Business Data Mining
Semester 2, 2019

Lecture 4
How to Conduct Data Understanding Phase?

Kim, Yang Sok

Dept. of MIS, Keimyung University

Table of Content

- **Introduction**
- **Tasks & Outputs**
  - Gathering data
  - Describing Data
  - Exploring Data
  - Verifying Data Quality
- **Exercises**
  - Exercise 1: Write a data description report
  - Exercise 2: Loading Dataset with Rapidminer
  - Exercise 3: Exploring Data with Rapidminer
- **Conclusion**

Introduction

Introduction

- **In the Data Understanding phase of CRISP-DM, you obtain data and verify that it is appropriate for your needs.**
  - You might identify issues that cause you to return to business understanding and revise your plan.
  - You may even discover flaws in your business understanding, another reason to rethink goals and plans.

- **The Data Understanding phase includes four tasks. These are**
  - Gathering data
  - Describing data
  - Exploring data
  - Verifying data quality

Tasks & Outputs

Tasks & Outputs

Gathering Data - Tasks

- **The Data Understanding phase requires you to acquire the data listed in the project resources.**

- **This initial collection includes data loading, if this is necessary for data understanding.**
  - For example, if you use a specific tool for data understanding, it makes perfect sense to load your data into this tool. In this lecture, we will use Rapidminer, so it is necessary load data with Rapidminer in this phase.
  - If you acquire **multiple data sources** then you need to consider how and when you're going to **integrate these**.

Tasks & Outputs

Gathering Data - Outputs

- **Just one deliverable exists for this task: the initial data collection report.**
- **You may do a lot of work to assemble the data you need before you can write this report. First, you will make your plan, as follows:**
  - **Outline data requirements**: Create a list of the types of data necessary to address the data mining goals. Expand the list with details such as the required time range and data formats.
  - **Verify data availability**: Confirm that the required data exists, and that you can use it. If some of the data you want is unavailable, decide how you will address that issue. Consider alternatives such as
    - Substituting with an alternative data source
    - Narrowing the scope of the project
    - Gathering new data
  - **Define selection criteria**: Identify the specific data sources (databases, files, documents, and so on.) you will use. Within those sources, specify the tables, fields, and case ranges that are relevant to this project.

Tasks & Outputs

Gathering Data - Outputs

- **Once you’ve gone through these steps, you must actually obtain the data.**
- **At this stage, import the data into the data-mining platform you’ll be using for the project to confirm that it is possible to do so and that you understand the process.**
- **In the course of this trial you may discover software (or hardware) limitations you had not anticipated, such as**
  - Limits on the number of cases or fields, or on the amount of memory you may use
  - Inability to read the data formats of your sources
  - Difficulty dealing with imperfections in the data (for example, you might encounter products that won’t import or analyze incomplete datasets)

Gathering Data - Outputs

- **Finally, summarize the gathering process in a report.**
- **The report should describe your requirements, and explain in some detail exactly what data you have gathered and from what sources.**
- **Here you confirm that you have actually obtained the data and that it is compatible with your data-mining platform.**
  - If you have run into difficulties, you'll explain what they were and how you have addressed them (using alternative sources, revising plans, changing formats).

Tasks & Outputs

Describing Data - Tasks

- **Now that you have data, prepare a general description of what you have.**
- **Examine the “gross” or “surface” properties of the acquired data and report on the results.**

Task & Outputs

Describing Data - Outputs

- **Data description report**
  - Describe the data that has been acquired including its format, its quantity (for example, the number of records and fields in each table), the identities of the fields and any other surface features which have been discovered.
  - Evaluate whether the data acquired satisfies your requirements.

Tasks & Outputs

Exploring Data - Tasks

- **During this stage you'll address data mining questions using querying, data visualization and reporting techniques. These may include:**
    - Distribution of key attributes (for example, the target attribute of a prediction task)
    - Relationships between pairs or small numbers of attributes
    - Results of simple aggregations
    - Properties of significant sub-populations
    - Simple statistical analyses

- **These analyses may directly address your data mining goals. They may also contribute to or refine the data description and quality reports, and feed into the transformation and other data preparation steps needed for further analysis.**

Tasks & Outputs

Exploring Data - Outputs

- **The deliverable for this task is the data exploration report.**

- **This report describe results of your data exploration, including first findings or initial hypothesis and their impact on the remainder of the project.**

- **If appropriate you could include graphs and plots here to indicate data characteristics that suggest further examination of interesting data subsets.**

- **This report should include a more detailed description of the data than the data description report, including distributions, summaries, and any signs of data quality problems.**

