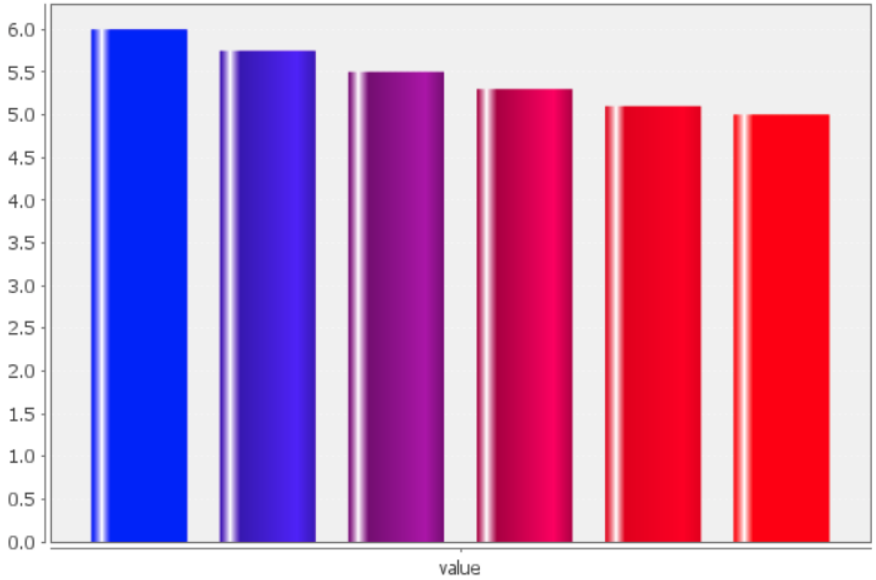
Simplification – be mindful of the cognitive load

Avoid 3D charts; Use 2D charts instead.



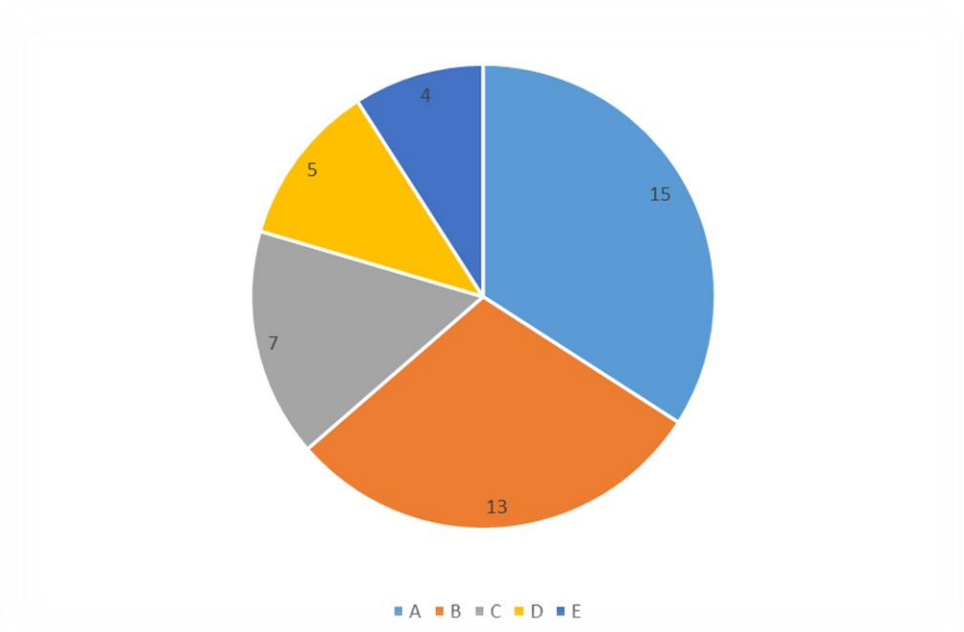
Simplification – remove extra effects

Remove extra effects like the shading on each of these bars.



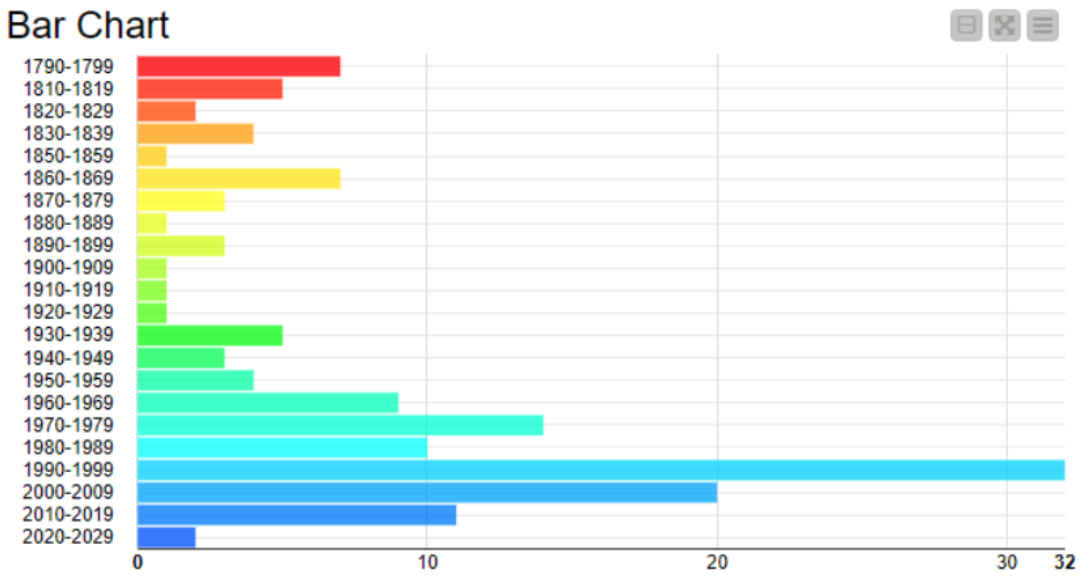
Simplification – choose charts with minimal processing

A pie chart with more than 2 categories is usually harder to understand immediately than a bar chart.



