Data Processing

Chapter 2 – Hands on Data Analytics for Everyone

October 17, 2022

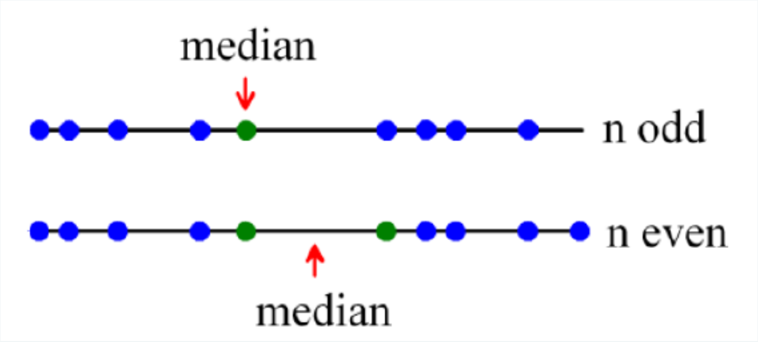
北京师范大学-香港浸会大学联合国际学院
United International College

Contents

Statistical measures can be used to describe a dataset:

- Range
- Min/max values
- Mean $\mu = \frac{1}{n}\sum_{i=1}^{n} x_i$
- Variance $\sigma^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \mu)^2$
- Standard deviation $\sigma = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \mu)^2}$
- Median (The middle number; found by ordering all data points and picking out the one in the middle - or if there are two middle numbers, taking the mean of those two numbers)
- Mode (Most frequently occurring value)
- Percentiles (Quartiles)
- Number of missing values
- …

Median, Quantiles, Quartiles, Interquartile Range



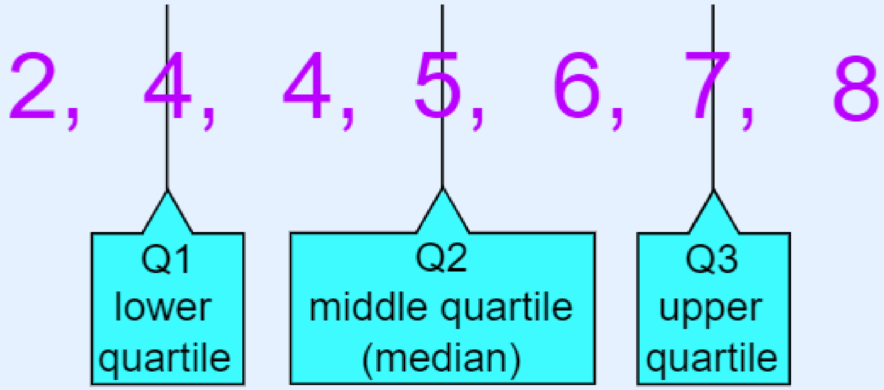
- **Median:** The value in the middle (for values sorted in increasing order)
- **q%-quantile** (0 < q < 100): The value for which q% of the values are smaller and 100-q% are larger. The median is the 50%-quantile
- **Quartiles**: 25%-quantile (1st quartile), median (2nd quantile), 75%-quantile (3rd quartile)
- **Interquartile range (IQR):** 3rd quartile – 1st quartile

Median, Quantiles, Quartiles, Interquartile Range - Example

Example: 5, 7, 4, 4, 6, 2, 8

Put them in order: 2, 4, 4, 5, 6, 7, 8

Cut the list into quarters:



And the result is:

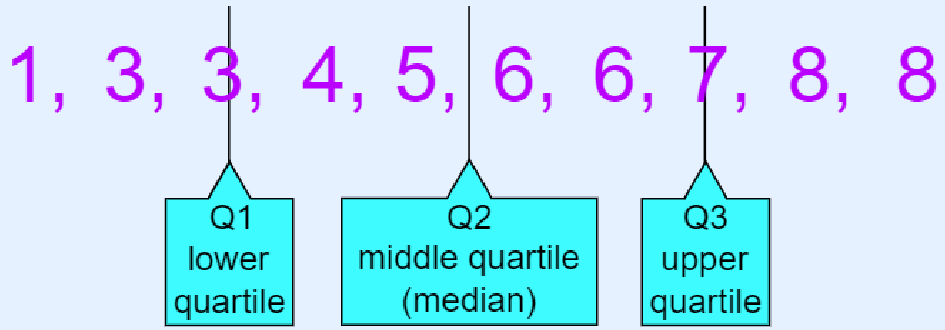
- Quartile 1 (Q1) = **4**
- Quartile 2 (Q2), which is also the Median, = **5**
- Quartile 3 (Q3) = **7**

Median, Quantiles, Quartiles, Interquartile Range - Example

Example: 1, 3, 3, 4, 5, 6, 6, 7, 8, 8

The numbers are already in order

Cut the list into quarters:



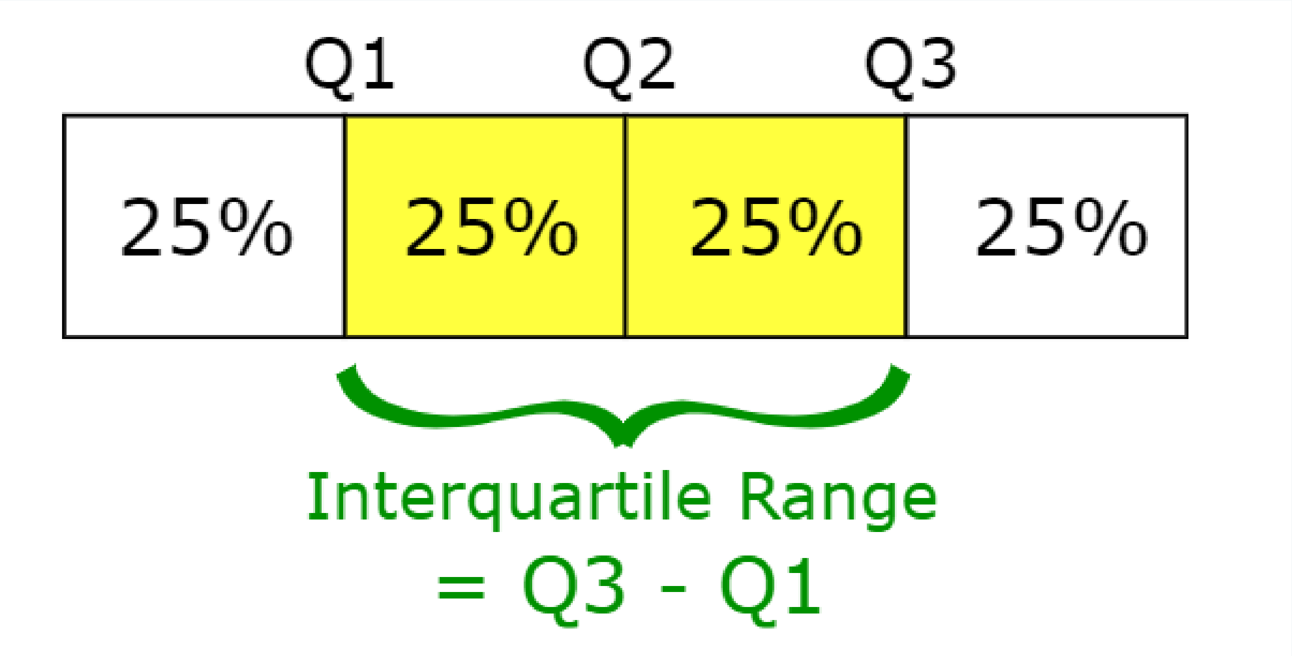
In this case Quartile 2 is half way between 5 and 6:

Q2 = (5+6)/2 = **5.5**

And the result is:

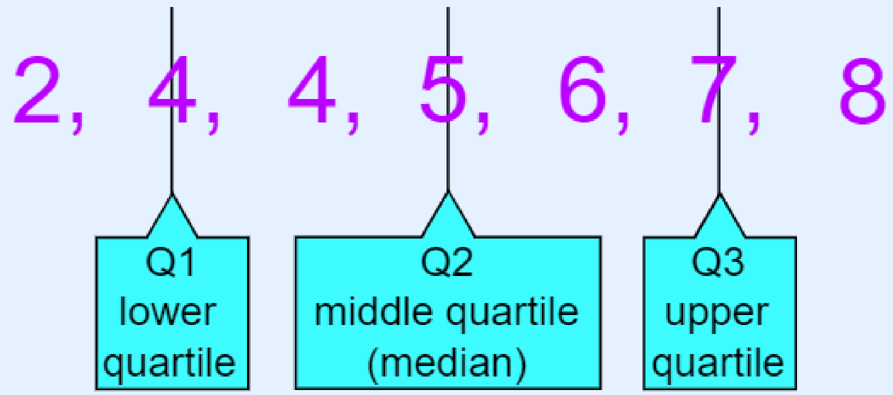
- Quartile 1 (Q1) = **3**
- Quartile 2 (Q2) = **5.5**
- Quartile 3 (Q3) = **7**

Median, Quantiles, Quartiles, Interquartile Range - Example



Median, Quantiles, Quartiles, Interquartile Range - Example

Example:



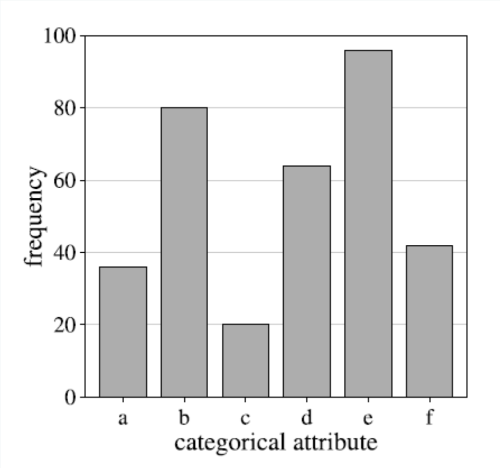
The **Interquartile Range** is:

Q3 − Q1 = 7 − 4 = **3**

https://www.mathsisfun.com/data/quartiles.html

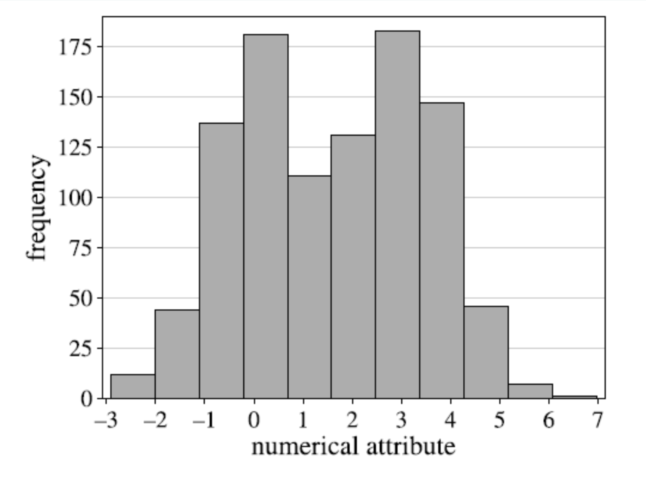
Contents

Bar Chart



A bar chart is a simple way to depict the frequencies of the values of a categorical attribute.

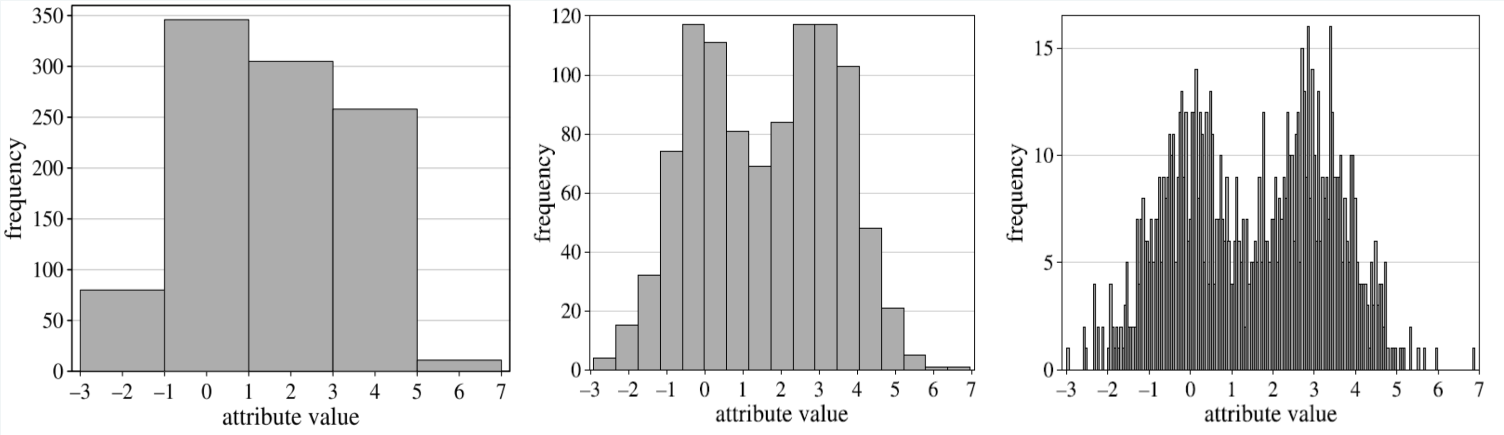
Histogram



- A histogram shows the frequency distribution for a numerical attribute.
- The range of the numerical attribute is discretized into a fixed number of intervals (bins), usually of equal length.
- For each interval, the (absolute) frequency of values falling into it is indicated by the height of a bar.

Choice of Number of Bins

The histogram in the figures resulted from a sample of size $n = 1000$



Choosing a low number of bins, the two peaks of the original distribution are no longer visible, and one gets the wrong impression that the distribution is unimodal

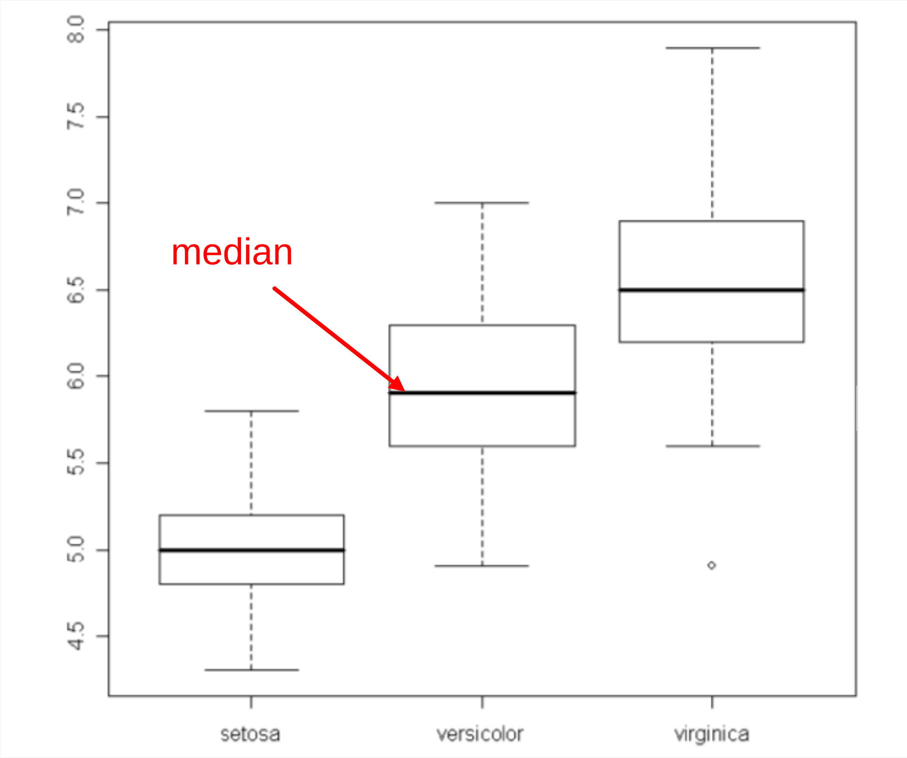
Choosing a high number of bins usually leads to a very scattered histogram in which it is difficult to distinguish true peaks from random peaks

Best choice for number k of bins in the histogram?

- Sturge's Rule $k = [log_2(n) + 1]$

Boxplots

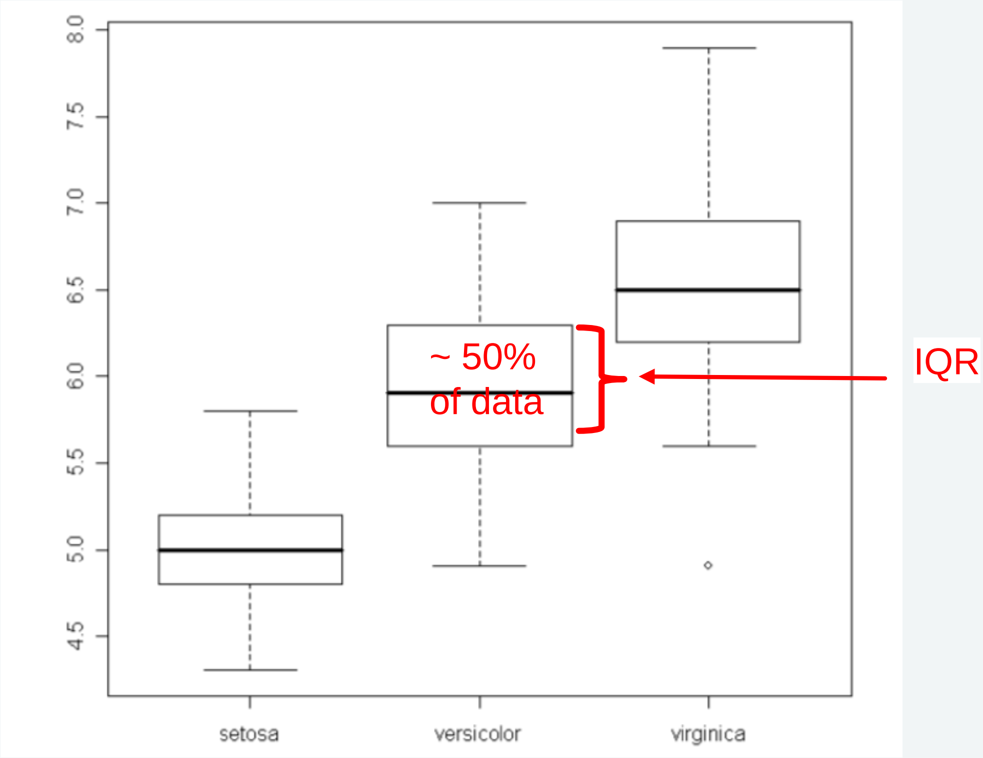
Boxplots are a very compact way to visualize and summarize the main characteristics of a numeric attribute, through the ***median***, the IQR, and possible outliers



https://www.khanacademy.org/math/statistics-robability/summarizing-quantitative-data