Tasks & Outputs

Verifying Data Quality - Tasks

- **You have the data and you've examined it, and now you have to determine whether it's good enough to support your goals.**
- **You will often have some quality problem to address yet still be able to move forward, but at times the data quality is so poor that it cannot support your plan and you'll have to look for alternatives.**
- **Some of the worst data problems would include**
  - The data you need doesn't exist. (Did it never exist, or was it discarded? Can this data be collected and saved for future use?)
  - It exists, but you can't have it. (Can this restriction be overcome?)
  - You find severe **data quality issues**.
    - Is the data complete (does it cover all the cases required)?
    - Is it correct, or does it contain errors and, if there are errors, how common are they?
    - Are there missing values in the data? If so, how are they represented, where do they occur, and how common are they?

Tasks & Outputs

Verifying Data Quality - Outputs

- **The deliverable for this task is the data quality report.**

- **This summarizes the data that you have, minor and major quality issues that you have found, and possible remedies for quality problems or alternatives (such as using an alternative data resource).**

- **If you are facing any really serious data quality issues and can't identify an adequate solution, you may have to recommend reconsidering goals or plans.**

Exercise 1:
Write a data description report

Exercise - Write a Data Description Report

Template

These three items need knowledge on data mining. So now you may skip these three items

Name	Name of dataset
Contributor	Dataset creator
Business Objective	Explain business objectives
Data Mining Objectives	**Explain data mining objectives**
Data Mining Tasks	**Prediction, classification, clustering, etc.**
Performance Measures	**Accuracy, Precision, Recall, F-Measure, RMSE, MAE, R2, …**
Dataset Characteristics	Multivariate, univariate, sequential, time-series, text, other
Number of Examples	Number of examples collected
Area	Life Sciences, Physical Sciences, CS / Engineering, Social Sciences, Business, Game, Other
Attribute Characteristics	Numerical, categorical, text, mixed
Number of Attributes	Number of attributes
Missing Values?	Yes/No
Version	1.0, …
Date of Creation	2019-08-05
Attribute Definition	

Name	Value Type	Role	Description	Allowable Values

Exercise - Write a Data Description Report

Template - Dataset Characteristics

- **Univariate data**
  - This type of data consists of only one variable.
  - The analysis of univariate data is thus the simplest form of analysis since the information deals with only one quantity that changes.
  - It does not deal with causes or relationships and the main purpose of the analysis is to describe the data and find patterns that exist within it.
  - The example of a univariate data can be height.
- **Bivariate data**
  - This type of data involves two different variables.
  - The analysis of this type of data deals with causes and relationships and the analysis is done to find out the relationship among the two variables.
  - Example of bivariate data can be temperature and ice cream sales in summer season.
- **Multivariate data**
  - When the data involves three or more variables, it is categorized under multivariate.
  - Example of this type of data is suppose an advertiser wants to compare the popularity of four advertisements on a website, then their click rates could be measured for both men and women and relationships between variables can then be examined.

Exercise - Write a Data Description Report

Template - Dataset Characteristics

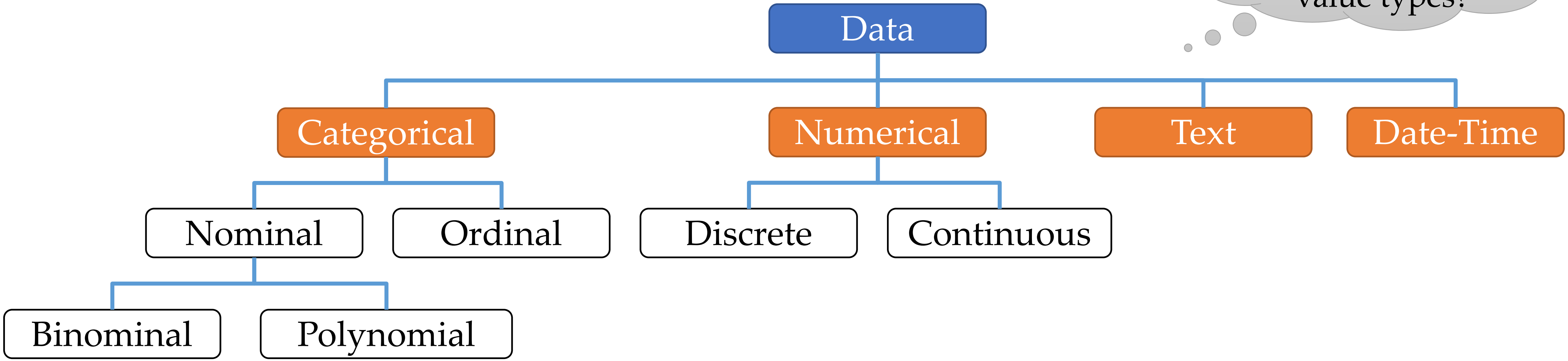
- **Sequential Data:**
  - The order of the data matters, but the time stamp is irrelevant or it doesn’t matter. (Example: DNA sequence. As you see the concept of time is irrelevant, so the order is not temporal.)
- **Temporal Sequence:**
  - In addition to the order of data, the time stamp also matters. (Example: Data collected from customers’ shopping behavior, considering their transaction time stamp as the temporal dimension.)
- **Time Series:**
  - The data is in order, with a fixed time-difference between occurrence of successive data points. (Example: Time series of the temperature of a surface being recorded every 120 seconds.)