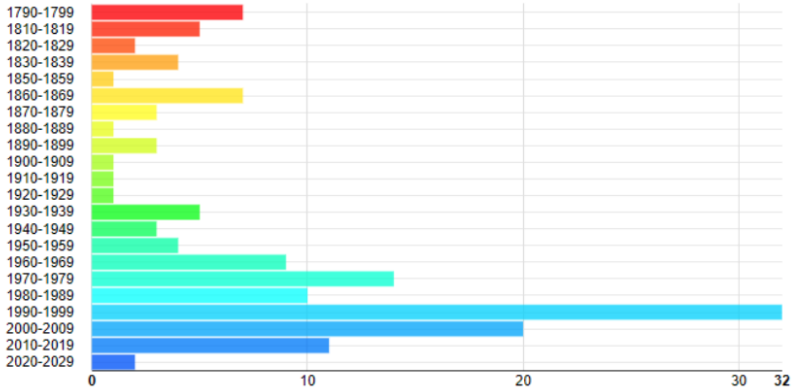
Color – make sure every color is needed

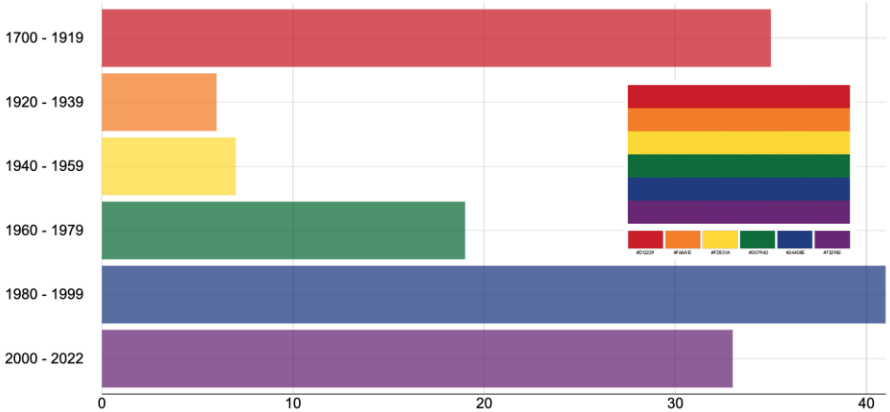
- Use color with intention; be able to explain each color choice or remove it



Color – reduce total colors used

- Image on left has too many colors



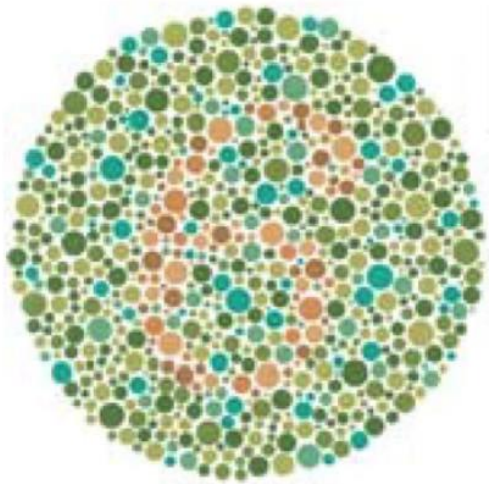


- Bin the years in a different way to show less but have more impact.

Color – be mindful of the colorblind

- Avoid Christmas colors (i.e., red and green) because the colorblind may not be able to distinguish them.

Typical Vision
(can see the red 6)



Colorblind Vision
(red and green are not separable)

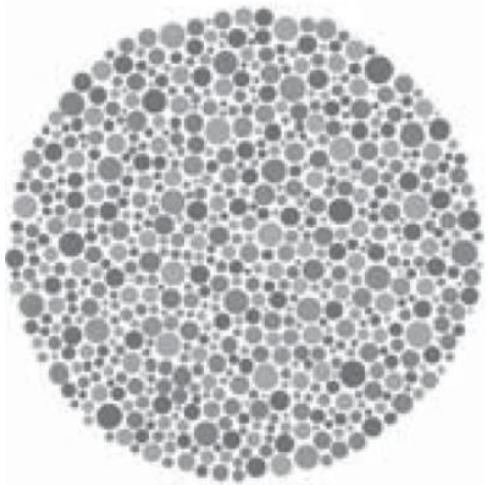
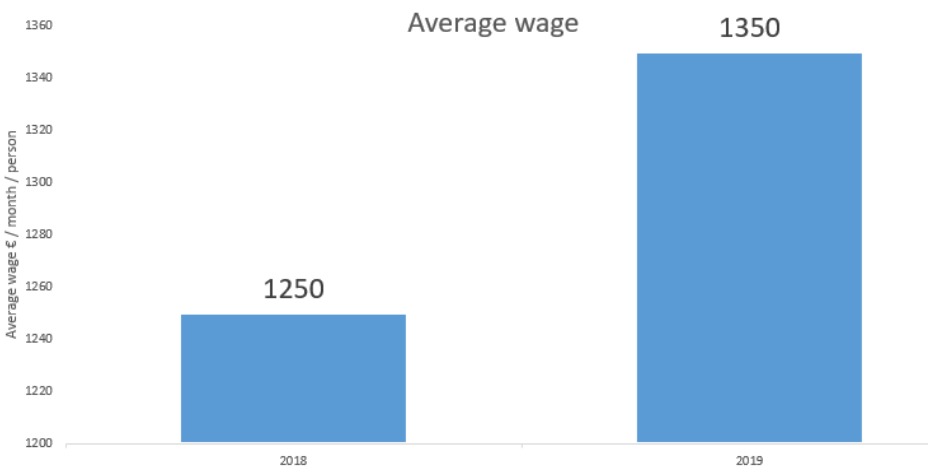
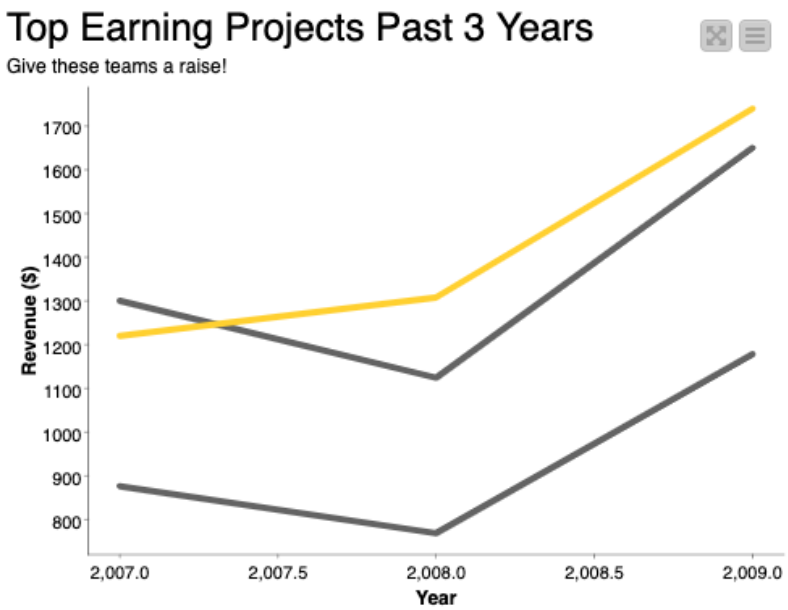