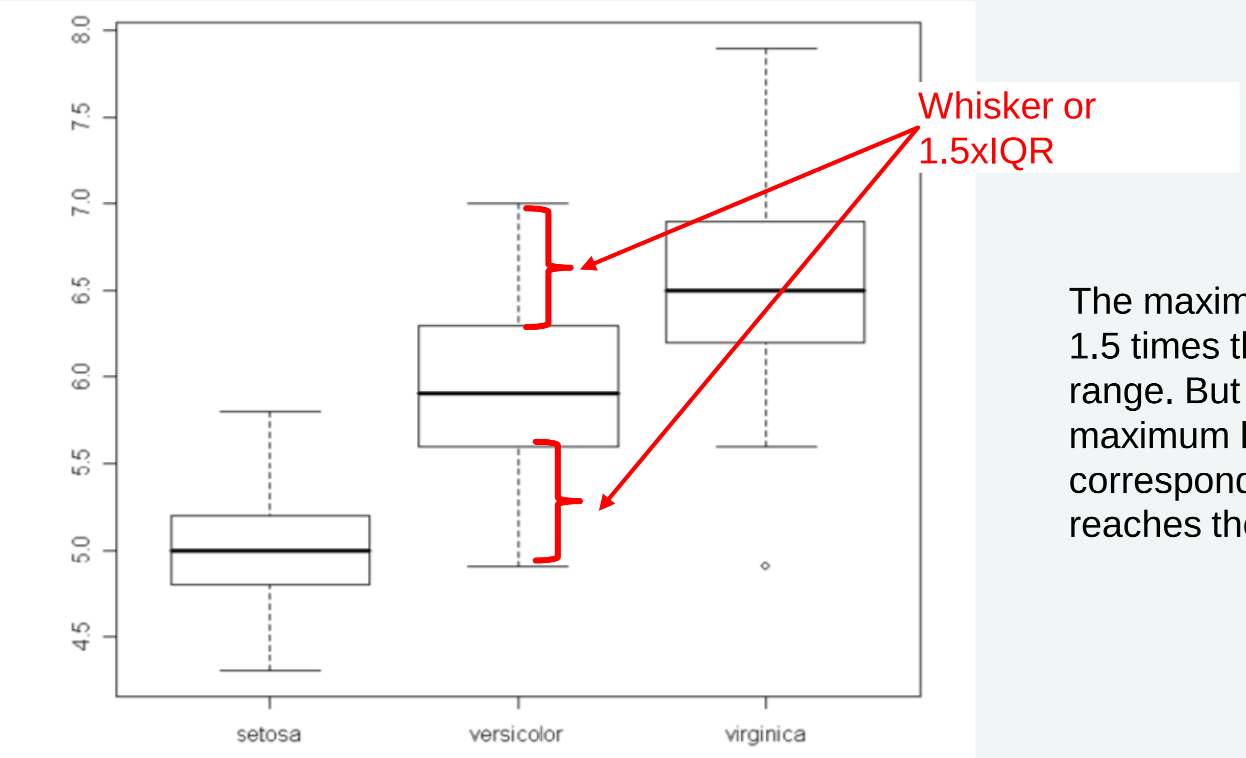
Boxplots

Boxplots are a very compact way to visualize and summarize the main characteristics of a numeric attribute, through the median, the ***IQR***, and possible outliers



Boxplots

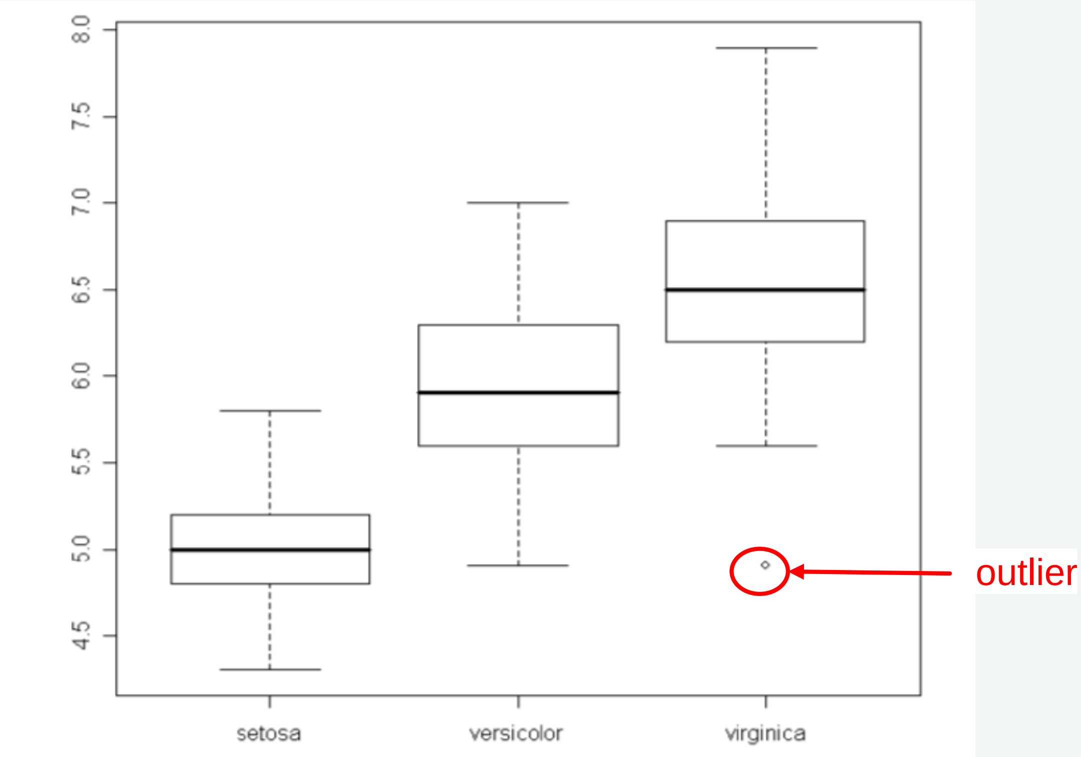
Boxplots are a very compact way to visualize and summarize the main characteristics of a numeric attribute, through the median, the ***IQR***, and possible outliers



The maximum length of each whisker is 1.5 times the length of the interquartile range. But if there is no data point at the maximum length of a whisker, the corresponding whisker is shortened until it reaches the next data point.

Boxplots

Boxplots are a very compact way to visualize and summarize the main characteristics of a numeric attribute, through the median, the *IQR*, and possible **outliers**

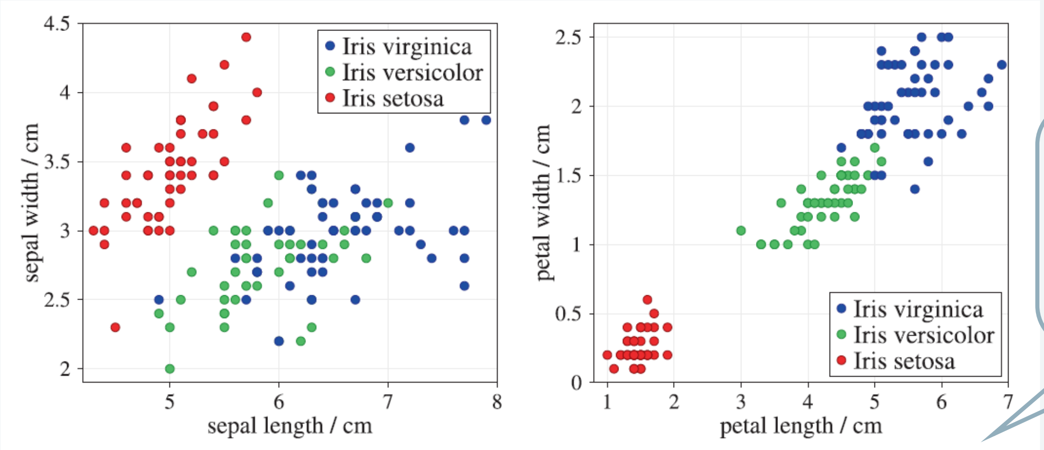


Data points lying outside the whiskers are considered as outliers and are indicated in the form of small circles.

博文雅志 真知笃行

In knowledge and in deeds, unto the whole person

Scatter Plot for Two Attributes

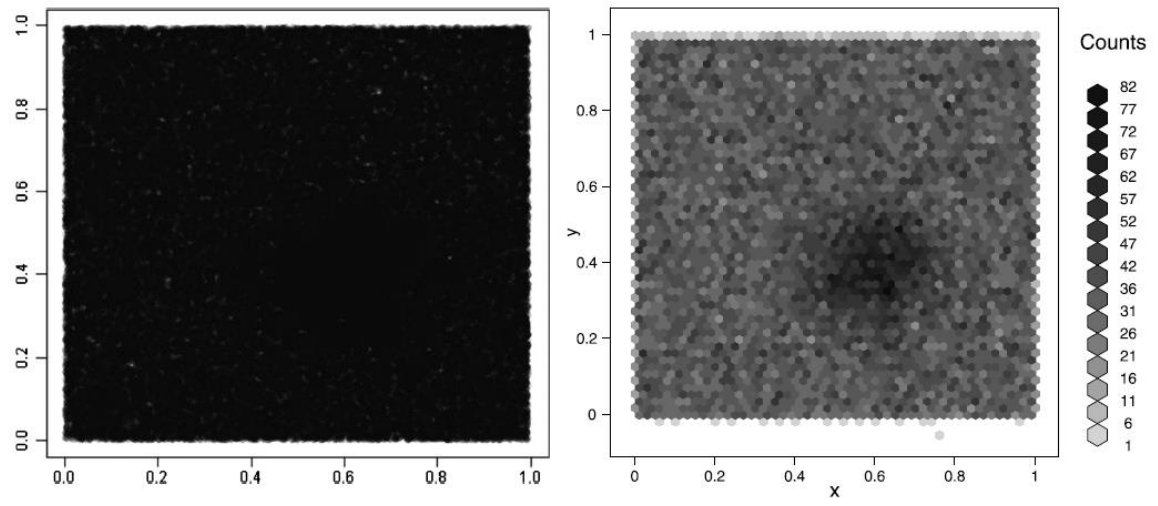


Petal length and width provide better class separation than sepal length and width

Scatter plots of the Iris(鸢(yuan)尾花) data set for sepal(萼片) length vs. sepalwidth (left) and for petal (花瓣) length vs. petal width (right). All quantities are measured in centimetres

- In scatter plots two attributes are plotted against each other
- Can be enriched with additional features (color, shape, size)
- Suitable for small number of points; not suitable for large datasets
- Points can hide each other -> add **Jitter** (a small random value to each point)

Scatter Plot for Two Attributes



Density plot (left) and a plot based on hexagonal binning (right) for a dataset with n = 100,000 instances

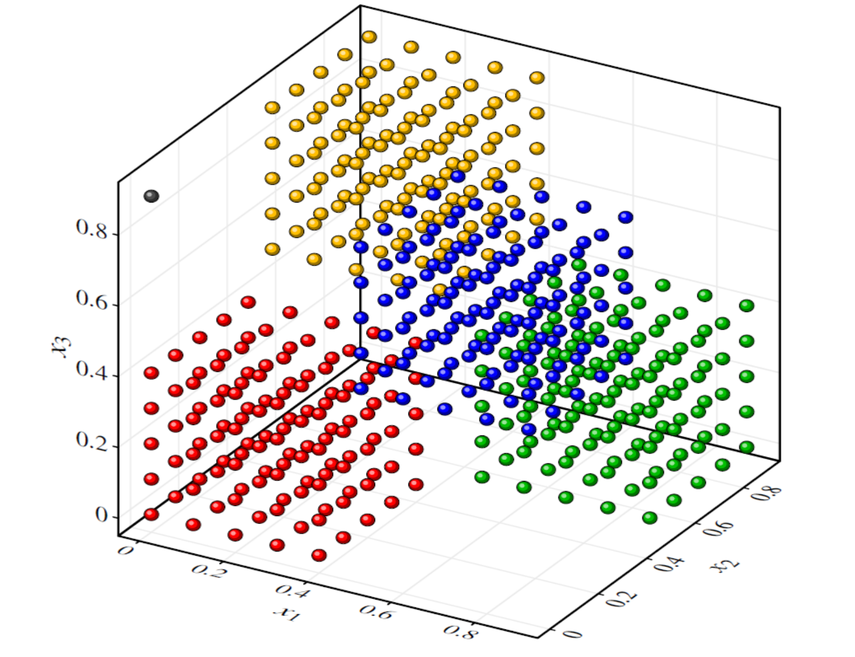
Both plots indicate a higher density of the data around the point (0.6, 0.4), which cannot be seen in the simple scatter plot

- Scatter plot is not suitable for large datasets
- Alternatives:
  - Density plot (*hexagonal binning*) for example using semi-transparent points: the more points in the same place the less transparent
  - Binning points into rectangles or hexagons and heat scale color

Contents

Visualization of Three-dimensional Data: 3D Plot

A display or plot is **by definition two-dimensional**, so that only maximum two axes (attributes) can be incorporated. **3D** techniques can be used to incorporate three axes (attributes).



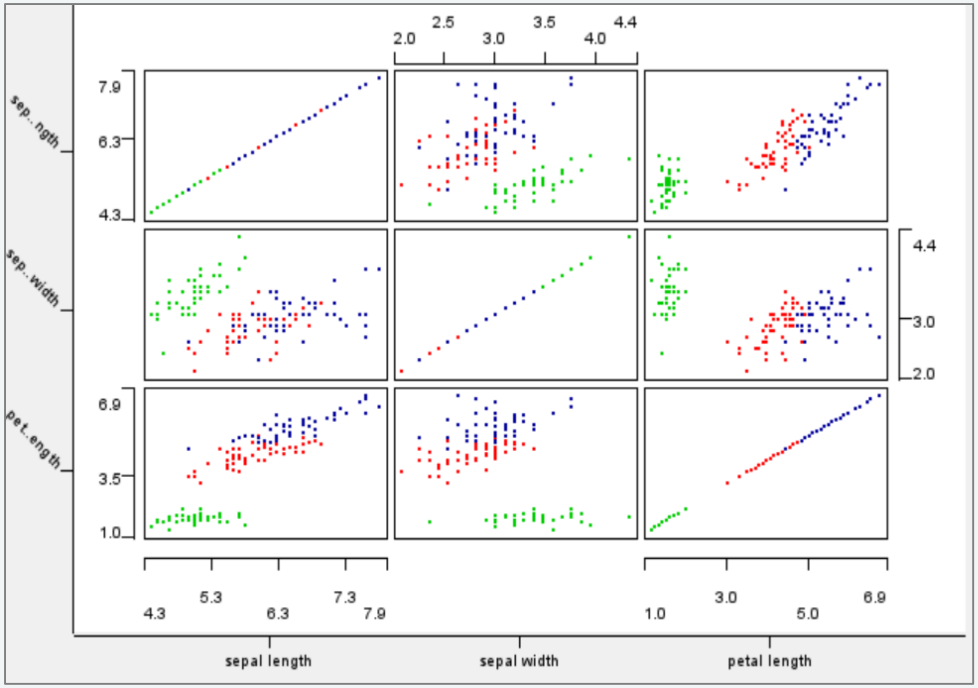
3D scatter plot

Example

- A data set distributed over a cube in a **chessboard-like pattern**.
- The colors are only meant to make the different cubes more easily discernible (可辨别的). They do not indicate classes.
- Note the outlier in the upper left corner

Visualization of Three-dimensional Data: Scatter Matrixes

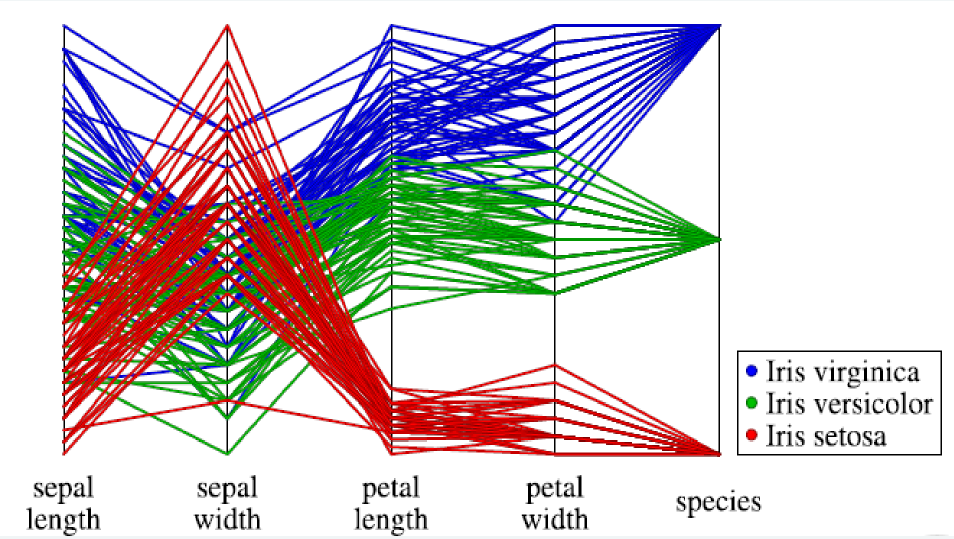
- A matrix of scatter plots $m \times m$ where m is the number of attributes (data dimensionality)
- For m attributes there are $\binom{m}{2} = m(m-1)/2$ possible scatter plots
- e.g. For 50 attributes there are 2450/2 scatter plots!



Scatter matrix

Parallel Coordinates Plot

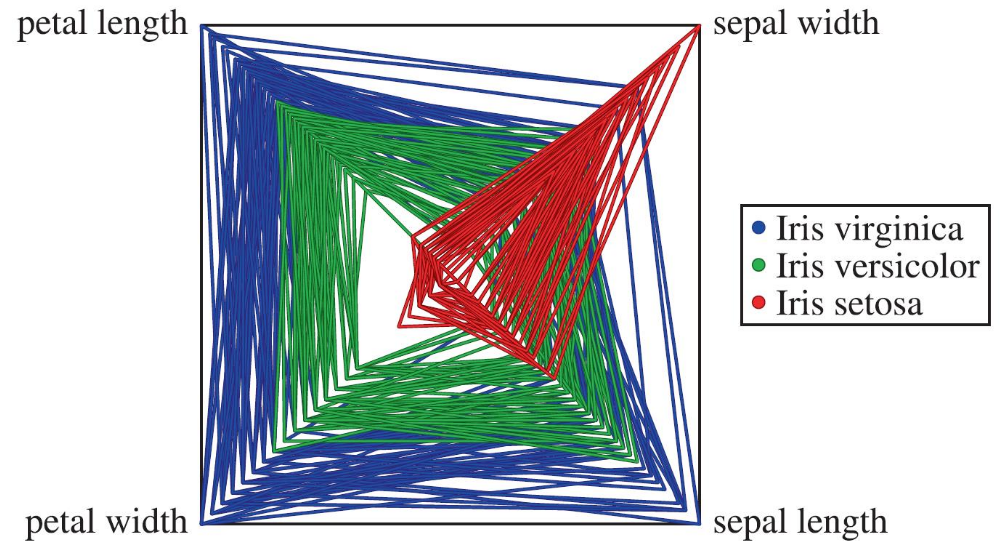
- Parallel coordinates draw the coordinate axes for each attribute parallel to each other, so that there is no limitation for the number of axes to be displayed.
- For each data object, a polyline is drawn connecting the values of the attributes on the corresponding axes.
- Maintains the original attributes
- Limited number of entries
- How do we spot correlation between features?