Exercise - Write a Data Description Report

Template – Attribute Definition

- **Name is a unique identifier, so names should not duplicate.**
- **Value type can be classified into four types**

Why I need to concern data value types?



Exercise - Write a Data Description Report

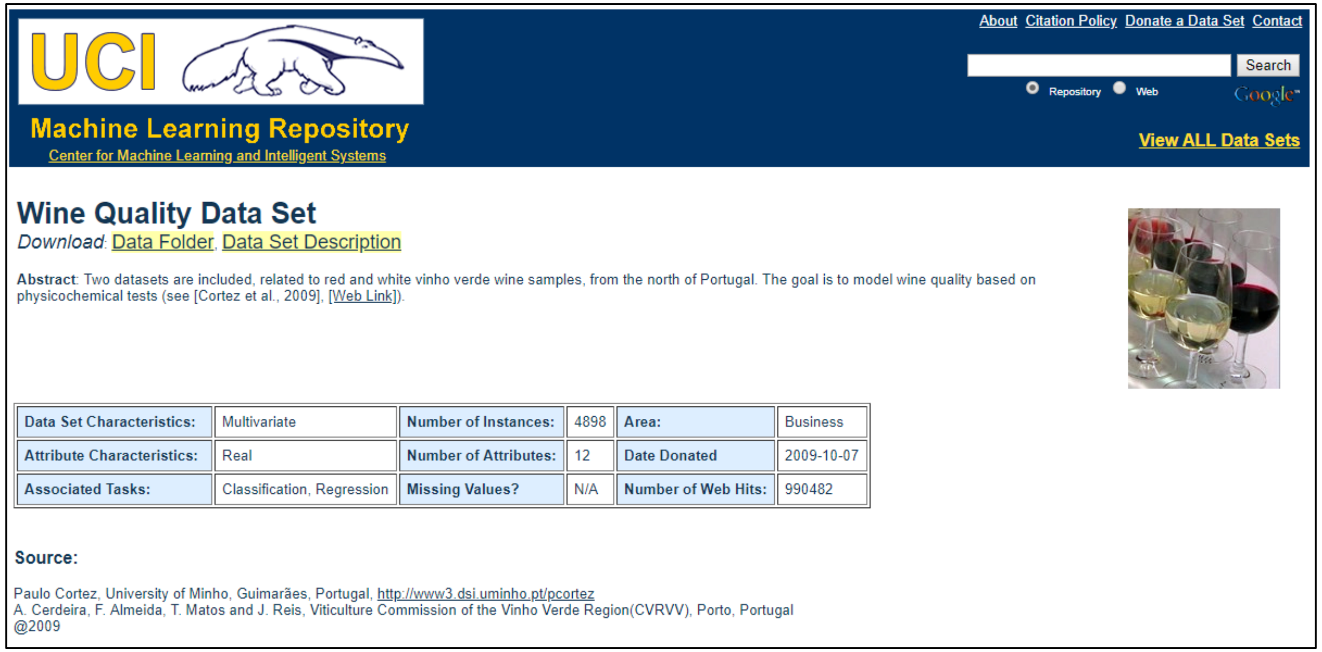
Template – Attribute Definition

- **Attribute is one of following roles: (general) attribute, label, id, weight, batch, cluster, prediction, outlier, cost, base value**
  - (general) attribute is an attribute to describe the example.
    - When data set load by read operators, Rapidminer regards all attributes as this type.
  - id is an attribute used to identify individual example.
  - label is an attribute that should be predicted by the model.
    - For predicting problems (classification and regression), an example must have a label attribute.