Image souce: article from Nature

Purpose – ask for a call to action

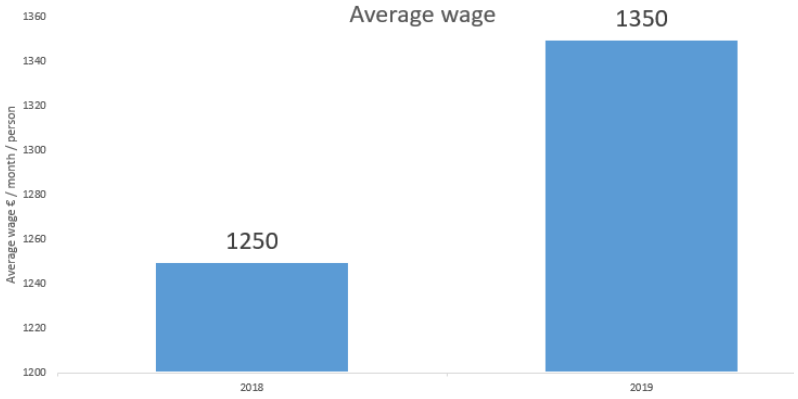
- Good visualizations usually ask us to do something.

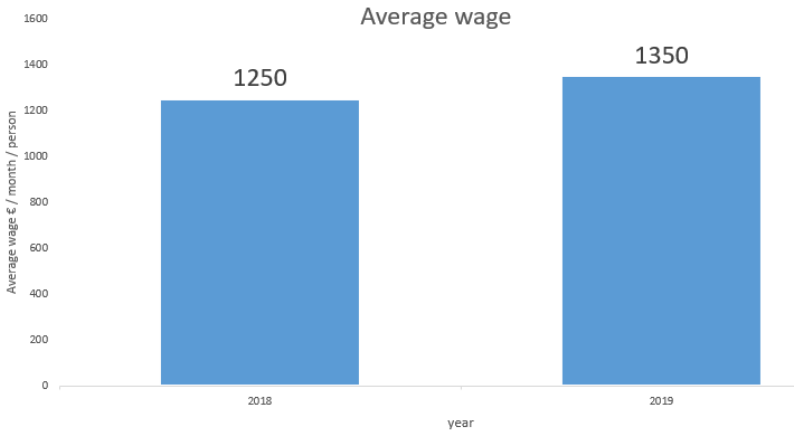




Purpose – do not purposefully mislead

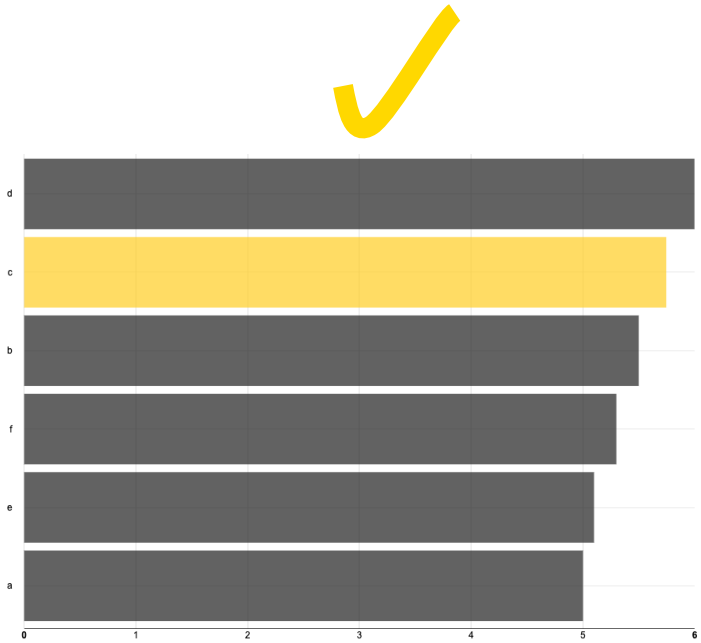
- Non-zero beginning on axis is deceptive





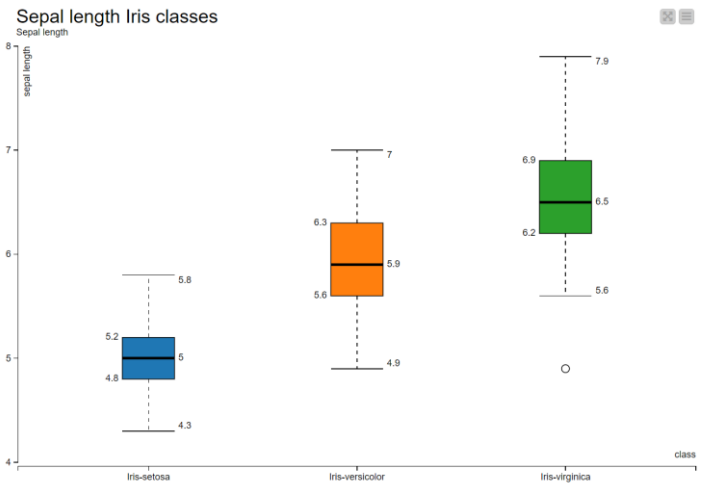
Purpose – true purpose is quick transfer of information