Parallel coordinates plot for the Iris data set

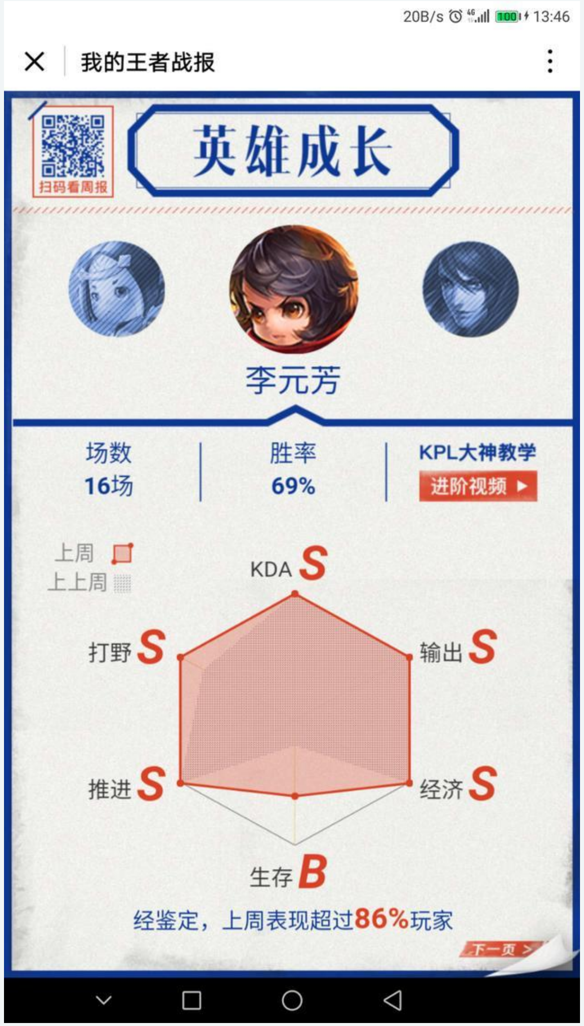
Radar Plot

- Similar idea of the Parallel Coordinates plot
- Axes are drawn in a star-like fashion intersecting in one point
- Also called spider plots
- Suitable for small datasets



Radar plot for the Iris data set

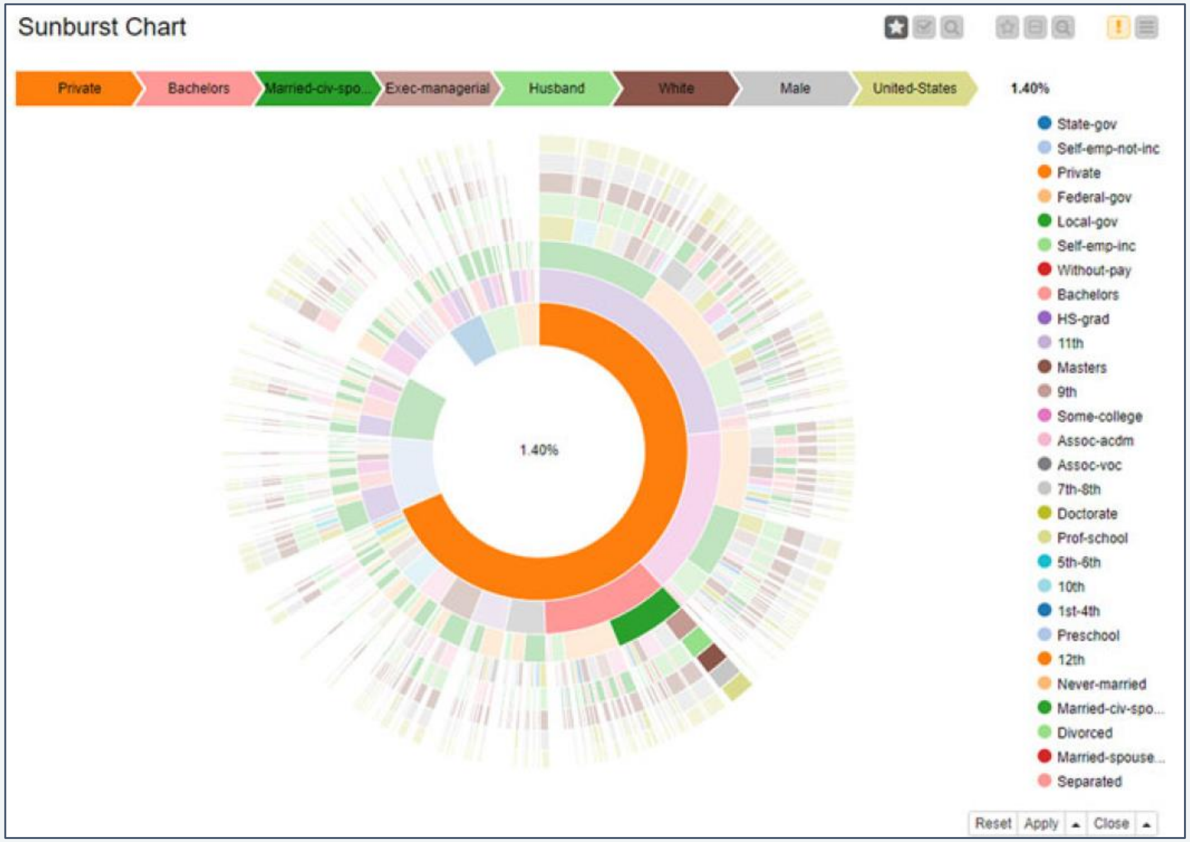
Radar Plot



20B/s 13:46
我的王者战报
扫码看周报
英雄成长
李元芳
场数
16场
胜率
69%
KPL大神教学
进阶视频
上周
上上周
KDA S
打野 S
输出 S
推进 S
经济 S
生存 B
经鉴定，上周表现超过86%玩家
下一页

当前赛季(S14)
全部比赛
标准模式
输出
生存
KDA
团战
发育
场次
7
胜率
42.9%
综合得分
74.3
详细数据
分享
全场最佳2
败方最佳2
三连决胜0
四连超凡0
五连绝世0
超神0

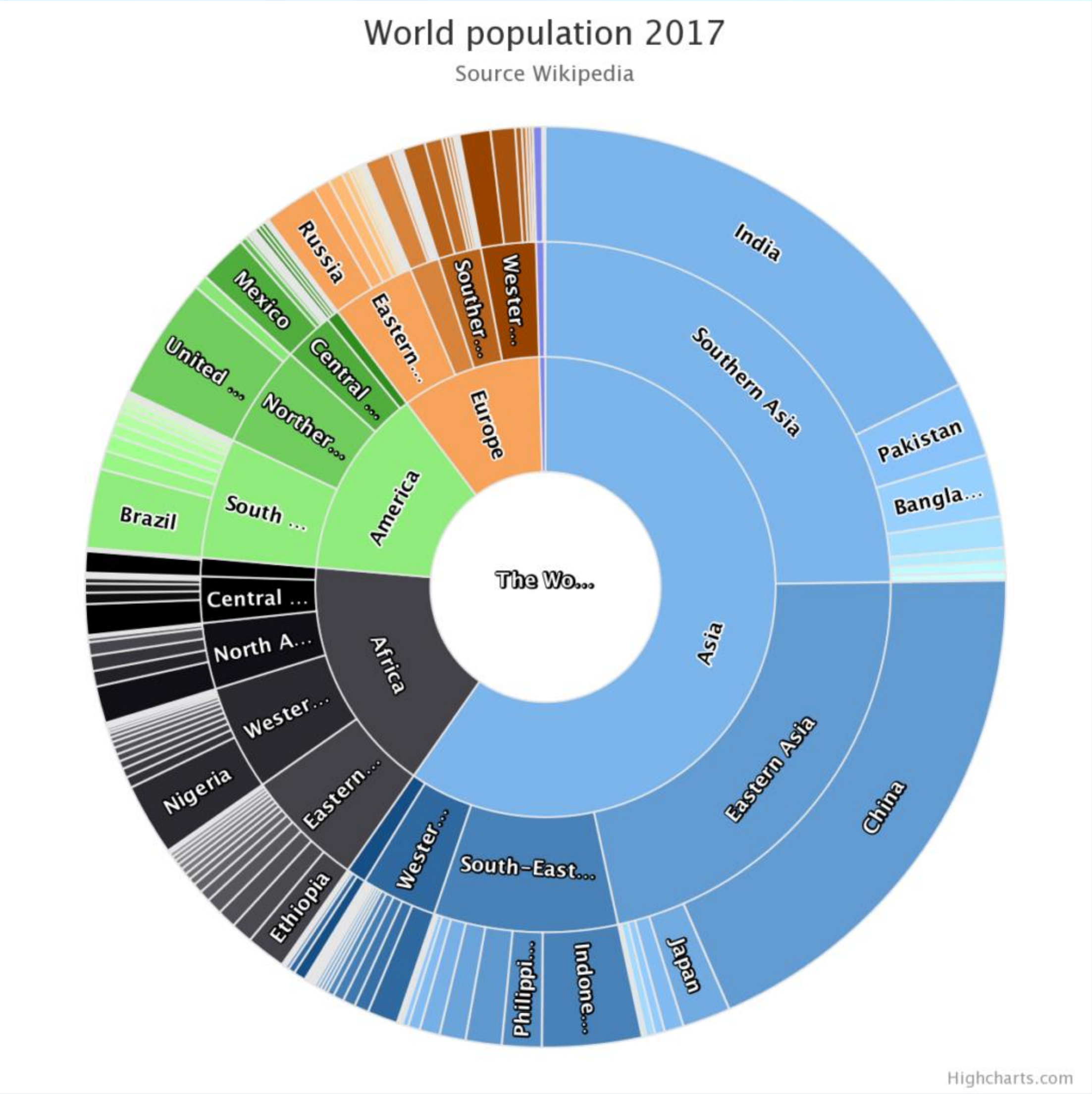
Sunburst Chart



- Display multidimensional hierarchical nominal data in a radial layout
- One section ⇔ one attribute

- Root attribute in the center, external sections are attributes located deeper in the hierarchy
- Area of a section represents the accumulated value of all descending sections

Sunburst Chart



Contents

Is there such a thing as “too much data”?

"Too much data":

- Consumes storage **space**
- Eats up processing **time**
- Is difficult to **visualize**
- Inhibits ML algorithm **performance**
- Beware of the model: **Garbage** in → Garbage out

Feature Selection vs. Dimensionality Reduction

Both methods are used for reducing the number of features in a dataset. However:

- Feature selection is simply selecting and excluding given features **without changing** them.
- Dimensionality reduction **might transform** the features into a lower dimension.
- Feature selection is often a somewhat more aggressive and more computationally expensive process.
  - **Backward** Feature Elimination
  - **Forward** Feature Construction

Feature Selection

- **Feature Selection**: choose a subset of the features (attributes) that is **as small as** possible and sufficient for the data analysis (= still informative!)

- Feature Selection includes:
  - Removing (more or less) **irrelevant** **features/fields** and
  - Removing **redundant** **features**

- **Evaluation** function to compare sets of attributes
- Strategy (heuristic) to select the possible feature subsets to be compared against each other with this measure

Feature Selection Techniques

- **Forward selection**

  Start with the **empty** set of features and **add** features one by one. In each step, add the feature that yields the best improvement of the **performance**.

- **Backward elimination**

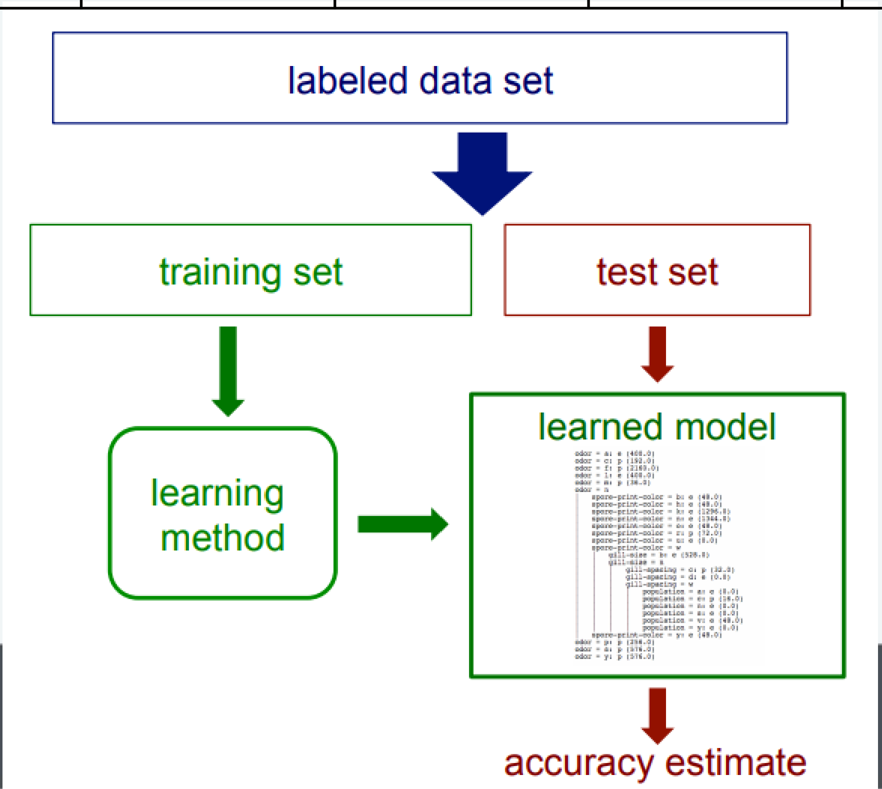
  Start with the **full** set of features and **remove** features one by one. In each step, remove the feature that results in the smallest decrease in **performance**.

Forward Feature Construction (Greedy Bottom-up)

1. First, train n separate models on one single input feature and keep the feature that produces the best **accuracy**.
2. Then, train $n-1$ separate models on 2 input features, the selected one and one more. At the end keep the additional feature that produces the best accuracy.
3. And so on ⋯ Continue until an acceptable error rate is reached.

Why we do Feature Selection? Model Evaluation and Feature Selection

Size in feet^2	# of bedrooms	# of floors	Age of home (years)	Price ($) in 1000's	Pet free?	In flood zone?
2104	5	1	45	460	Y	N
1416	3	2	40	232	N	Y
1534	3	2	30	315	Y	N
852	2	1	36	178	N	N
…	…	…	…	…	…	…



Backward Feature Elimination (Greedy Top-down)

1. First train one model on *n* input features
2. Then train *n* separate models each on $n-1$ input features and remove the feature whose removal produced the least **disturbance**
3. Then train $n-1$ separate models each on $n-2$ input features and remove the feature whose removal produced the least disturbance
4. And so on. Continue until desired **maximum error rate** on *training* data is reached.

Dimensionality Reduction Techniques

- Measure based
  - Ratio of missing values
  - Low variance
  - High Correlation

Dimensionality Reduction Based on Missing Values Ratio

First partition (as defined in dialog) - 0:337:0:276 - Partitioning (80% vs. 20%)

File Hilite Navigation View

Table "default" - Rows: 40000 | Spec - Columns: 231 | Properties | Flow Variables

Row ID	D Var16	I Var17	I Var18	I Var19	S Var20	I Var21	I Var22	I Var23	I Var24	I Var25	I Var26	I Var27	D Var28
Row0	?	?	?	?	?	464	580	?	14	128	?	?	166.56
Row1	?	?	?	?	?	168	210	?	2	24	?	?	353.52
Row2	?	?	?	?	?	1212	1515	?	26	816	?	?	220.08
Row4	?	?	?	?	?	64	80	?	4	64	?	?	200
Row7	?	?	?	?	?	32	40	?	2	16	?	?	230.56
Row8	?	?	?	?	?	200	250	?	2	64	?	?	300.32
Row10	?	?	?	?	?	92	115	?	6	112	?	?	133.12
Row11	?	?	?	?	?	236	295	?	8	40	?	?	133.12
Row12	?	?	?	?	?	0	0	?	?	0	?	?	240.56
Row13	?	?	?	?	?	480	600	?	10	216	?	?	176.56
Row14	?	?	?	?	?	148	185	?	0	8	?	?	236.08
Row16	?	?	?	?	?	584	730	?	6	320	?	?	220.08
Row17	?	?	?	?	?	168	210	?	2	32	?	?	166.56
Row18	?	?	?	?	?	12	15	?	2	0	?	?	253.52
Row20	?	?	?	?	?	168	210	?	2	56	?	?	272.08
Row21	?	?	?	?	?	20	25	?	2	0	?	?	86.96
Row22	?	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	?	?	166.56
Row23	?	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	?	?	198.88
Row24	?	?	?	?	?	216	270	?	8	128	?	?	200
Row25	?	?	?	?	?	152	190	?	4	16	?	?	20.08
Row26	?	0	0	0	?	?	?	?	?	?	?	?	?
Row28	?	?	?	?	?	0	0	?	?	0	?	?	257.28
Row29	?	?	?	?	?	312	390	?	0	120	?	?	200
Row30	?	?	?	?	?	112	140	?	4	56	?	?	166.56
Row31	?	?	?	?	?	28	35	?	0	16	?	?	285.2
Row33	?	?	?	?	?	160	200	?	4	40	?	?	[illegible]
Row36	?	?	?	?	?	612	765	?	14	360	?	?	200
Row37	?	?	?	?	?	380	475	?	4	208	?	?	336.56
Row38	?	?	?	?	?	76	95	?	0	16	?	?	213.36
Row40	?	?	?	?	?	228	285	?	22	56	?	?	200
Row41	?	?	?	?	?	120	150	?	10	80	?	?	133.12
Row42	?	5	0	0	?	?	?	?	?	?	?	?	?
Row43	?	?	?	?	?	72	90	?	0	40	?	?	191.36
Row44	?	?	?	?	?	0	0	?	?	0	?	?	120.4
Row47	?	?	?	?	?	0	0	?	?	0	?	?	186.64
Row48	?	?	?	?	?	172	215	?	4	200	?	?	137.68
Row49	?	?	?	?	?	0	0	?	?	0	?	?	274.16

Missing Value

IF (% missing value > threshold) THEN remove column

Dimensionality Reduction Based on Low Variance

Output table - 0:347:0:337 - Missing Value (Numeric: 0)

File Hilite Navigation View

Table "default" - Rows: 40000 | Spec - Columns: 231 | Properties | Flow Variables

Row ID	20	I Var21	I Var22	I Var23	I Var24	I Var25	I Var26	I Var27	D Var28
Row51	h	336	420	0	8	72	0	0	133.12
Row52	h	120	150	0	0	16	0	0	286.96
Row54	h	124	155	0	0	0	0	0	234.72
Row55	h	184	230	0	4	64	0	0	642.64
Row56	h	268	335	0	4	88	0	0	133.12
Row57	h	128	160	0	0	96	0	0	198.88
Row59	h	132	165	0	0	112	0	0	253.52
Row60	h	44	55	0	0	24	0	0	186.64
Row61	h	104	130	0	4	72	0	0	166.56
Row62	h	212	265	0	6	136	0	0	379.6
Row63	h	20	25	0	0	0	0	0	166.56
Row65	h	492	615	0	18	256	0	0	133.12
Row66	h	148	185	0	2	8	0	0	186.64
Row68	h	140	175	0	2	40	0	0	176.56
Row69	h	0	0	0	0	0	0	0	166.56
Row71	h	0	0	0	0	0	0	0	392.08
Row72	h	124	155	0	6	88	0	0	153.2
Row73	h	152	190	0	0	32	0	0	253.52
Row74	h	324	405	0	8	104	0	0	186.64
Row75	h	0	0	0	0	0	0	0	0
Row76	h	60	75	0	6	0	0	0	200
Row77	h	180	225	0	4	88	0	0	166.56
Row78	h	232	290	0	4	144	0	0	200
Row79	h	16	20	0	0	16	0	0	313.68
Row81	h	152	190	0	0	48	0	0	220.08
Row82	h	108	135	0	4	88	0	0	166.56

Note: requires min-max-normalization, and only works for **numeric** columns

If column has **constant** value (variance = **0**), it contains no useful information

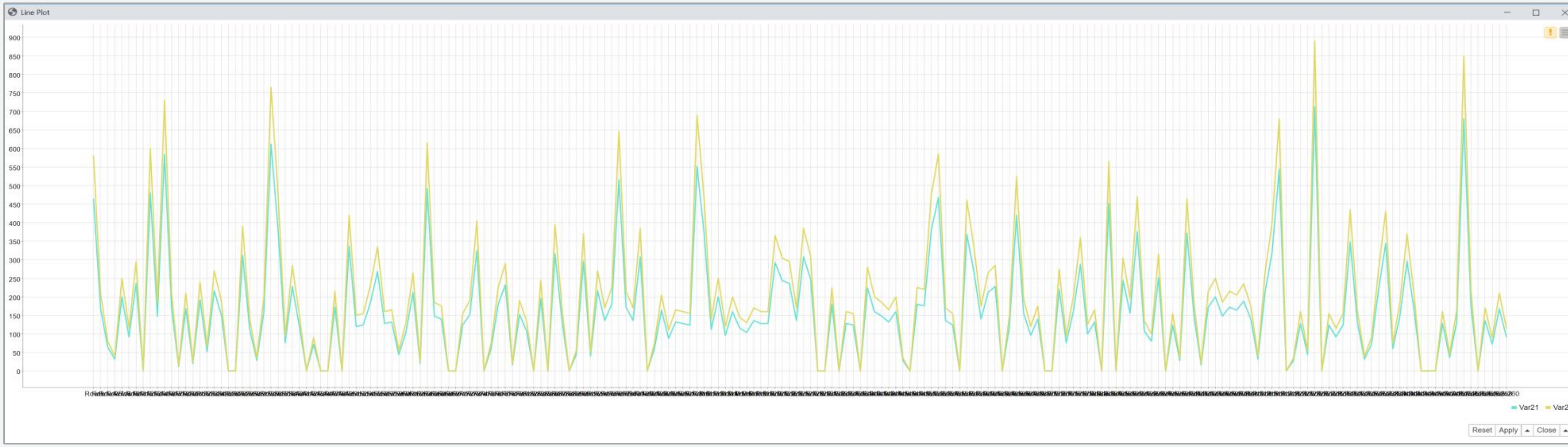
In general: **IF (variance < threshold) THEN remove column**

Dimensionality Reduction Based on High Correlation

Two **highly correlated** input variables probably carry similar information

$$IF\ (corr(var1, var2) > threshold\ \Rightarrow remove\ var1$$

Note: requires min-max-normalization of numeric columns



Contents

Missing Value Imputation: Motivation

Data is not always available

- E.g., many tuples have no recorded value for several attributes, such as weight in a people database

Missing data may be due to:

- Equipment malfunctioning (broken sensors)
- Inconsistency with other recorded data and thus deleted
- Data not entered (manually)
- Data not considered important at the time of collection
- Data format / contents of database changes
- Refusal to answer a question
- Irrelevant attribute for the corresponding object (pregnant (yes/no) for men)
- Missing value might not necessarily be indicated as missing (instead: zero or default values).