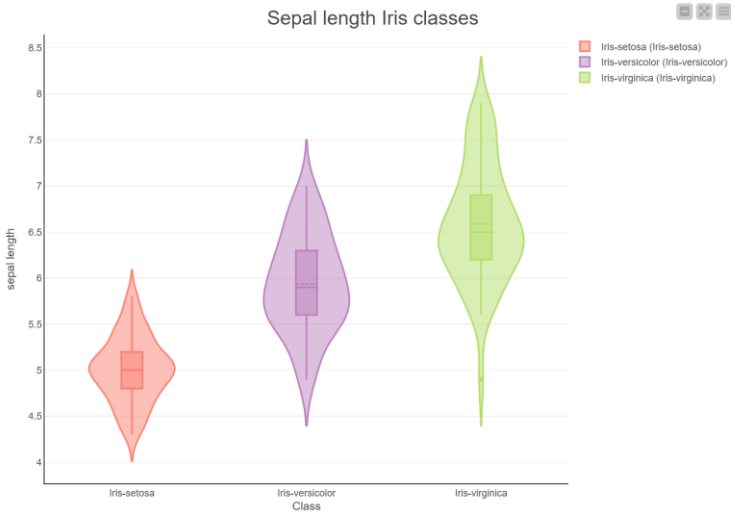
- Order your data to lower the cognitive load on the viewer



KNIME Knowledge Check 03

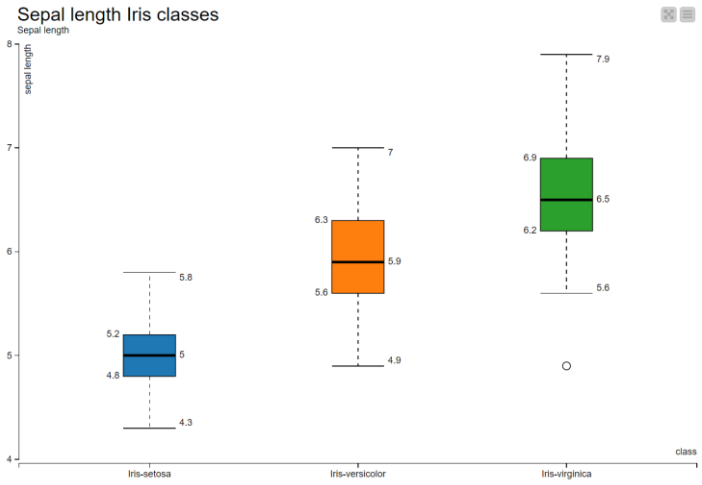
- From a visualization perspective, what advantage does the violin plot have over the box plot and vice-versa?





KNIME Knowledge Check 03: Solution

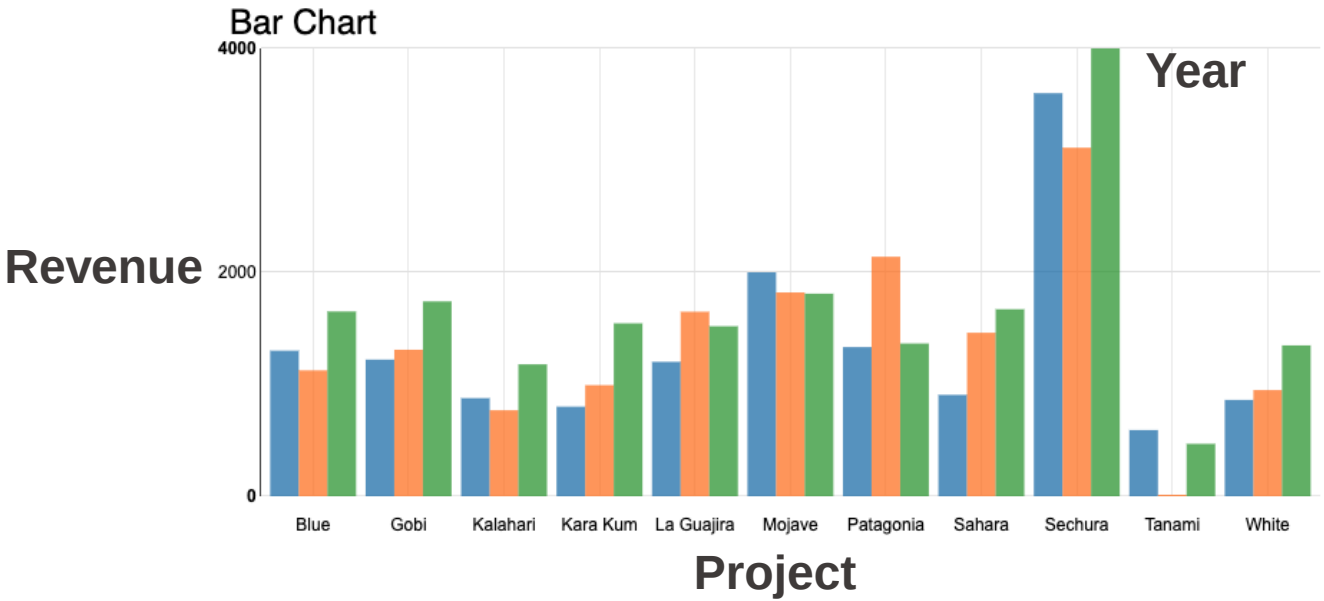
- Advantage of the Box Plot
  - Hides distribution (causing less cognitive overload)



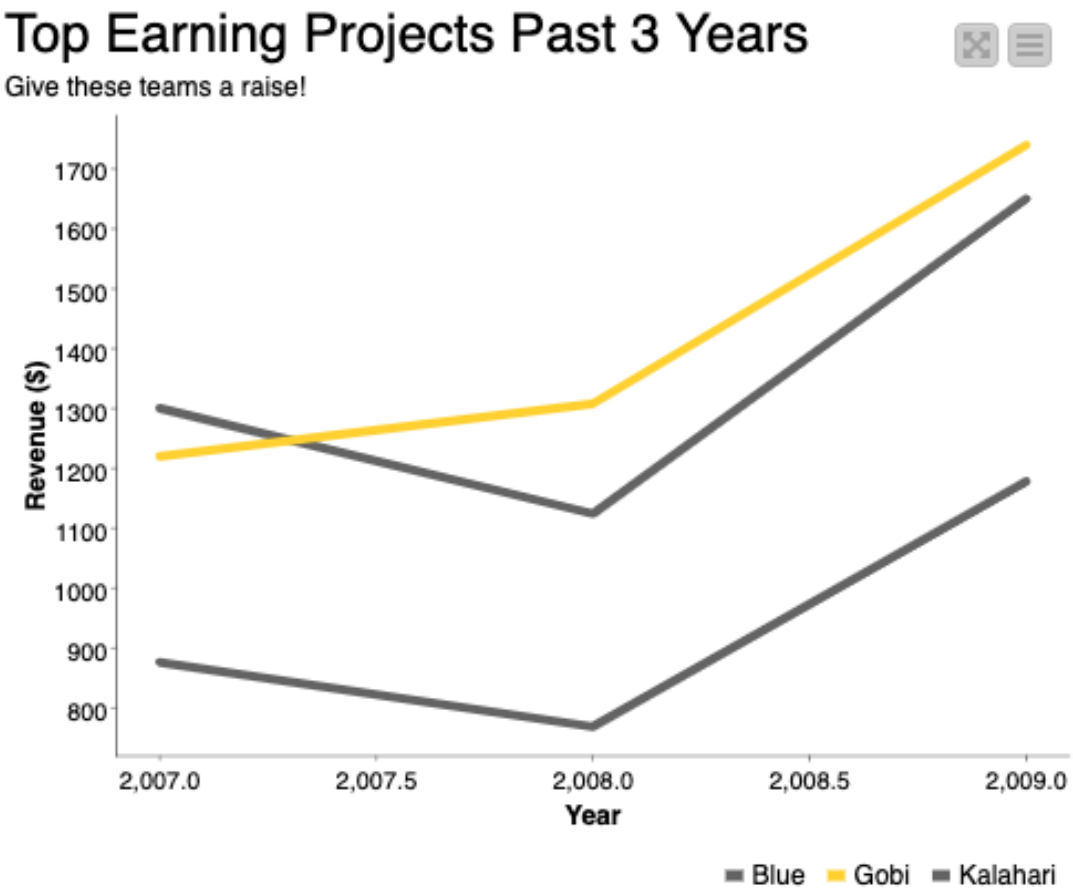
- Advantages of Violin Plot
  - Displays box plots and
  - Displays data distribution

KNIME Knowledge Check 04

- How can this visualization be improved?



KNIME Knowledge Check 04: Solution



Why is this better?

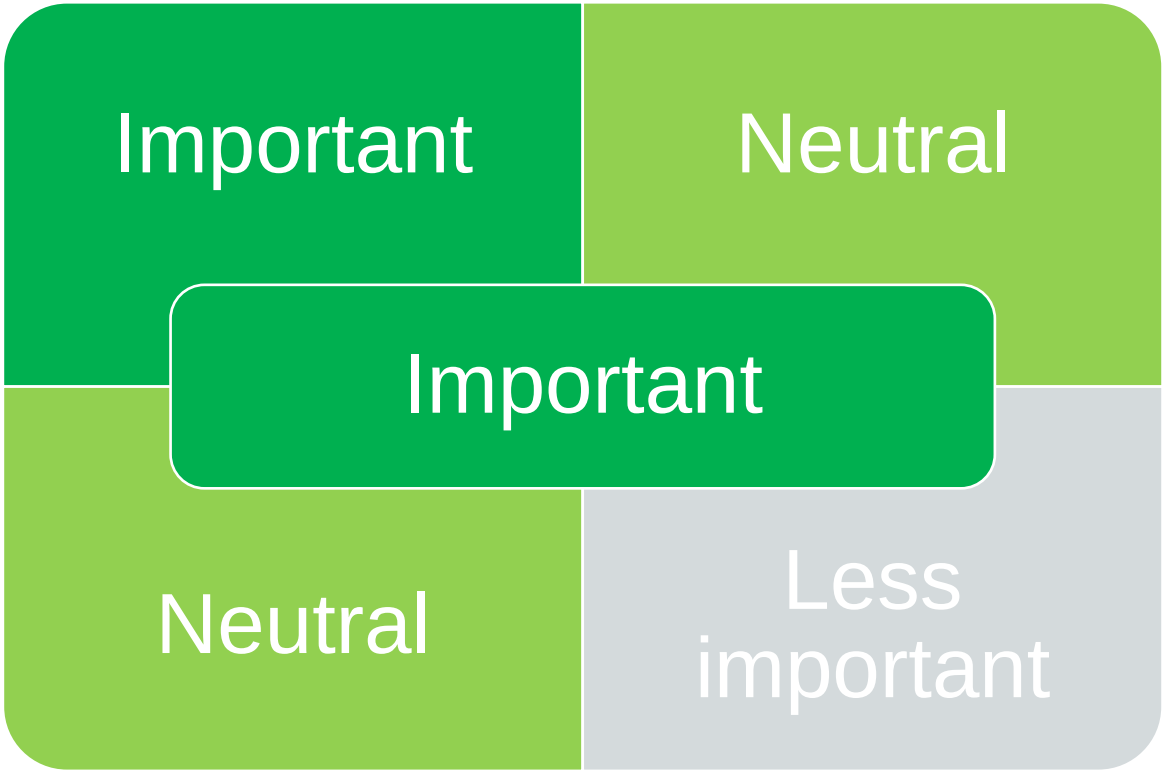
1. Simplified (less irrelevant data)
2. Color (highlighted one project for discussion)
3. Purpose (Advocating for raises)

Dashboard rules

1. Avoid excess detail
2. Make your dashboard fit on one page
3. Follow the standard dashboard distribution

Dashboard distribution

Placement of items in a dashboard is critical for rapid, natural comprehension.
Notice: key points on left-top and center.



Example 1: Support Tickets Dashboard

Exemplar dashboard distribution: Statistics on top, key plot on left with purpose, details on right

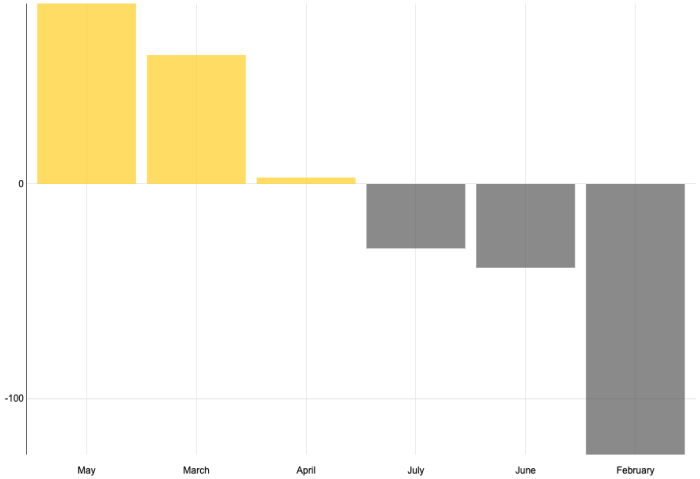
Support Tickets

Mean Duration Hours	Mean Duration Mins	Total Tickets	Closed tickets	Open tickets
7.7	461.88	10000	9987	13