Missing Values: Types

Types of missing values:

Example: Suppose you are modeling weight Y as a function of sex X

- **Missing Completely At Random** (MCAR): the probability that a value for X is missing does neither depend on the value of X nor on other variables.
  There may be no particular reason why some people told you their weights and others didn't.

- **Missing At Random** (MAR): the probability that Y is missing depends only on the value of X.
  One sex X may be less likely to disclose its weight Y.

- **Not Missing At Random** (NMAR): the probability that Y is missing depends on the unobserved value of Y itself.
  Heavy (or light) people may be less likely to disclose their weight.

Missing Values Imputation

How to handle missing values?

- Ignore/delete the record
- Fill in (impute) missing value as:
- **Fixed value**: e.g., “unknown”, -9999, -1 when only positive numbers in the domain, etc.
- Attribute **mean / median / mode**
- Attribute **most frequent value**
- **Next / previous /avg interpolation / moving avg value** (in time series)
- **A predicted value** based on the other attributes (inference-based such as Bayesian, Decision Tree, ...)

Contents

What are outliers?

- An **outlier** is a value or data object that is far away or very different from all or most of the other data.
- Errors in measurements or exceptional conditions that don't describe the common functioning of the underlying system
- Outliers are supposed to be rare

Outlier Detection

Causes for outliers:

- Data **quality** problems (erroneous data coming from wrong measurements or typing mistakes)
- Exceptional or unusual **situations**/data objects.

Outlier handling :

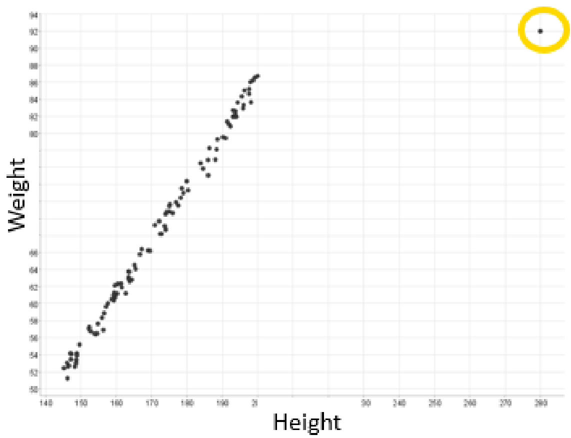
- Outliers coming from erroneous data should be **excluded** from the analysis.
- Even if the outliers are correct (exceptional data), it is sometimes useful to exclude them from the analysis.
- For example, a single extremely large outlier can lead to completely misleading values for the mean value.

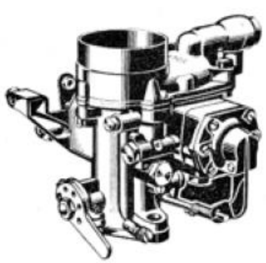
Outlier Detection

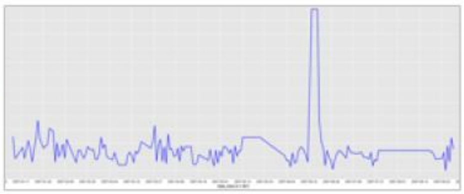
An outlier could be, for example, a rare behaviour, a system defect, a measurement error, or a reaction to an unexpected event

Image: https://www.nytimes.com/2016/06/25/business/international/brexit-financial-economic-impact-leave.html

Brexit referendum

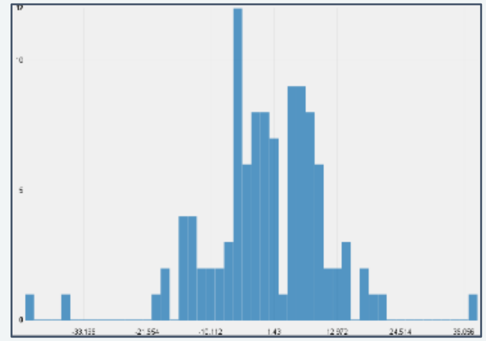


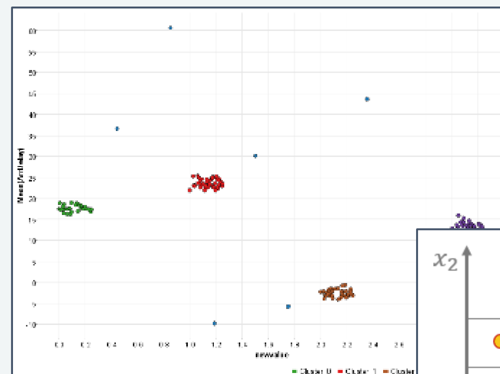


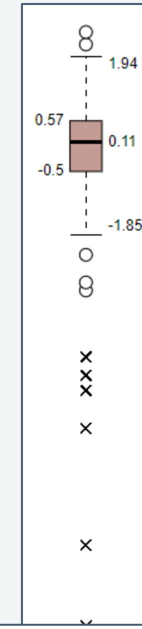


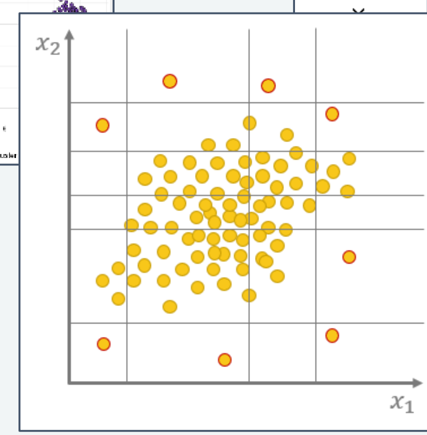
Outlier Detection Techniques

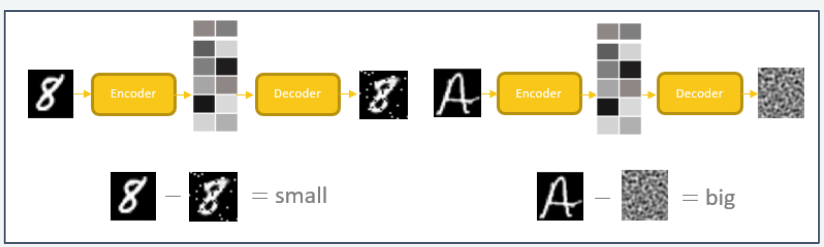
- Knowledge-based
  - We know that a 200 year old person must be a mistake
  - We know that "A" in a number corpus is an outlier
- Statistics-based
  - Distance from the median
  - Position in the distribution tails
  - Distance to the closest cluster center
  - Error produced by an autoencoder
  - Number of random splits to isolate a data point from other data











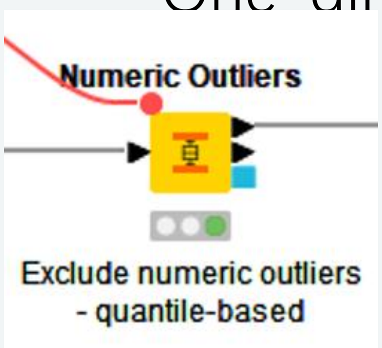
Statistical Methods for Outlier Detection

- Quantile-based: **Box plot**
- Distribution-based: **Z-Score**
- Cluster-based: **DBSCAN**
- Neural Autoencoder
- Isolation Forest
- …

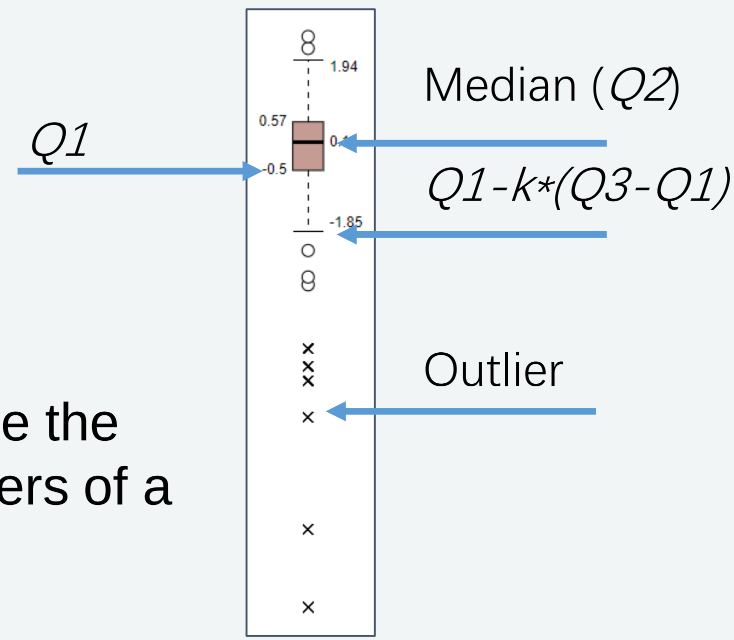
Quartile based (Box Plots)

Challenges:

- Outliers in the data expand the quantiles
- Skewed data might require different *k* to detect upper and lower outliers
- One-dimensional



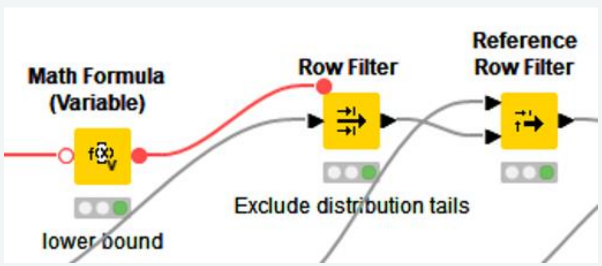
Flag data points outside the upper and lower whiskers of a box plot as outliers

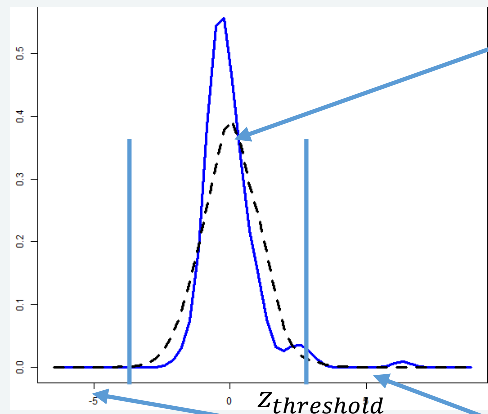


Z-score Based Outlier

Challenges:

- Normality assumption
- The parameters of the distribution are sensitive to outliers
- Doesn’t work for data with a trend and seasonality





Standard Normal Distribution

Flag data points in the distribution tails as outliers

Outliers

Contents

From Categorical to Numerical

- Binary attribute: numerical attribute with the values 0 and 1.
- Ordinal attribute (“sortable”): enumerate in the correct order 1, . . . , *k*
- Categorical attribute (not ordinal) with more than two values, say $a_1, \ldots, a_k$, should **not be converted into a single numerical attribute instead:** **convert to k attributes $A_1, \ldots, A_k$ with values 0 and 1.**
- a_i is represented by $a_i = 1$ and $a_j = 0$ for $i \neq j$ (1−of−*n* encoding).

From Numerical to Categorical

Splitting a numerical range into a number of bins

- **Equi-width discretization:** splits the range into intervals (bins) of the same length.
- **Equi-frequency discretization:** splits the range into intervals such that each interval (bin) contains (roughly) the same number of records.
- **V-optimal discretization:** minimizes $\sum_i n_i V_i$ where n_i is the number of data objects in the i-th interval and V_i is the sample variance of the data in this interval.

Contents

Normalization: Motivation

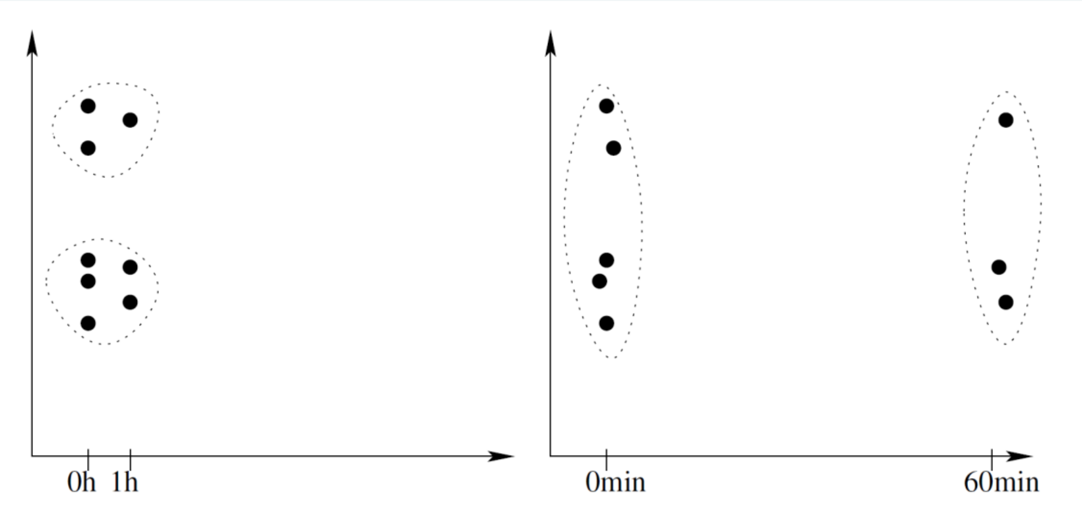
- For some data analysis techniques (e.g. PCA, MDS; cluster analysis) the influence of an attribute depends on the scale or measurement unit.
- To guarantee impartiality, some kind of **standardization** or **normalization** should be applied.

Example:

- Lengths in cm (100 – 200) and weights in kilogram (30 – 150) fall both in approximately the same scale
- What about lengths in m (1-2) and weights also in gram (30000 – 150000)?
  → The weight values in mg dominate over the length values for the similarity of records!

Normalization: Motivation

- 0h vs 1h can be expressed as 0min vs 60min



Goal of normalization:

- Transformation of attributes to make record ranges **comparable**

Construct Data: Assure Impartiality

- In absence of domain knowledge, different techniques can be applied

- **min–max normalization** $n : dom\ X \rightarrow [0,1],\ x \mapsto \frac{x - minX}{max_X - minX}$

Both sensitive to outliers!

- **z-score standardization** $s : dom\ X \rightarrow \mathbb{R},\ x \mapsto \frac{x - \hat{\mu}_X}{\hat{\sigma}_X}$

- **robust z-score standardization** $s : dom\ X \rightarrow \mathbb{R},\ x \mapsto \frac{x - \tilde{x}}{IQR_X}$

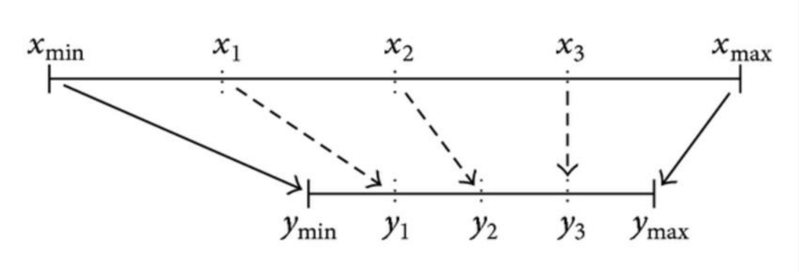
- **decimal scaling** $d : dom\ X \rightarrow [0,1],\ x \mapsto \frac{x}{10^s}$

Normalization: Techniques

- **min-max normalization**

$n: dom(X) \rightarrow [0,1]$

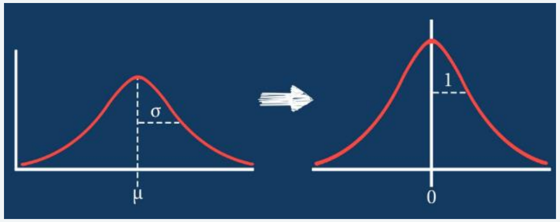
$$y = \frac{x - x_{min}}{x_{max} - x_{min}} (y_{max} - y_{min}) + y_{min}$$



- **z-score normalization**

$s: dom(X) \rightarrow \mathbb{R}$

$$y = \frac{x - \hat{\mu}(X)}{\hat{\sigma}(X)}$$



- **normalization by decimal scaling**

$d: dom(X) \rightarrow [0,1]$

$y = \frac{x}{10^j}$ where j is the smallest integer value larger than $\log_{10}(\max(X))$

Contents

Unstructured String Data

Data in string format is difficult to process (see unstructured data)

We can extract some information from string if the string feature has some pattern. For example the data below contains several useful information if properly cleaned

Job Title	Salary Estimate	Job Description	Rating	Company Name	Location	Headquarters	Size	Founded	Type of ownership	Sector	F
Senior Data Scientist	$111K-$181K (Glassdoor est.)	ABOUT	3.5	Hopper	New York, NY	Montreal, Canac	501 to 1000 employees	2007	Company - Private	Travel & Tourisn	U
Data Scientist, Product Analytics	$111K-$181K (Glassdoor est.)	At Noom, we	4.5	Noom US	New York, NY	New York, NY	1001 to 5000 employees	2008	Company - Private	Consumer Servic	U
Data Science Manager	$111K-$181K (Glassdoor est.)	Decode_M	-1	Decode_M	New York, NY	New York, NY	1 to 50 employees	-1	Unknown	-1	U
Data Analyst	$111K-$181K (Glassdoor est.)	Sapphire	3.4	Sapphire Digital	Lyndhurst, NJ	Lyndhurst, NJ	201 to 500 employees	2019	Company - Private	Information Tec	U
Director, Data Science	$111K-$181K (Glassdoor est.)	Director, Data	3.4	United	New York, NY	New York, NY	51 to 200 employees	2007	Company - Private	Business Service	U
Data Scientist	$111K-$181K (Glassdoor est.)	Job Brief	2.9	IFG Companies	New York, NY	Hartford, CT	201 to 500 employees	1985	Company - Private	Insurance	U
Quantitative Researcher	$111K-$181K (Glassdoor est.)	Experience:	4.4	PDT Partners	New York, NY	New York, NY	51 to 200 employees	1993	Company - Private	Finance	U
Quantitative Research Associate	$111K-$181K (Glassdoor est.)	Seeking a	-1	Enlightenment Resea	New York, NY	New York, NY	1 to 50 employees	-1	Unknown	-1	U
AI Scientist	$111K-$181K (Glassdoor est.)	Paige is a	5	Paige	New York, NY	New York, NY	1 to 50 employees	2018	Company - Private	Information Tec	U
Quantitative Researcher	$111K-$181K (Glassdoor est.)	About the	4.8	Jane Street	New York, NY	New York, NY	501 to 1000 employees	2000	Company - Private	Finance	U
Data Scientist	$111K-$181K (Glassdoor est.)	Company	3.9	Quartet Health	New York, NY	New York, NY	201 to 500 employees	2014	Company - Private	Information Tec	U
Data Scientist/Machine Learning	$111K-$181K (Glassdoor est.)	PulsePoint,Ñ¢,	4.4	PulsePoint	New York, NY	New York, NY	51 to 200 employees	2011	Company - Private	Information Tec	$
Data Scientist, Acorn AI Labs	$111K-$181K (Glassdoor est.)	Medidata:	4.3	Medidata Solutions	New York, NY	New York, NY	1001 to 5000 employees	1999	Company - Public	Information Tec	$
Data Scientist	$111K-$181K (Glassdoor est.)	A Career with	3.9	Point72	New York, NY	Stamford, CT	1001 to 5000 employees	2014	Company - Private	Finance	U
Data Scientist - Alpha Insights	$111K-$181K (Glassdoor est.)	Two Sigma is a	4.4	Two Sigma	New York, NY	New York, NY	1001 to 5000 employees	2001	Company - Private	Finance	U
Data Scientist	$111K-$181K (Glassdoor est.)	Data Scientist	3	Affinity Solutions	New York, NY	New York, NY	51 to 200 employees	1998	Company - Private	Business Service	U
Data Scientist, Analytics	$111K-$181K (Glassdoor est.)	Company Descri	3.6	Etsy	Brooklyn, NY	Brooklyn, NY	501 to 1000 employees	2005	Company - Public	Retail	$
Data Scientist/ML Engineer	$111K-$181K (Glassdoor est.)	Data	3.3	PA Consulting	New York, NY	London, United I	1001 to 5000 employees	1943	Company - Private	Business Service	$
Data Scientist	$111K-$181K (Glassdoor est.)	Job Description	3.6	Etsy	New York, NY	Brooklyn, NY	501 to 1000 employees	2005	Company - Public	Retail	$
VP, Data Science	$111K-$181K (Glassdoor est.)	We are looking	3.9	7Park Data	New York, NY	New York, NY	51 to 200 employees	2012	Company - Private	Business Service	U
Data Scientist, Disney+ Personaliz	$111K-$181K (Glassdoor est.)	Job Summary:Cc	4	Walt Disney Co.	New York, NY	Burbank, CA	10000+ employees	1923	Company - Public	Media	$
Senior Data Scientist, Data Scienc	$111K-$181K (Glassdoor est.)	We,Äôre	3.4	Squarespace	New York, NY	New York, NY	1001 to 5000 employees	2003	Company - Private	Information Tec	U
Quantitative Researcher ,Äì Intern	$111K-$181K (Glassdoor est.)	Job Description	4.1	Citadel Securities	New York, NY	Chicago, IL	201 to 500 employees	2002	Company - Private	Finance	U
Senior Data Engineer (Healthcare	$111K-$181K (Glassdoor est.)	Key	3.4	Enterprise	New York, NY	Jacksonville, FL	51 to 200 employees	1998	Company - Private	Information Tec	$
Data Scientist	$111K-$181K (Glassdoor est.)	Job Description	4.4	WITHIN	New York, NY	New York, NY	51 to 200 employees	2015	Company - Private	Business Service	U
Data Scientist, Marketplace Econo	$111K-$181K (Glassdoor est.)	We are looking f	3.8	Spotify	New York, NY	Stockholm, Swe	1001 to 5000 employees	2006	Company - Public	Information Tec	U
Data Scientist	$111K-$181K (Glassdoor est.)	About Datadog:\	4.1	Datadog	New York, NY	New York, NY	1001 to 5000 employees	2010	Company - Public	Information Tec	$
Lead Data Scientist	$111K-$181K (Glassdoor est.)	Description: Its	3.3	Aetna	New York, NY	Hartford, CT	10000+ employees	1853	Company - Public	Insurance	$
Data Scientist, Personalization	$111K-$181K (Glassdoor est.)	About	5	Hungryroot	New York, NY	New York, NY	1 to 50 employees	2015	Company - Private	Consumer Servic	$
Principal Data Scientist	$111K-$181K (Glassdoor est.)	Description: Its	3.3	Aetna	New York, NY	Hartford, CT	10000+ employees	1853	Company - Public	Insurance	$
Data Scientist	$120K-$140K (Glassdoor est.)	Caserta is a best	4.3	Caserta	New York, NY	New York, NY	51 to 200 employees	2001	Company - Private	Information Tec	U
Data Scientist, Decisions	$120K-$140K (Glassdoor est.)	At Lyft, our	3.7	Lyft	New York, NY	San Francisco, C	5001 to 10000 employees	2012	Company - Public	Information Tec	U

Regex

One way to clean the string data is through data **deletion** and **replacement** through the use of Regular Expressions (regex)

Regular expression is a pattern defining a class of strings. Some examples:

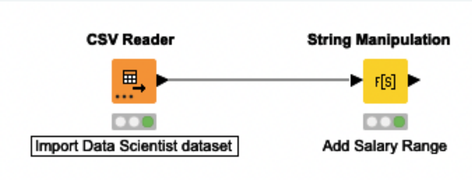
Given a column of strings
- "AnyWord"   search for pattern "AnyWord"
- "^AnyWord" search for values starting with "AnyWord"
- "AnyWord$" search for values ending with "AnyWord"
- "[a-zA-Z]" search for values containing any non numeric character
- "[a-zA-Z]{3}" search for values containing at least 3 non-numeric character
- "Any.*Word" search for values containing Any and Word and anything inbetween the two words.
- "[^0-9]" search for values containing any numeric character

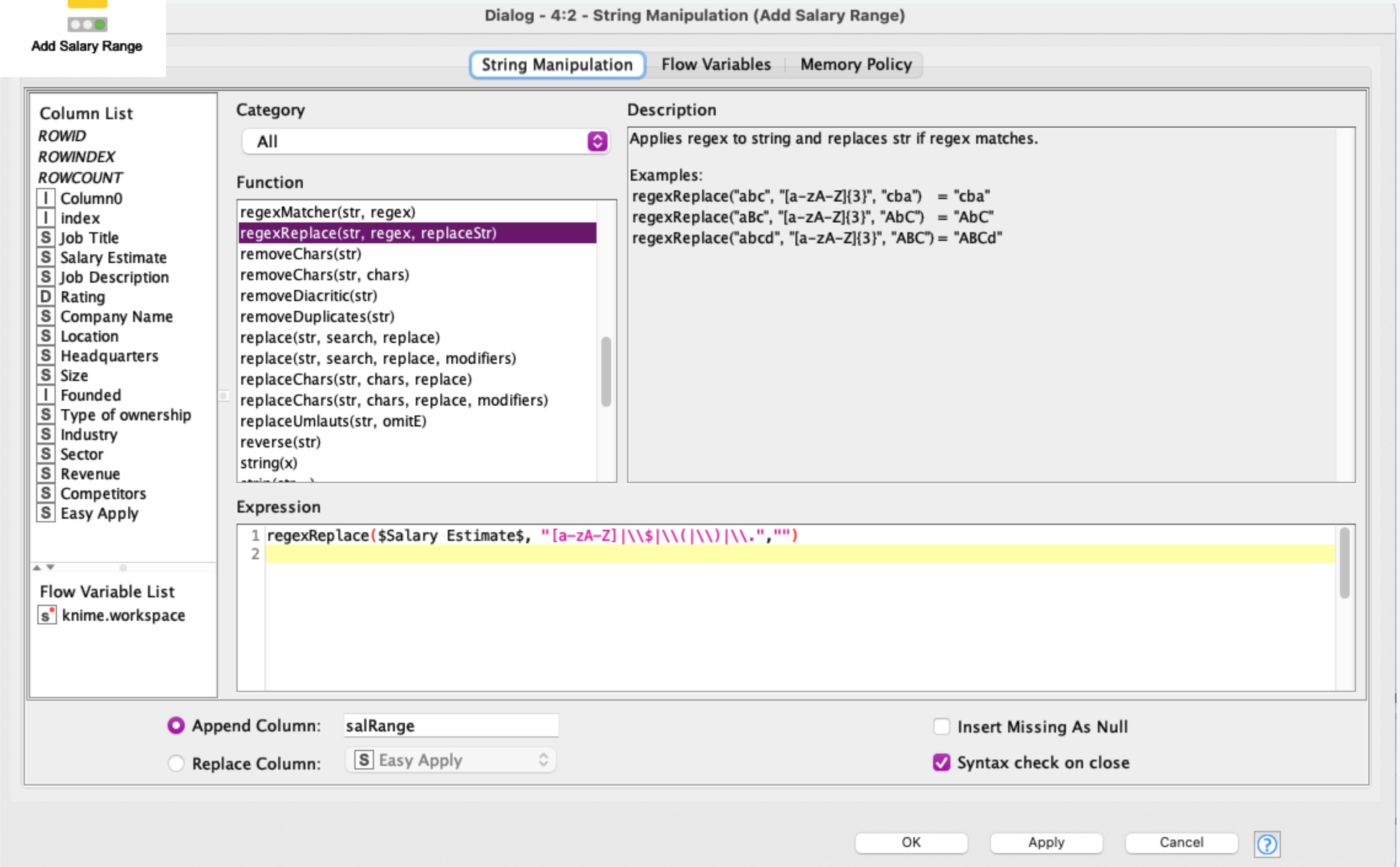
Regex in KNIME

The following nodes have Regex compatibility to transform string data:

- String Manipulation
- Regex Split

Example String Manipulation in KNIME

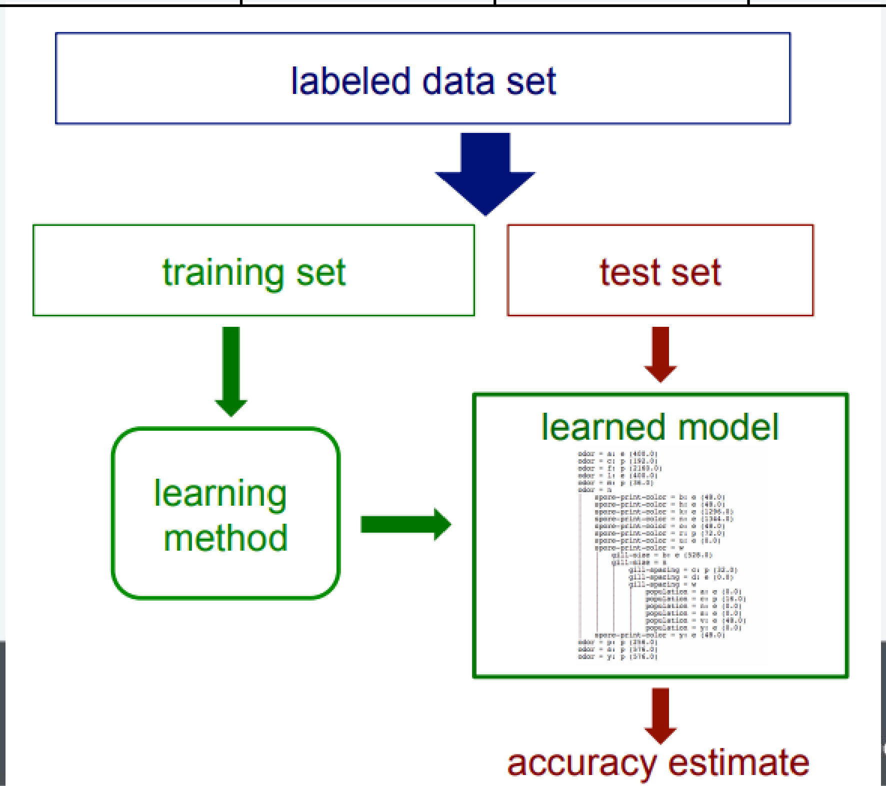




Contents

Why we do Feature Engineering? Model Evaluation and Feature Engineering

Size in feet^2	# of bedrooms	# of floors	Age of home (years)	Price ($) in 1000's	Pet free?	In flood zone?
2104	5	1	45	460	Y	N
1416	3	2	40	232	N	Y
1534	3	2	30	315	Y	N
852	2	1	36	178	N	N
…	…	…	…	…	…	…

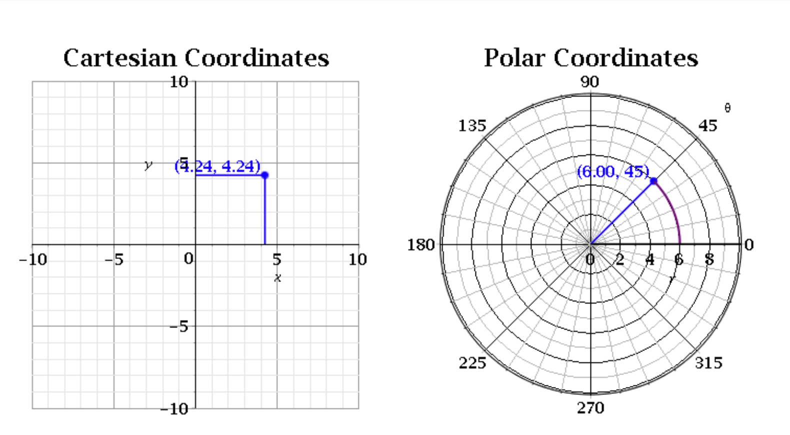


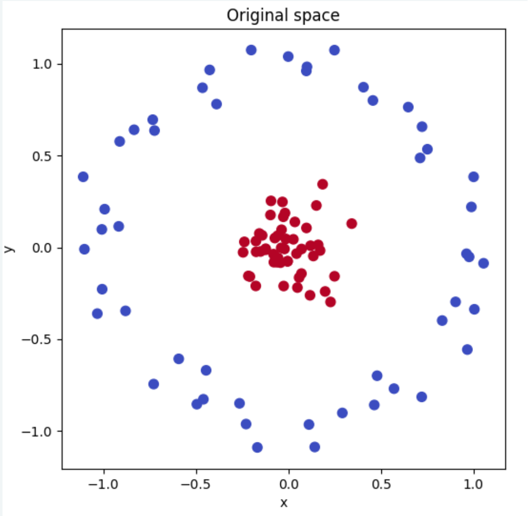
Add more columns to the dataset

Feature Engineering Example

Sometimes transforming the original data leads to better modeling results

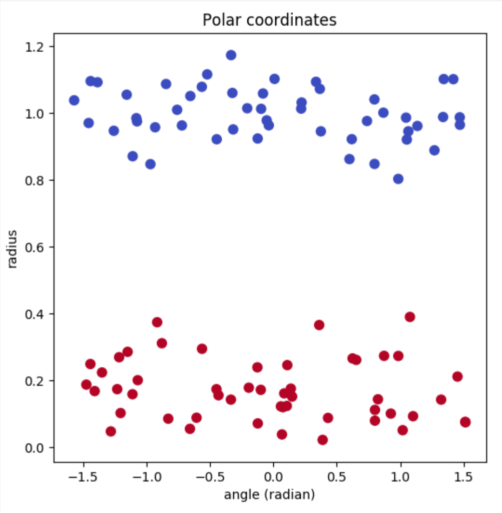
Euclidean to polar coor





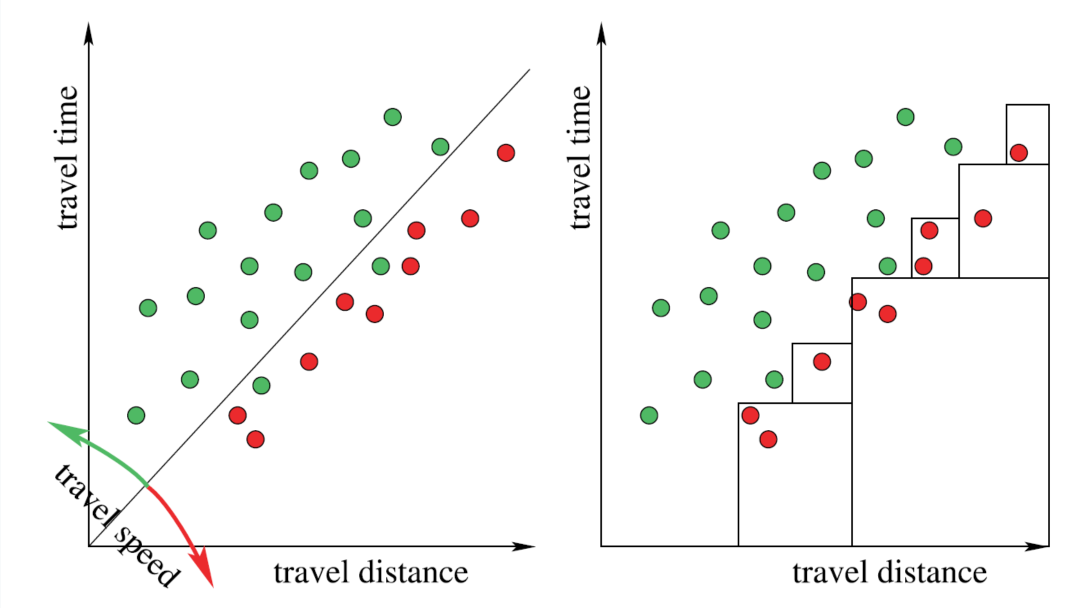
Radius $r = \sqrt{x^2 + y^2}$

Angle $\theta = Tan^{-1}(y/x)$



Feature Engineering Example

Offering new features that were difficult to represent by the original model



Example for the usefulness of derived features:
for a number of journeys, the travel time and distance are shown; the color indicates whether the driver was ticketed or not. Discriminating both classes with axis-parallel rectangles is laborious, but easy with a new attribute for travel speed

Feature Engineering in practice

Feature engineering includes all **transformation techniques** of existing attributes and and **construction** of new attributes that may (or may not) replace the original attributes

- Exploit domain knowledge to improve the model results

Example: Feature Engineering by Scale Conversion

Scale Conversion

- Categorical → Numerical: map categorical and ordinal values to a set of binary values
- Numerical → Categorical: **Discretization** (equal-width, equal-depth, V-optimal)

Example: Engineering Ratio Attributes

Feature Engineering refers to generating new features from the existing ones

Example: **Find the best workers in a company.**

Attributes available:

- the tasks, a worker has finished within each month,
- the number of hours he has worked each month,
- the number of hours that are normally needed to finish each task.

These attributes do *contain* information about the efficiency of the worker. But instead of using these three “raw” attributes, it might be more useful to define a new attribute ***efficiency***.

$$efficiency = \frac{hours\ actually\ spent\ to\ finish\ the\ tasks}{hours\ usually\ spent\ to\ finish\ the\ tasks}$$

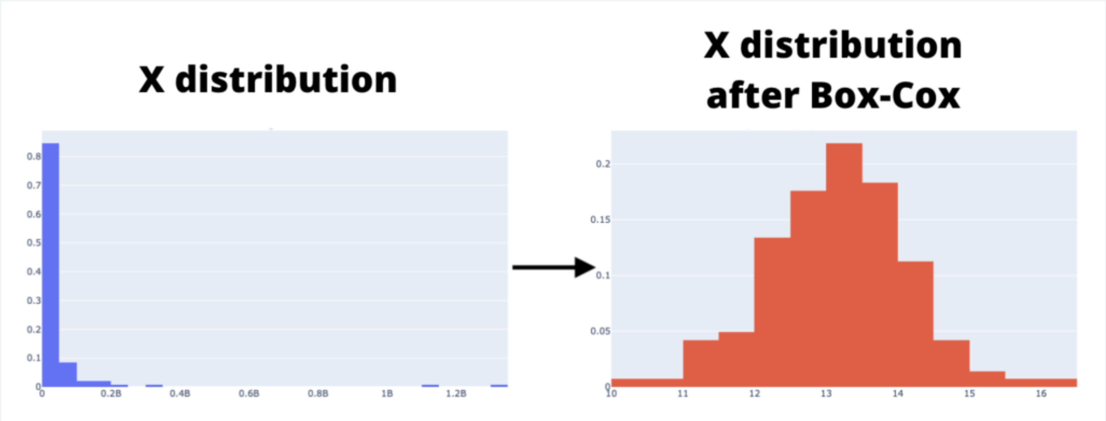
Example: Feature Engineering with Box-Cox Transform

Typical assumption: some variables obey a certain distribution (e.g. Gaussian)

- Transform the data to better approximate the distribution using the **power transform (Box-Cox Transform)**

$$y \mapsto \begin{cases} \dfrac{y^{\lambda}-1}{\lambda \overline{y}^{(\lambda-1)}} & if\ \lambda \neq 0 \\ \overline{y} \log y & if\ \lambda = 0 \end{cases}$$

Note: Only idea behind those techniques



Complex Data Types and Feature Extraction

Complex data types:

- Texts
- Graphs
- Images
- Molecules
- Other Objects

Especially for complex data types, **feature extraction is required**

- **Text data analysis.** Frequency of keyword, . . .
- **Time series data analysis.** Fourier or wavelet coefficients, . . .
- **Image data analysis.** Fourier or wavelet coefficients, . . .
- **Graph data analysis.** Number of vertices, number of edges, . . .