Year Analysis

Ticket Count Change Comparing Prior Month

More debuggers needed March through May



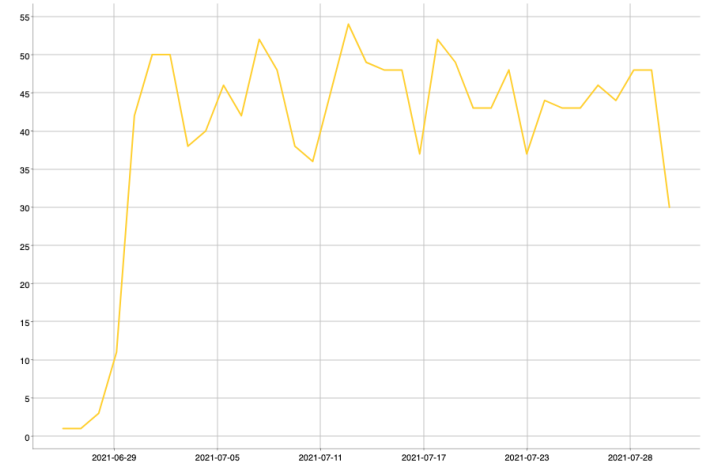
Select year:

2021

Month Analysis

Ticket counts

by date



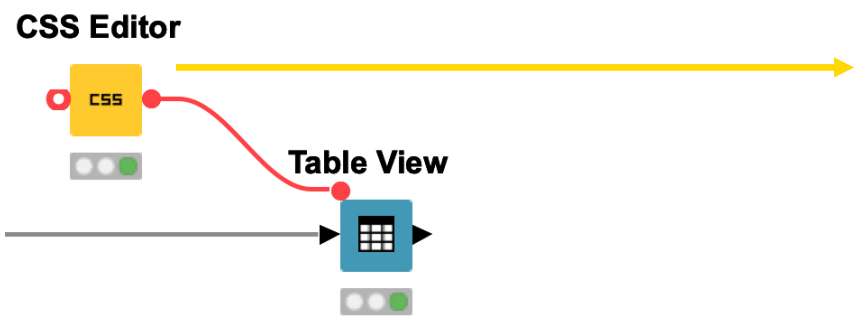
Select month:

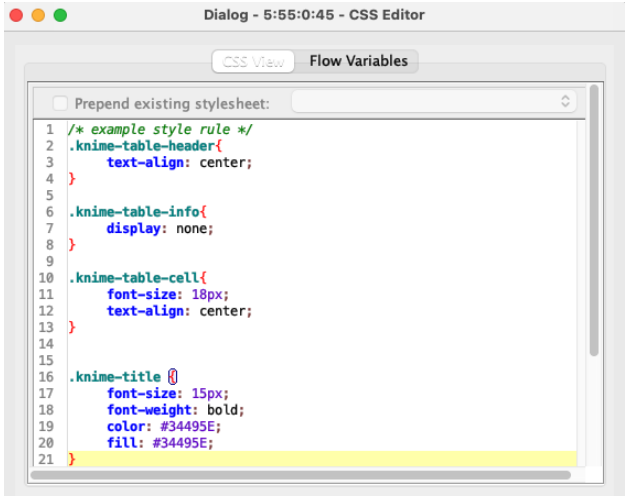
July

How was the topmost table made?

Support Tickets

Mean Duration Hours	Mean Duration Mins	Total Tickets	Closed tickets	Open tickets
7.7	461.88	10000	9987	13





Additional dashboard customization

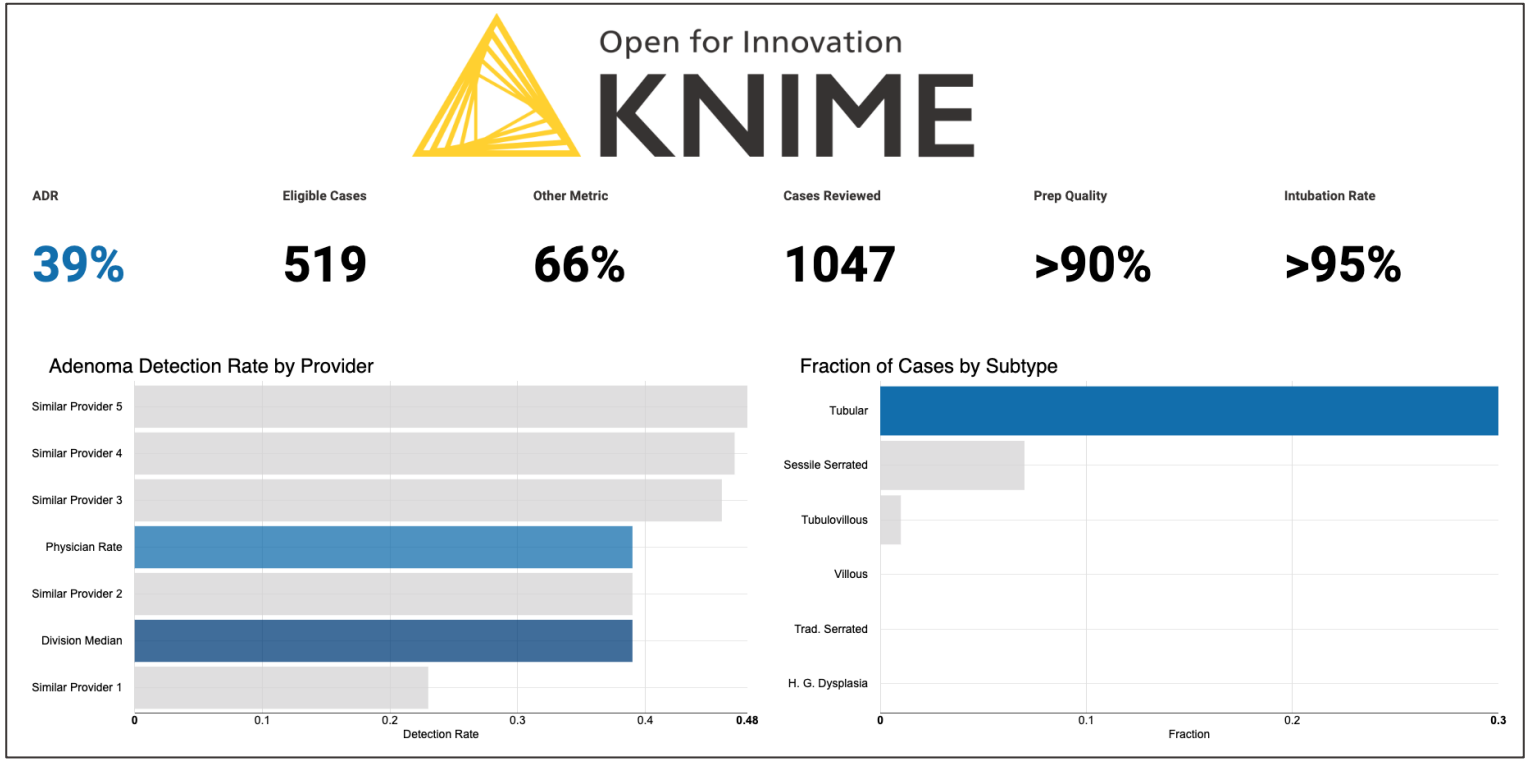
1. Add images
    - For instance, integrating a company logo

2. Modify HTML
    - To add more certain text bigger or add color

3. Change CSS
    - To change the size of font for certain views

Example 2: Medical Dashboard

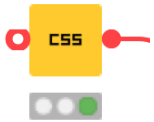
Additional customization: Adding images, modifying HTML, changing CSS



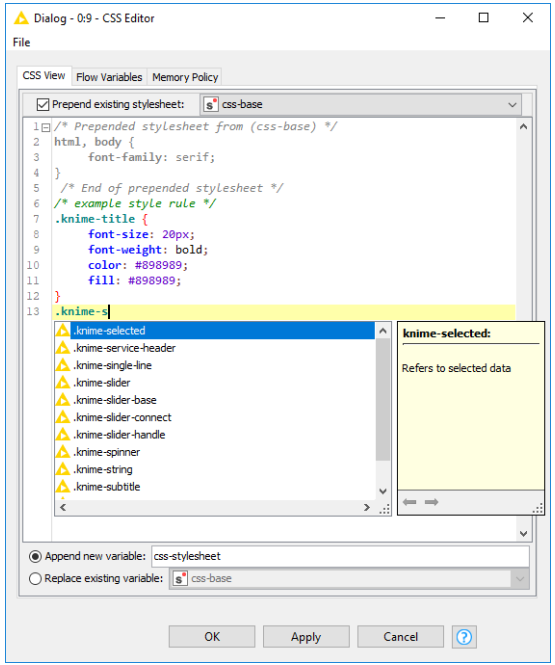
What is CSS?

- CSS, Cascading Style Sheets, is a language for presentation in HTML or XML.

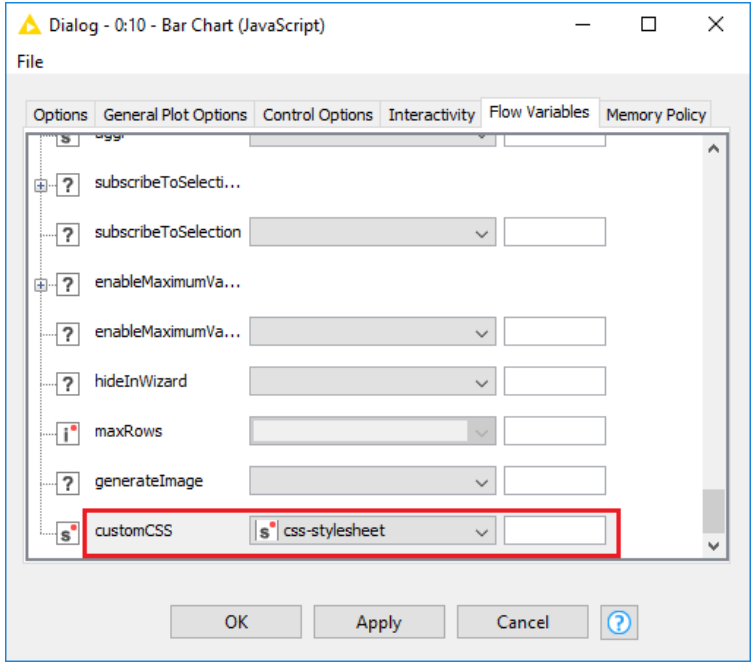
CSS Editor



Define font size, etc.



Apply your definitions

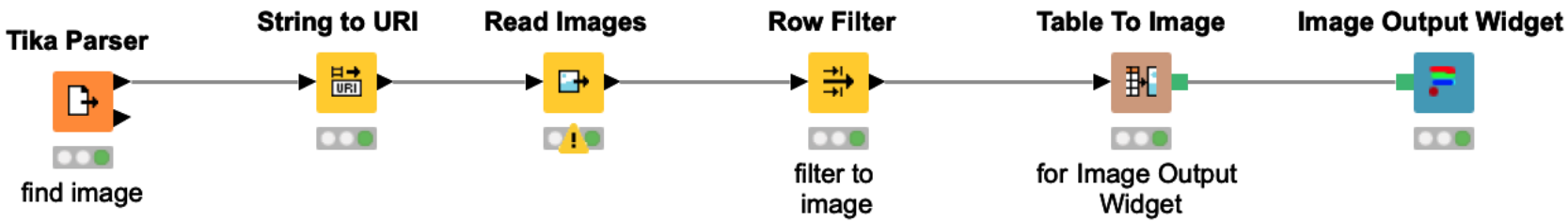


[See the CSS guide for more tips]

Additional customization – Adding an image

To show a logo like:

We can use this workflow (found in the demos folder):

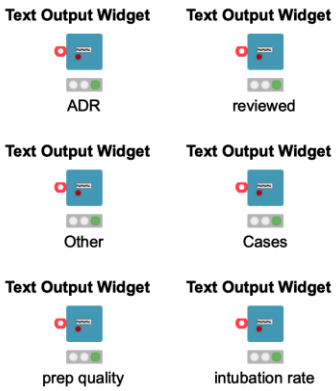


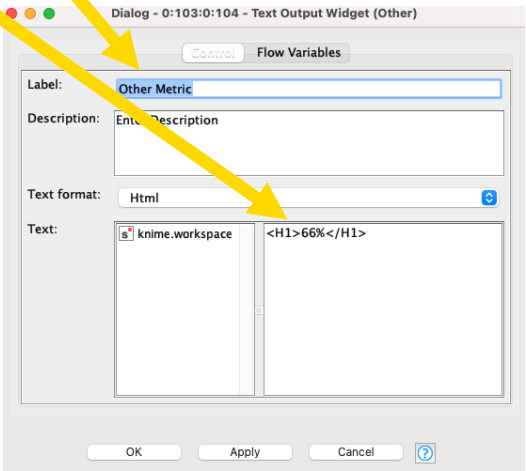
Additional customization – Modifying the HTML