Contents

Data Integration

Vertical Data Integration (Concatenation)

- Unify database structures
- Remove duplicates

id	Last name	First name	Gender
p2	Mayer	Susan	F
p5	Smith	Walter	M
p7	Brown	Jane	F
…	…	…	…

+

Shopper id	Item id	Price
p2	i254	12.50
p5	i4245	1.99
p5	i32123	1.29
p5	i254	12.50
p5	i21435	5.99
p7	i254	12.50
…	…	…

Horizontal Data Integration (Join)

- Overrepresentation of items
- Data explosion

Item id	Price	Last name	First name	Gender
i254	12.50	Mayer	Susan	F
i4245	1.99	Smith	Walter	M
i32123	1.29	Smith	Walter	M
i254	12.50	Smith	Walter	M
i21435	5.99	Smith	Walter	M
i254	12.50	Brown	Jane	F
…	…	…	…	…

The two data sets on top contain information about customers and product purchases. The joint data set at the bottom combines these two tables. Note how we loose information about individual customers and how a lot of duplicate information is introduced. In reality this effect is, of course, far more dramatic

Contents

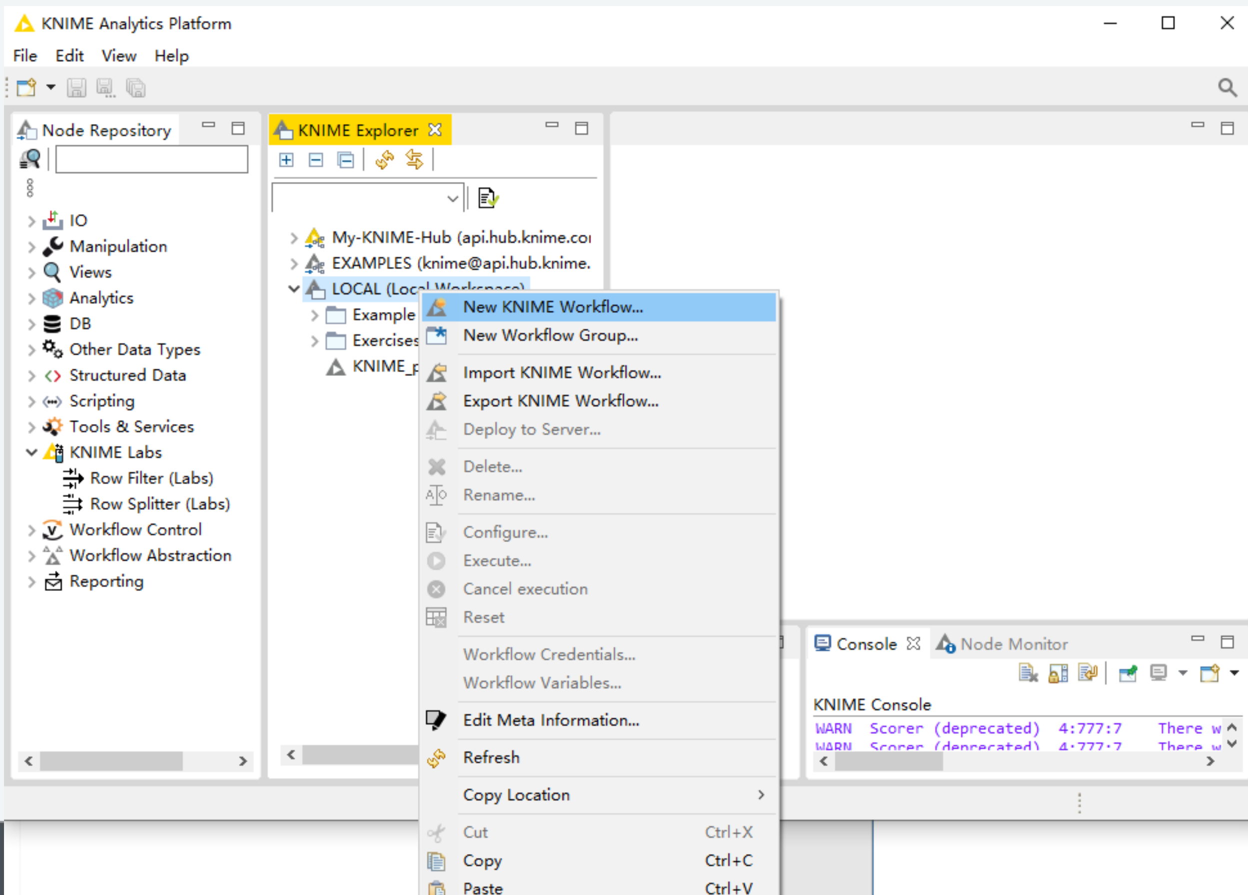
Lab: Visualization of sales data

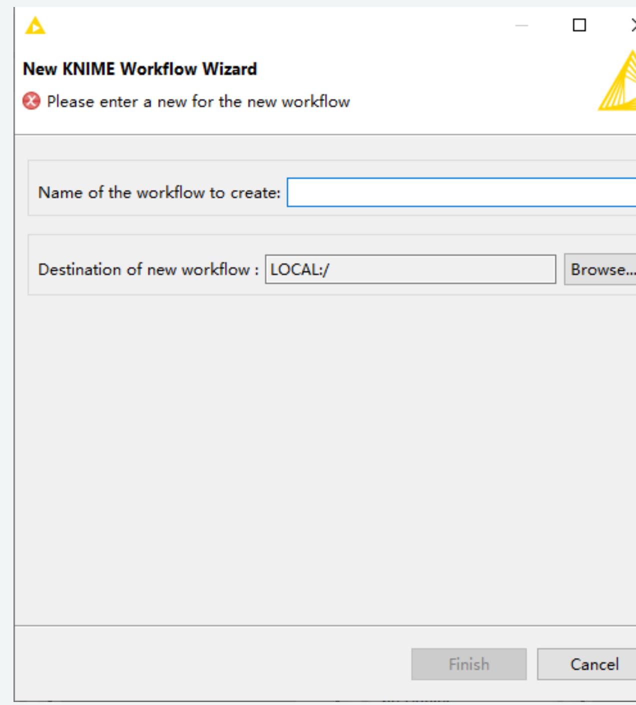
You will learn how to do basic data preparation in KNIME. You will learn:

1. How to read csv data file
2. How to filter columns and rows
3. How to visualize your results in different charts.

Step 1: Create a new KNIME workflow

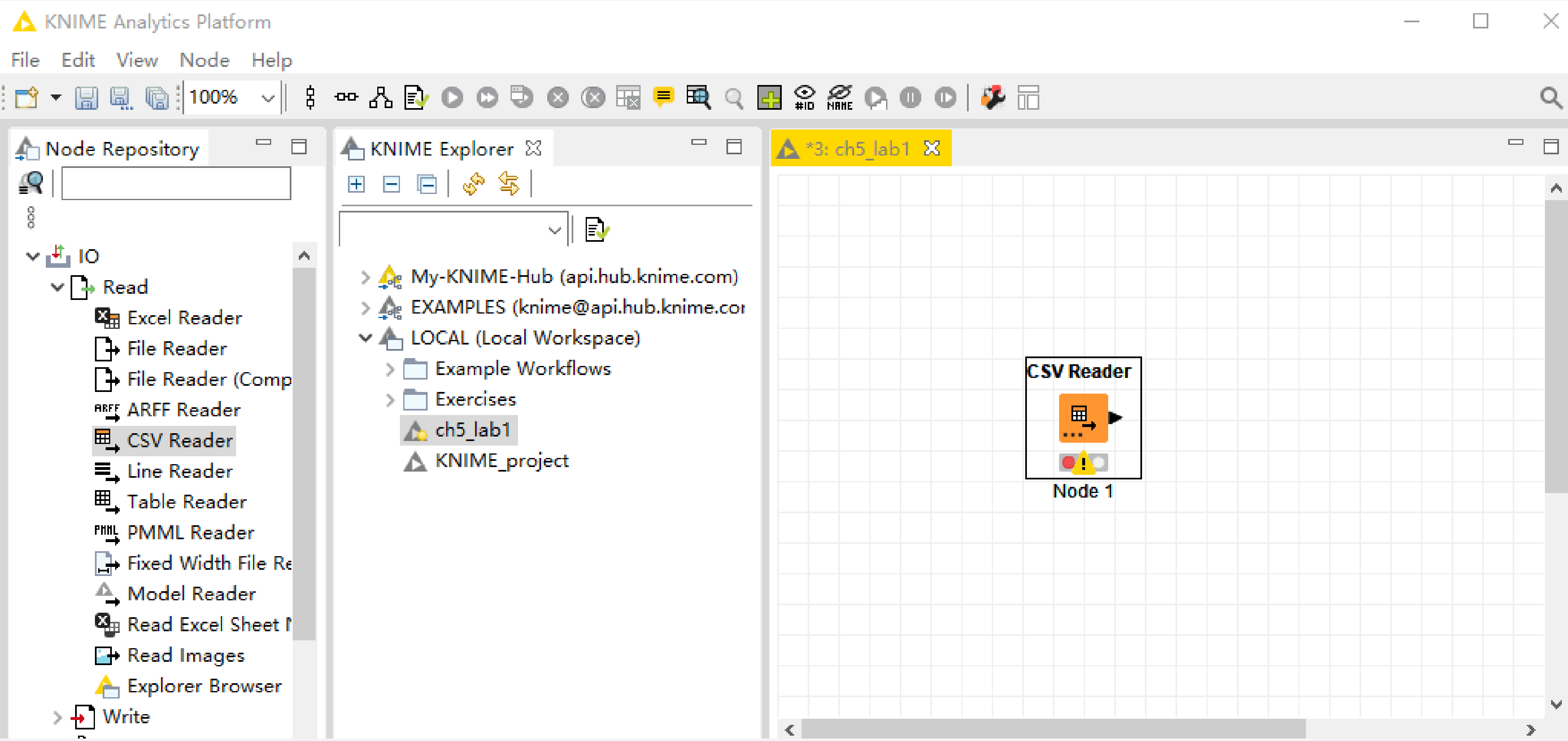
- In KNIME Explorer, under LOCAL menu, right click your mouse, it will pop up a window, select New KNIME Workflow menu
- After click the New Workflow menu, the following window will pop up. You need to provide a name for your workflow. Then click finish button.





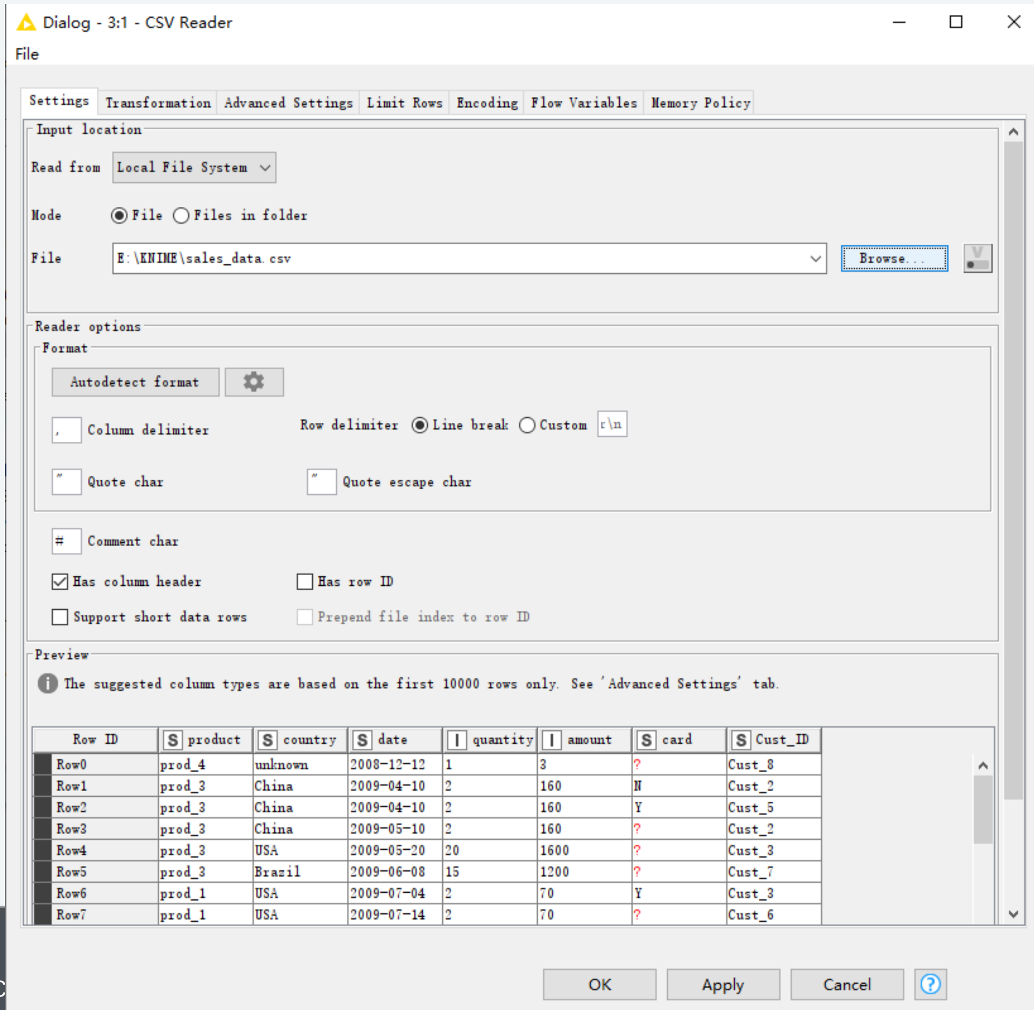
Step 2: Add a CSV Reader Node

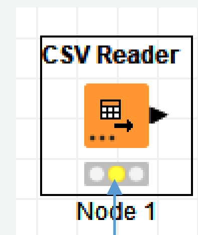
- An empty working space is shown, allowing you to drag and drop some nodes in side.
- In Node Repository window, find IO menu, select CSV Reader, drag and drop it into your empty working sheet.



Step 3: Configure CSV Reader Node

- Right click your mouse on the node. A window will pop up, click the configuration menu.
- You need to provide your csv data file location on your machine. For example, out data file is sales_data.csv, and located in E:/KNIME, you can type: E:/KNIME/sales_data.csv. The contents will displayed and click ok/apply button, and done. Your Node 1 becomes yellow indicting the data is ready.

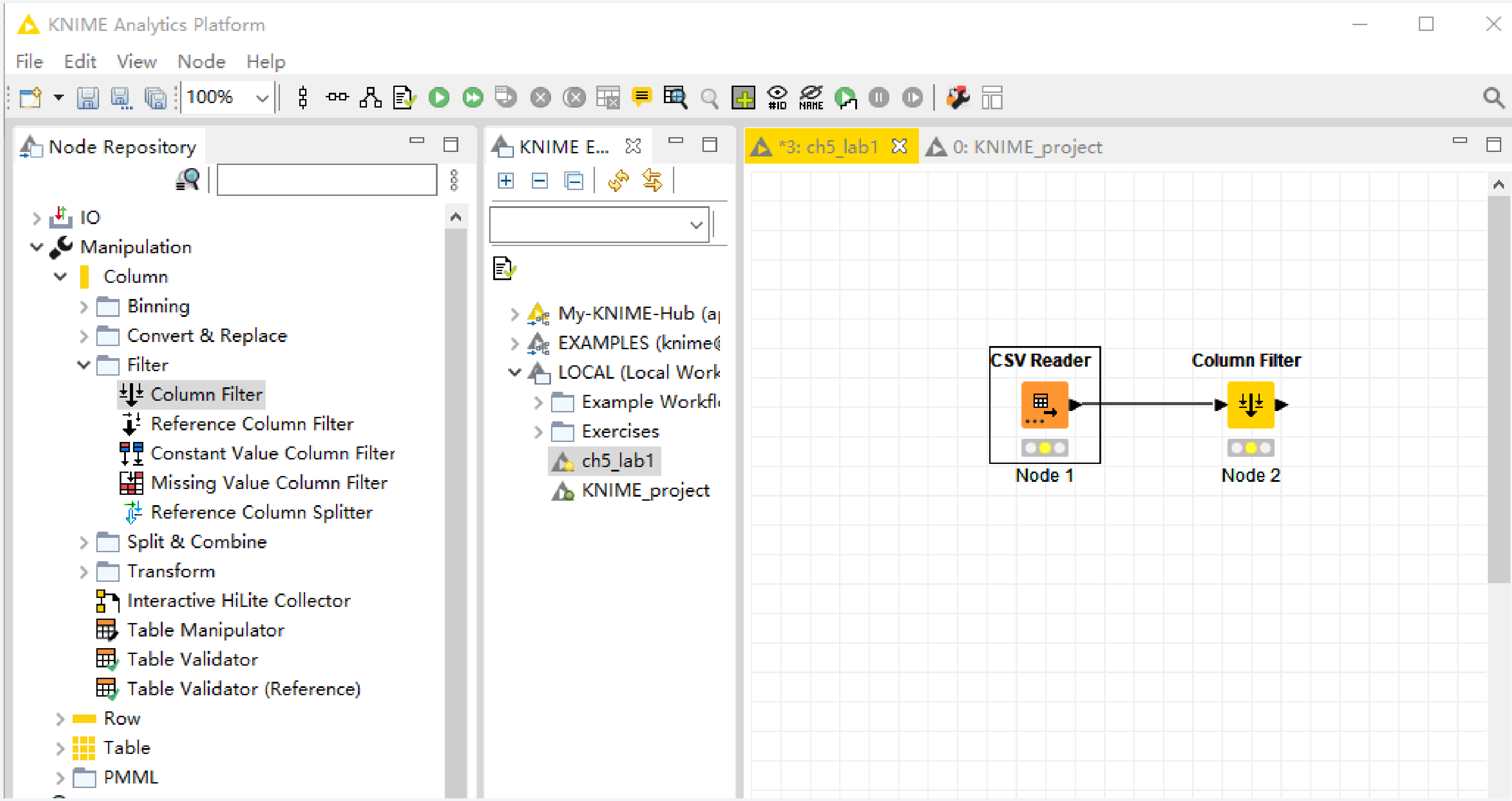




yellow

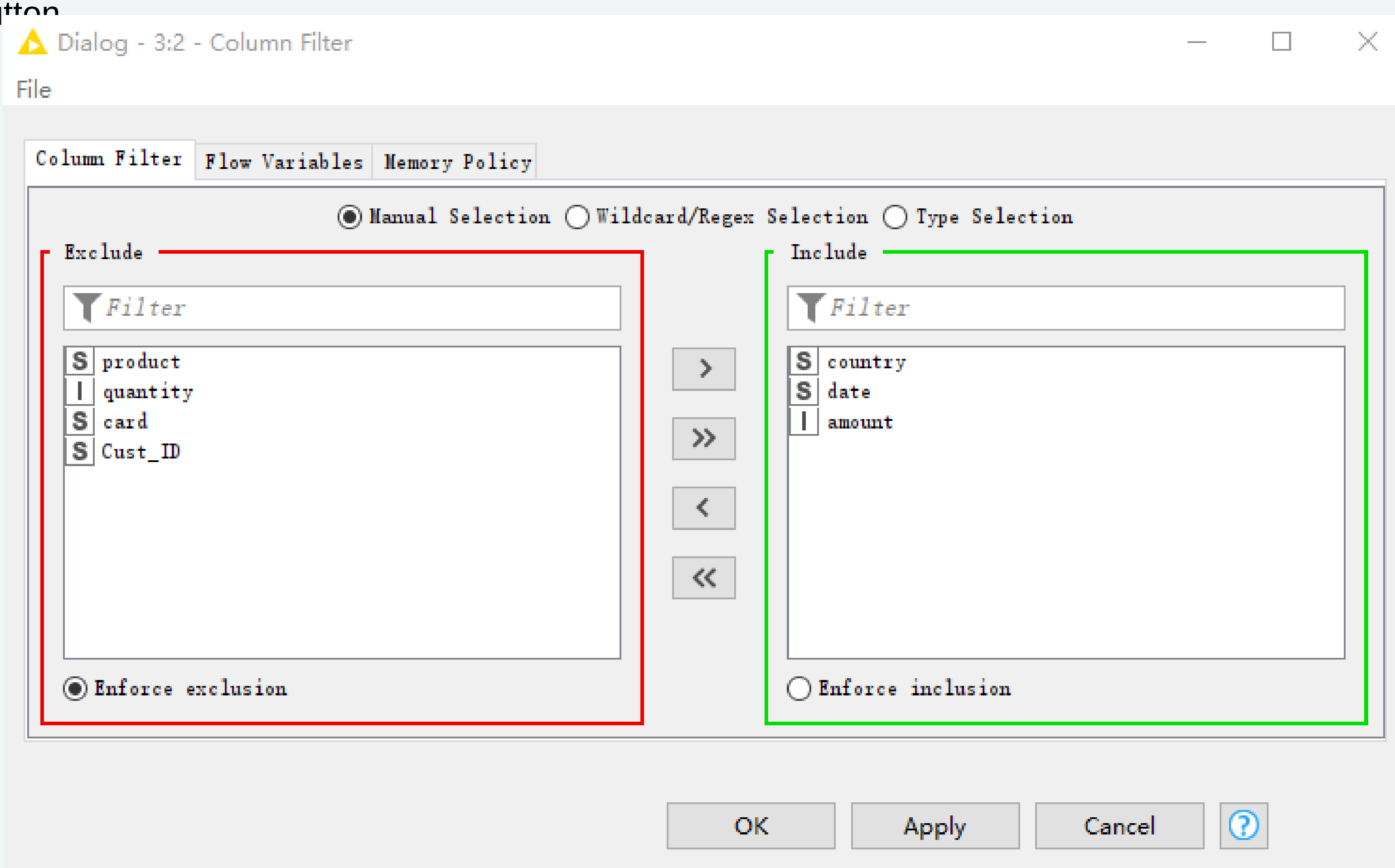
Step 4: Add a column Filter Node

- In Node Repository, find Column Filter under Column Filter of Manipulation menu. Drag and drop a Column Filter node into sheet, and connect CSV Reader to this Column Filter Node, as figure shows:



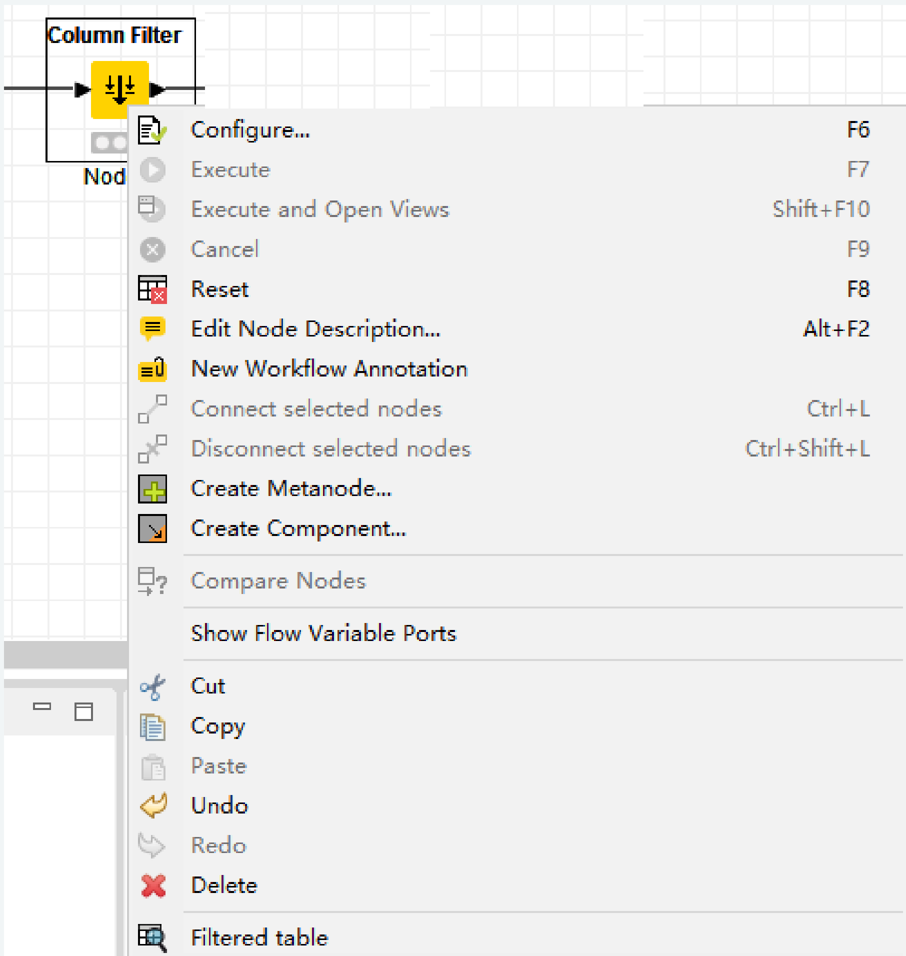
Step 5: Configure column Filter Node

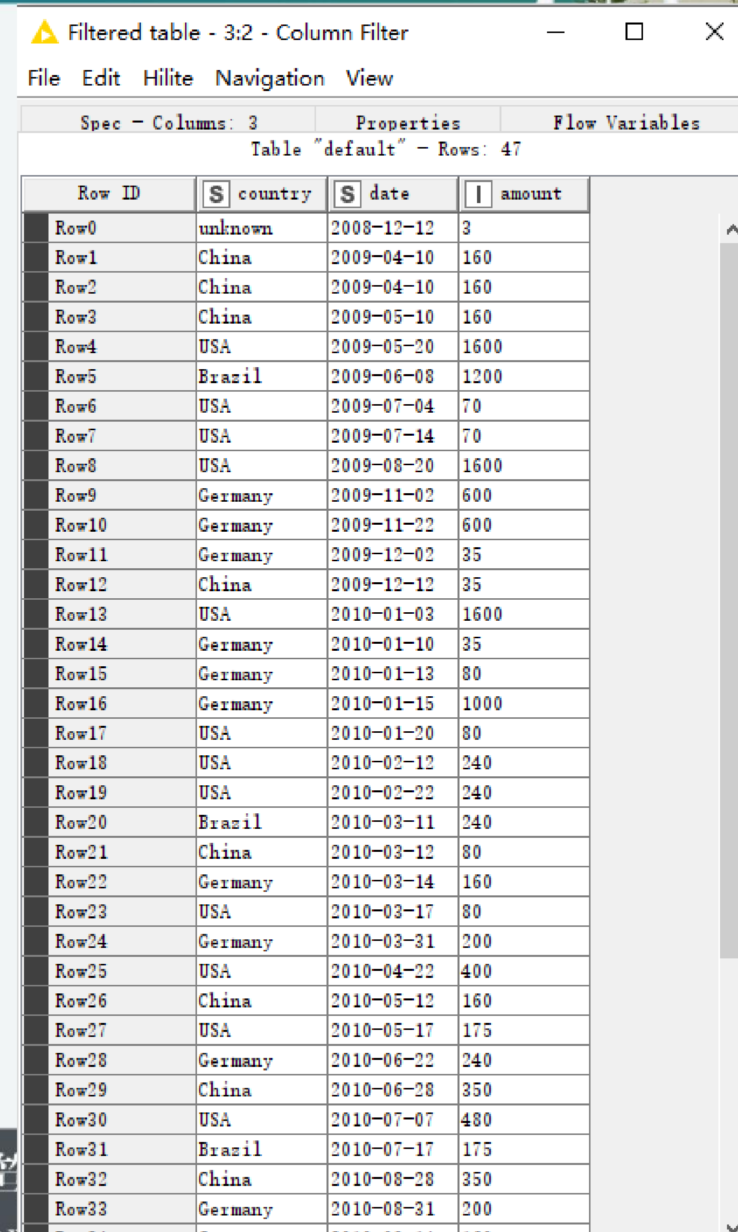
- Right click on node Colum Filter, a window pop up, click configure menu, the following configuration window pop up. We will exclude a few columns from right to left, as figure shown. Click Ok or Apply button



Step 6: Check the Filtered Data

- Right click the node Colum Filter, a window pop up, click Filtered table.

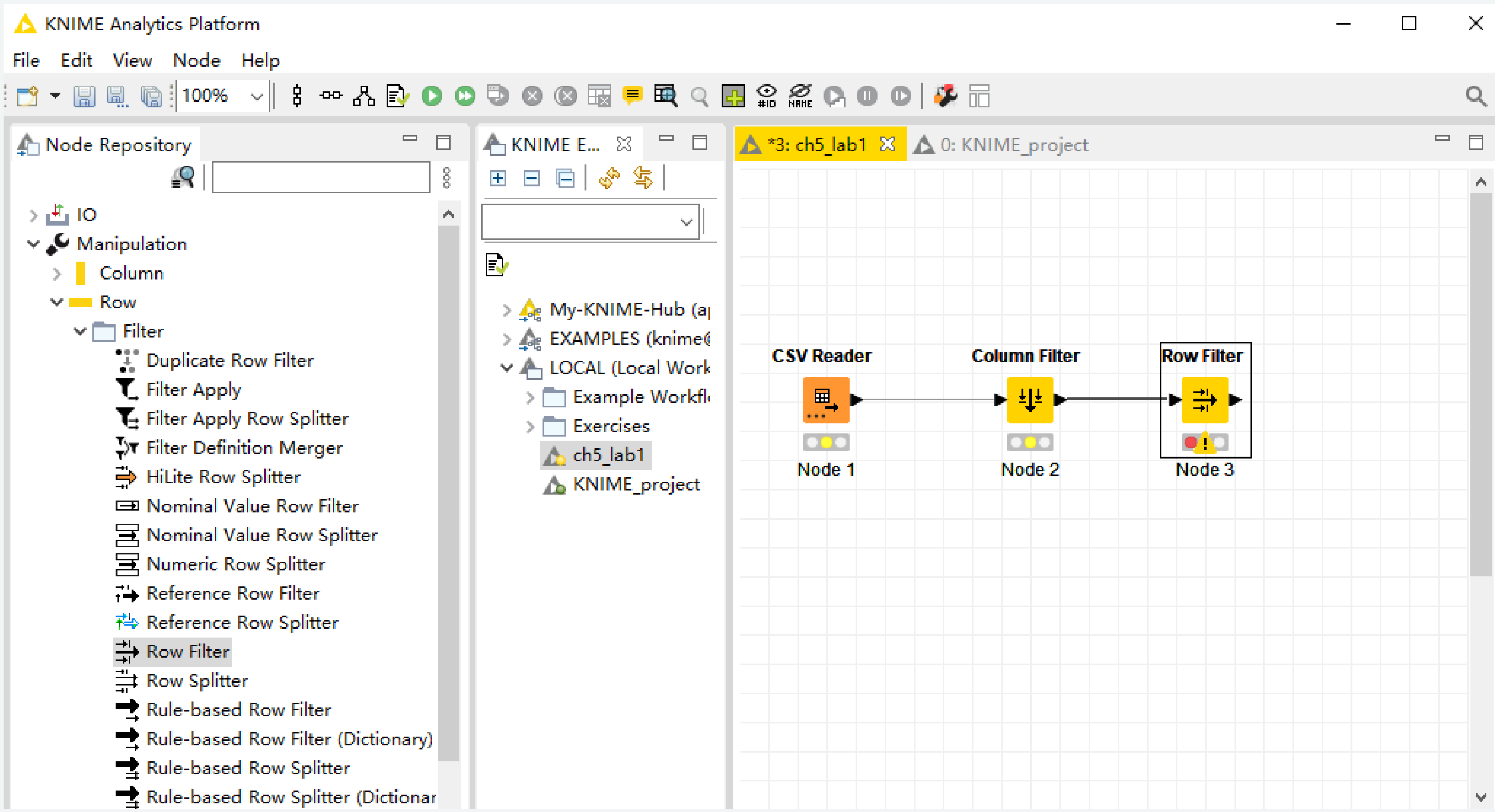




Row ID	S country	S date	I amount
Row0	unknown	2008-12-12	3
Row1	China	2009-04-10	160
Row2	China	2009-04-10	160
Row3	China	2009-05-10	160
Row4	USA	2009-05-20	1600
Row5	Brazil	2009-06-08	1200
Row6	USA	2009-07-04	70
Row7	USA	2009-07-14	70
Row8	USA	2009-08-20	1600
Row9	Germany	2009-11-02	600
Row10	Germany	2009-11-22	600
Row11	Germany	2009-12-02	35
Row12	China	2009-12-12	35
Row13	USA	2010-01-03	1600
Row14	Germany	2010-01-10	35
Row15	Germany	2010-01-13	80
Row16	Germany	2010-01-15	1000
Row17	USA	2010-01-20	80
Row18	USA	2010-02-12	240
Row19	USA	2010-02-22	240
Row20	Brazil	2010-03-11	240
Row21	China	2010-03-12	80
Row22	Germany	2010-03-14	160
Row23	USA	2010-03-17	80
Row24	Germany	2010-03-31	200
Row25	USA	2010-04-22	400
Row26	China	2010-05-12	160
Row27	USA	2010-05-17	175
Row28	Germany	2010-06-22	240
Row29	China	2010-06-28	350
Row30	USA	2010-07-07	480
Row31	Brazil	2010-07-17	175
Row32	China	2010-08-28	350
Row33	Germany	2010-08-31	200
Row34	Germany	2010-09-14	160

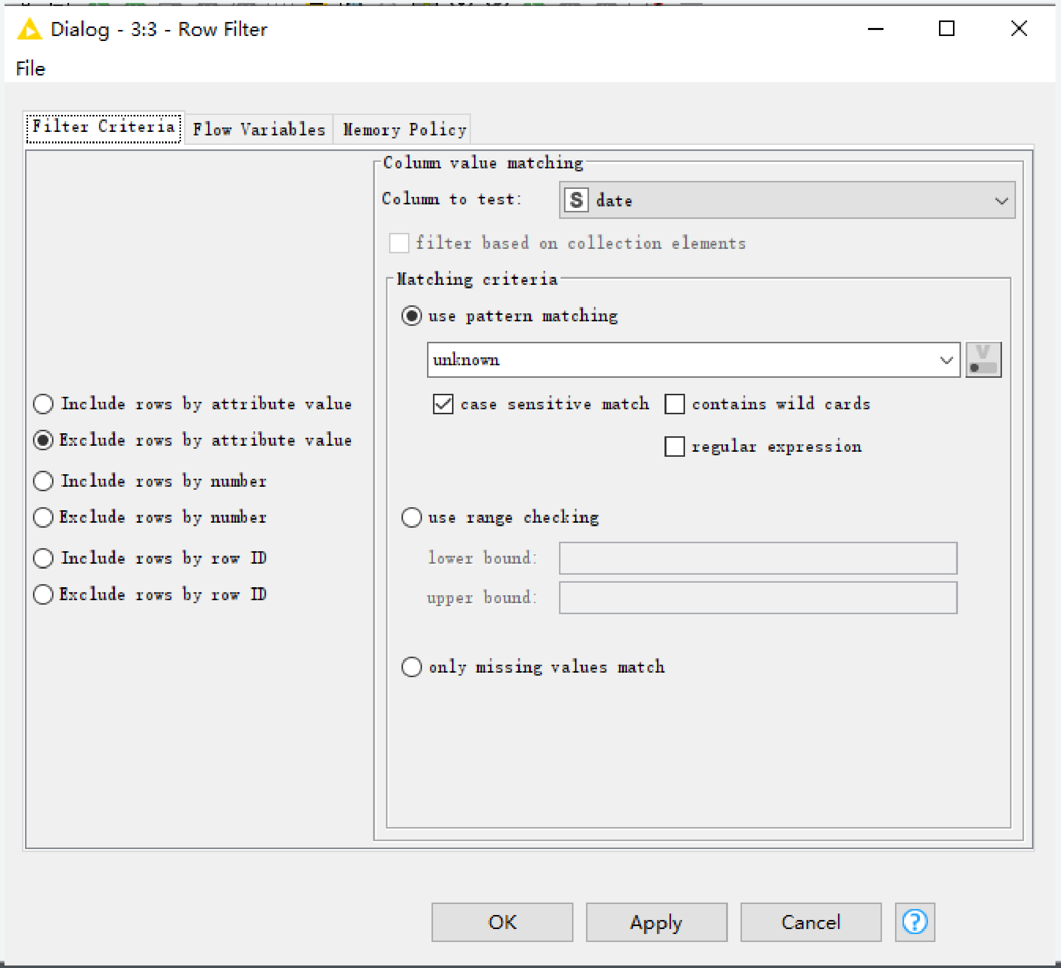
Step 7: Add a Row Filter Node

- In Node Repository, find Row Filter under Row Filter of Manipulation menu. Drag and drop a Row Filter node into sheet, and connect Column Filter Node 2 to Row Filter Node 3, as figure shows:



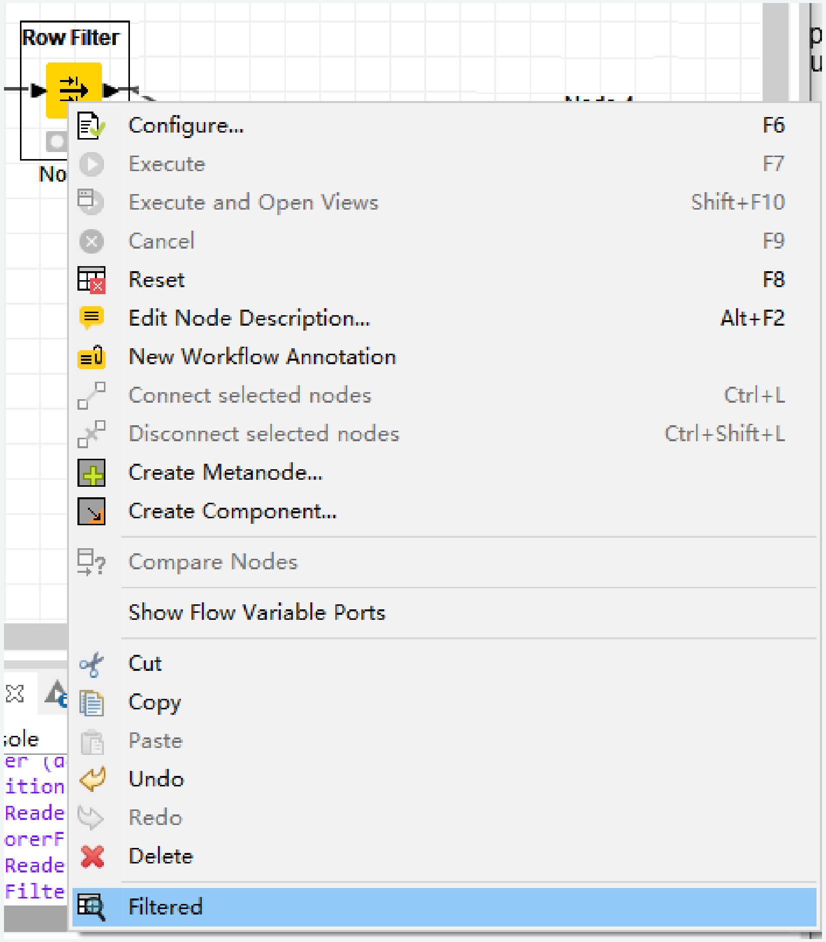
Step 8: Configure the Row Filter Node

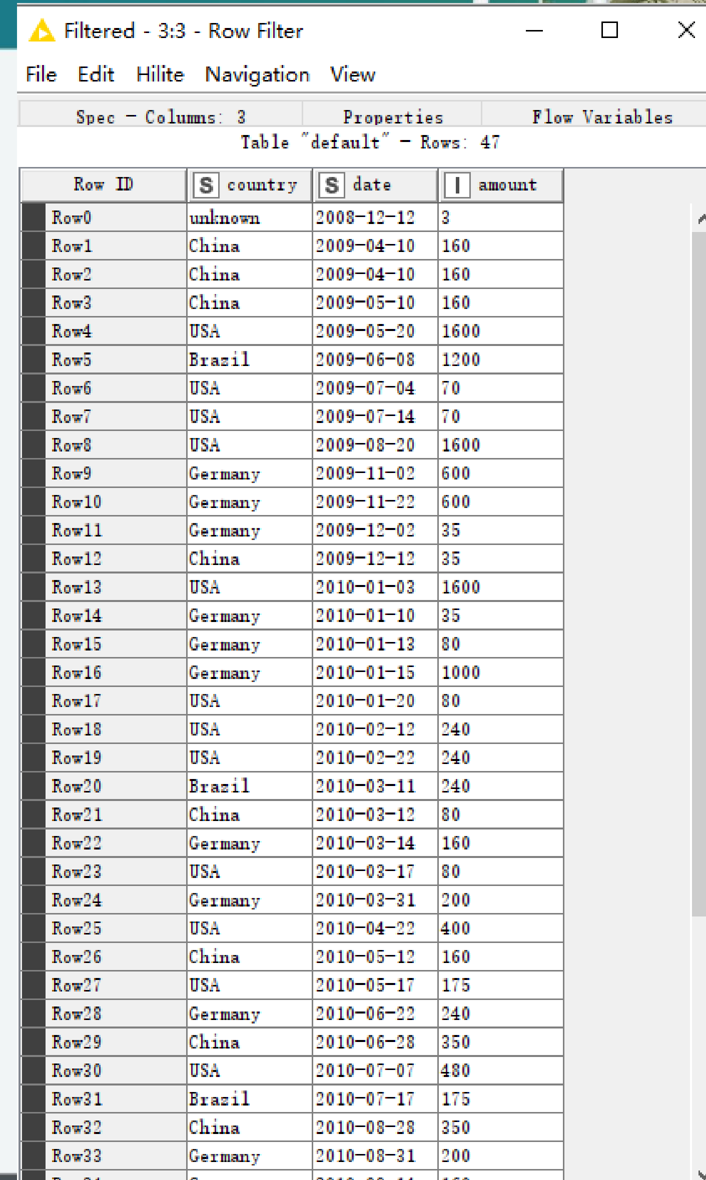
- Right click node Row Filter, a window pop up, click configure menu, the following configuration window pop up. We will exclude rows by attribute value for unknow for all columns (amount, country, date). Click OK or Apply.



Step 9: Check the data output of the Row Filter Node

- Right click node Row Filter, a window pop up, click Filtered menu.

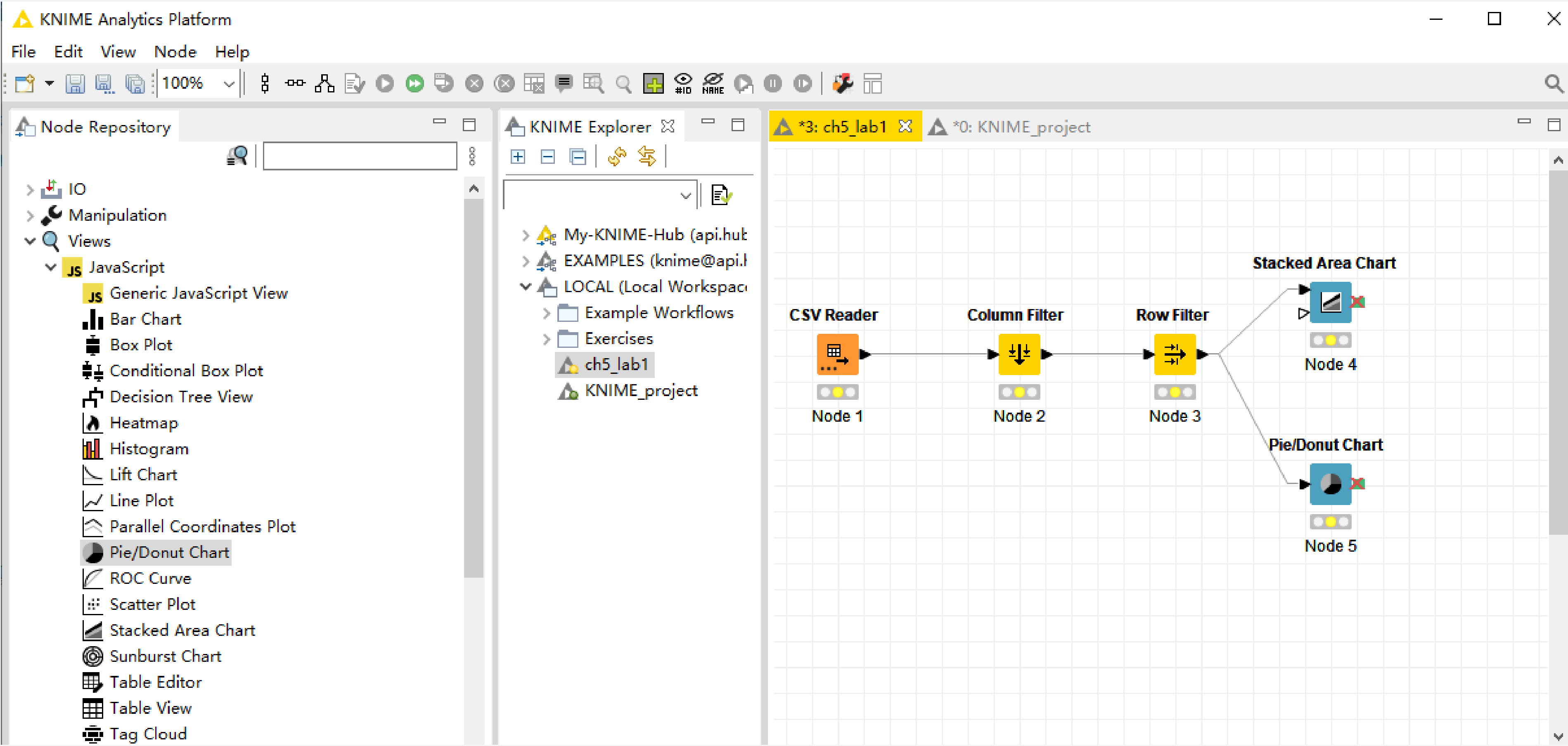




Row ID	S country	S date	I amount
Row0	unknown	2008-12-12	3
Row1	China	2009-04-10	160
Row2	China	2009-04-10	160
Row3	China	2009-05-10	160
Row4	USA	2009-05-20	1600
Row5	Brazil	2009-06-08	1200
Row6	USA	2009-07-04	70
Row7	USA	2009-07-14	70
Row8	USA	2009-08-20	1600
Row9	Germany	2009-11-02	600
Row10	Germany	2009-11-22	600
Row11	Germany	2009-12-02	35
Row12	China	2009-12-12	35
Row13	USA	2010-01-03	1600
Row14	Germany	2010-01-10	35
Row15	Germany	2010-01-13	80
Row16	Germany	2010-01-15	1000
Row17	USA	2010-01-20	80
Row18	USA	2010-02-12	240
Row19	USA	2010-02-22	240
Row20	Brazil	2010-03-11	240
Row21	China	2010-03-12	80
Row22	Germany	2010-03-14	160
Row23	USA	2010-03-17	80
Row24	Germany	2010-03-31	200
Row25	USA	2010-04-22	400
Row26	China	2010-05-12	160
Row27	USA	2010-05-17	175
Row28	Germany	2010-06-22	240
Row29	China	2010-06-28	350
Row30	USA	2010-07-07	480
Row31	Brazil	2010-07-17	175
Row32	China	2010-08-28	350
Row33	Germany	2010-08-31	200
Row34	Germany	2010-09-14	160

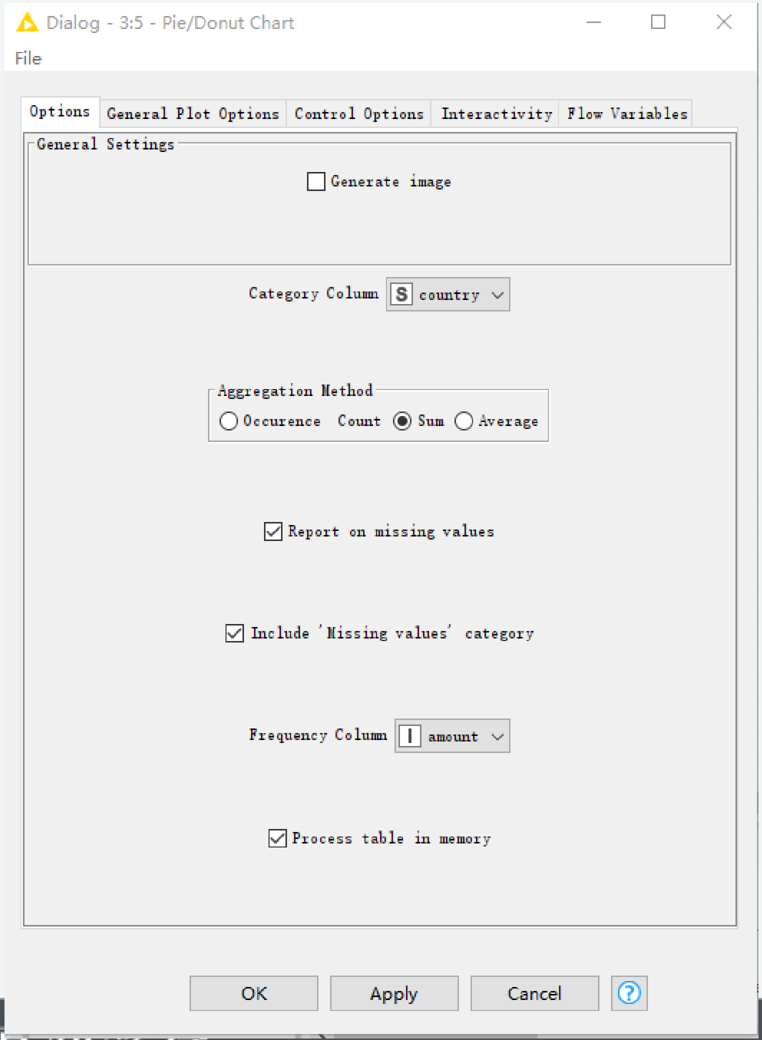
Step 10: Add a Stacked Area Chart Node and Pie/Donut Chart Node

- In Node Repository, find Stacked Area Chart and Pie/Donut Chart under JavaScript of Views menu. Drag and drop a Stacked Area Chart node and a Pie/Donut Chart Node into sheet, and connect Row Filter Node 3 to these two Nodes, as figure shows. We are ready to go.



Step 11: Configure the Chart Nodes

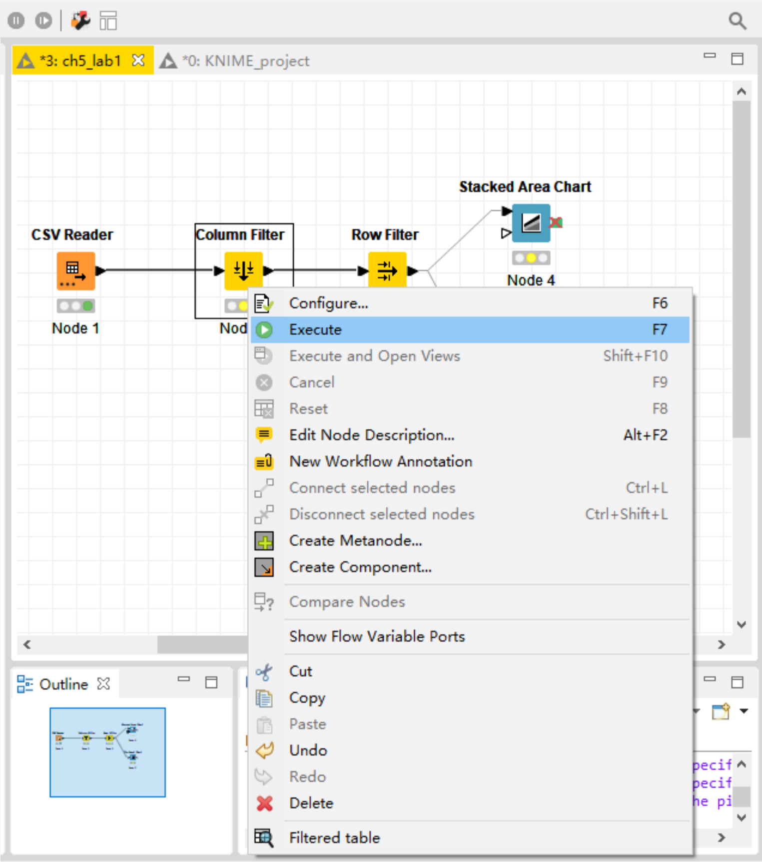
- Right Stacked Area Chart Node, We don't need to configure anything this time. But for Pie/Donut Chart, we configure as figure shows. Click OK or Apply.

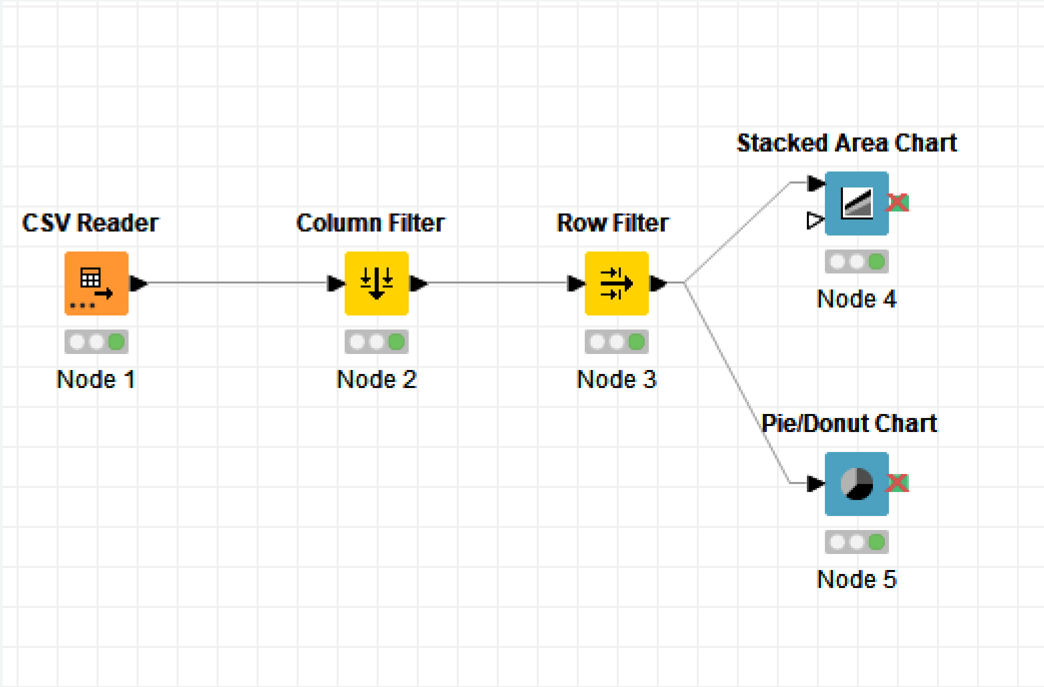


博文雅志 真知笃行
In knowledge and in deeds, unto the whole person

Step 12: Run each node one by one

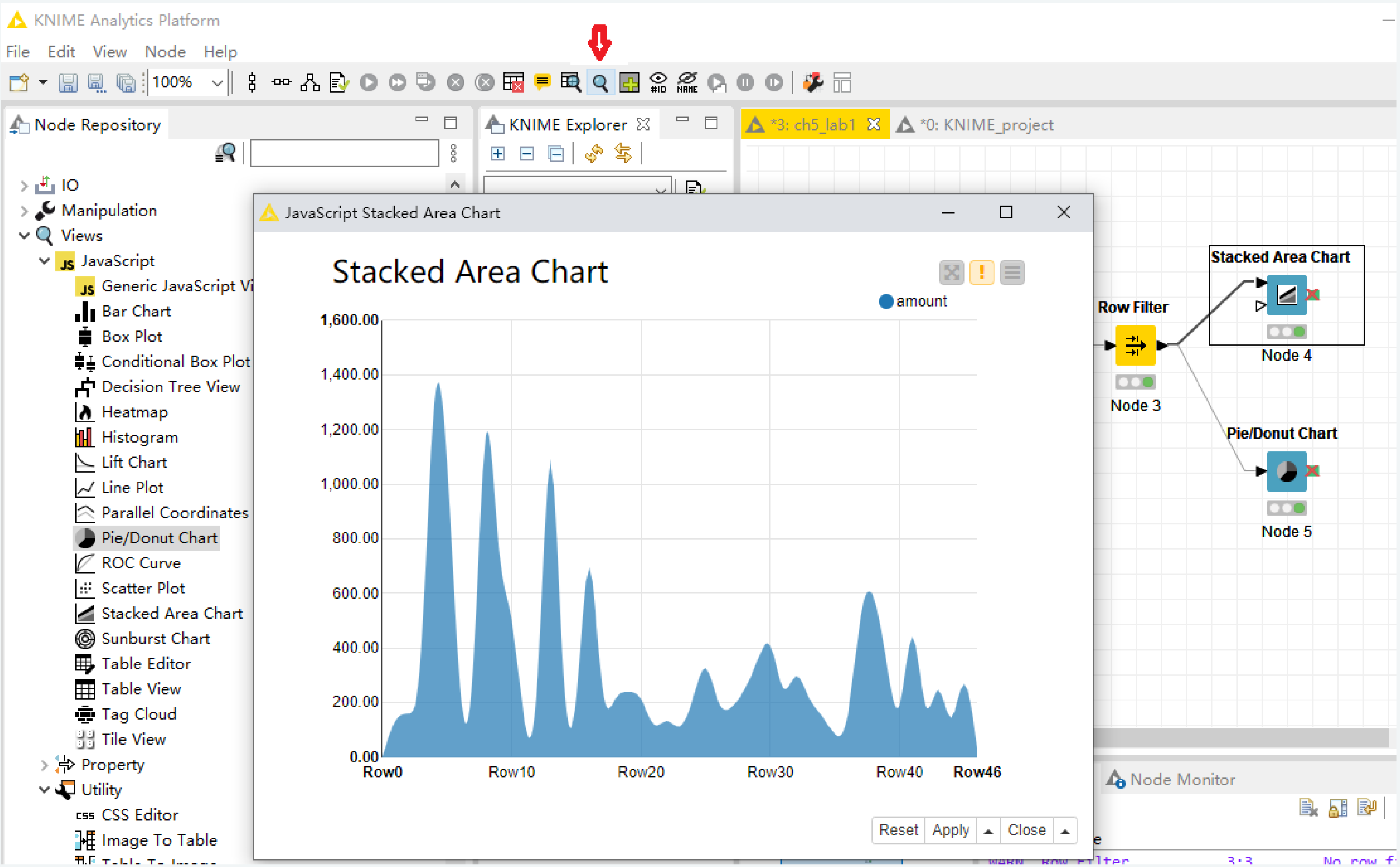
- Right click each Node, from the pop up menu, click Execute, the Node should become green.
- Eventually, all Nodes should be green.





Step 13: Check the Stacked Area Chart

- Click the Stacked Area Chart Node first, then click the toolbar, the Stacked Area Chart will pop up.



博文雅志 真知笃行

In knowledge and in deeds, unto the whole person

Step 14: Check the Pie Chart

- Click the Pie/Donut Chart Node first, then click the toolbar, the Pie Chart will pop up.