Change the HTML to affect size and color of text in Text Output Widget

ADR	Eligible Cases	Other Metric	Cases Reviewed	Prep Quality	Intubation Rate
39%	519	66%	1047	>90%	>95%

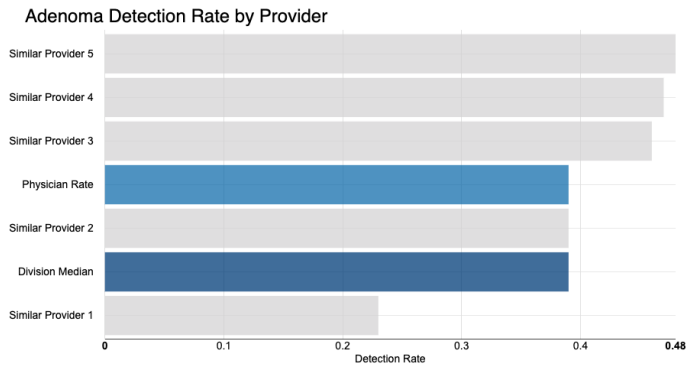
How to build

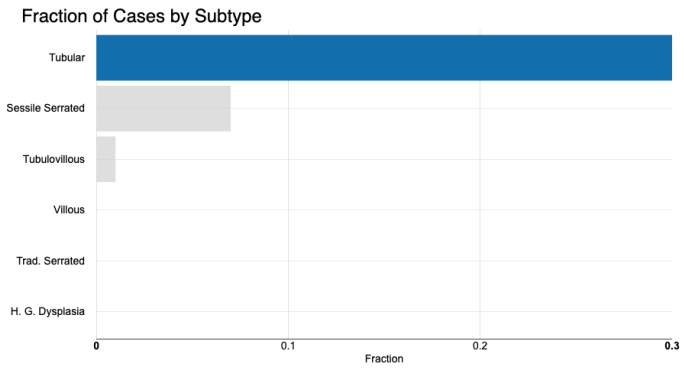




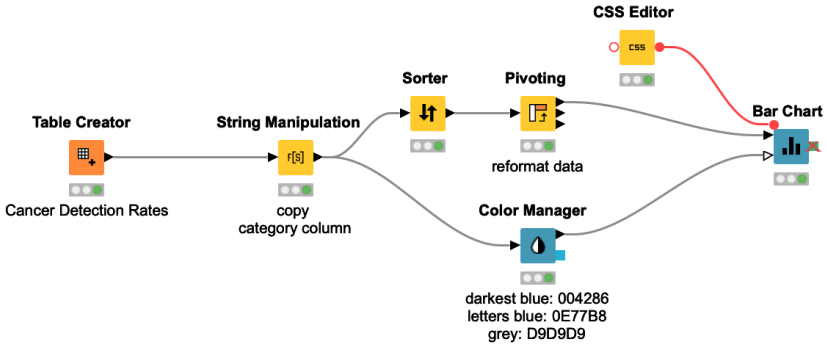
Additional customization – Changing the CSS

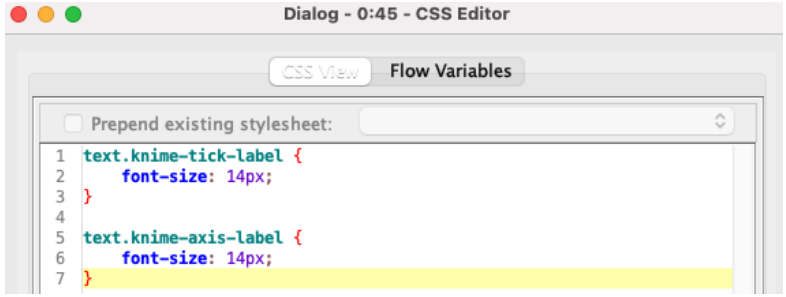
Change the CSS of the bar chart to have larger font





How to build





Data Visualization Summary

- Simplify
  - Show only relevant information
  - Avoid a high cognitive load
- Emphasize
  - Significant elements only
  - Use color with intent
- Select the right visualization type
  - Based on the type (category, number, etc.)
  - And on the number of variables (1, 2, or 3)

Summary of Section 4

Now you should be able to:

1. Assess data quality via outlier detection

2. Implement best practices for data visualization

Course objectives

With the completion of this course, you should now be able to:

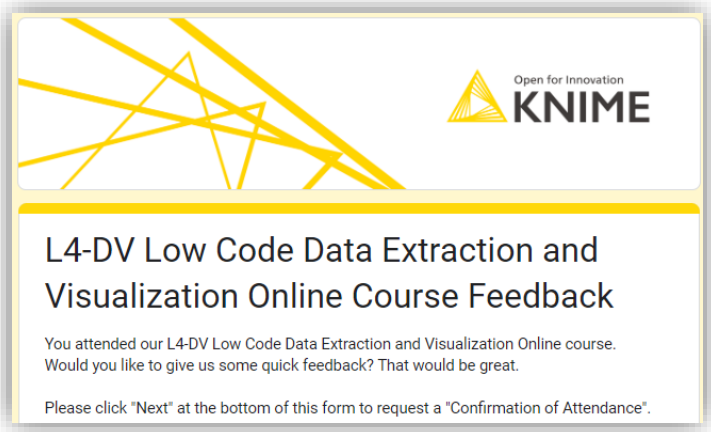
- Collect data via **REST APIs**, **web text scraping**, and an **interactive** data collection **tool**
- Explore and visualize data
- Extract data and images from **PDF** documents
- Write regular expressions (**regex**)
- Identify and correct errors in data via **outlier detection**
- Build effective and **beautiful visuals**

Confirmation of Attendance and Survey

- If you would like to get a “Confirmation of Attendance” please click on the link below*

Confirmation of Attendance and Survey

- The link also takes you to our course feedback survey. Filling it in is optional but highly appreciated!



Thank you!

*Please send your request within the next 3 days

Thank you!