- **Allowable Values are defined by min and max values for the numerical attribute value type and by a list of values for the categorical attribute value type.**

Exercise - Write a Data Description Report Template

- Go to https://archive.ics.uci.edu/ml/datasets/wine+quality


UCI
Machine Learning Repository
Center for Machine Learning and Intelligent Systems
About Citation Policy Donate a Data Set Contact
Search
Repository
Web
Google
View ALL Data Sets

Wine Quality Data Set

Download: Data Folder, Data Set Description

Abstract: Two datasets are included, related to red and white vinho verde wine samples, from the north of Portugal. The goal is to model wine quality based on physicochemical tests (see [Cortez et al., 2009], [Web Link]).

Data Set Characteristics:	Multivariate	**Number of Instances:**	4898	**Area:**	Business
Attribute Characteristics:	Real	**Number of Attributes:**	12	**Date Donated**	2009-10-07
Associated Tasks:	Classification, Regression	**Missing Values?**	N/A	**Number of Web Hits:**	990482

Source:

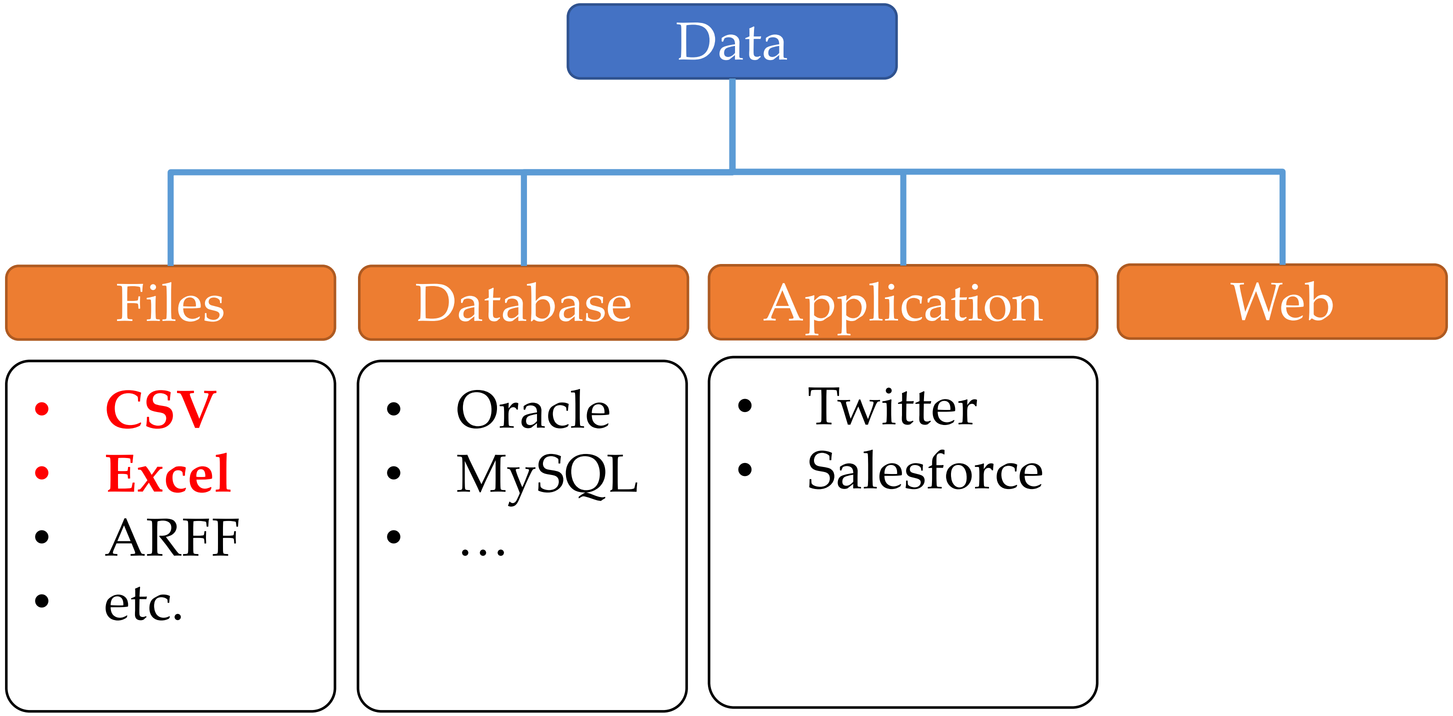
Paulo Cortez, University of Minho, Guimarães, Portugal, http://www3.dsi.uminho.pt/pcortez
A. Cerdeira, F. Almeida, T. Matos and J. Reis, Viticulture Commission of the Vinho Verde Region(CVRVV), Porto, Portugal
@2009

- Use information in the web page, the paper, and the dataset in order to write the data description report for the Wine Quality Prediction project.

Exercise 2: Loading Dataset with Rapidminer

Data Sources

- **Data can be stored into different sources with different formats:**



Data

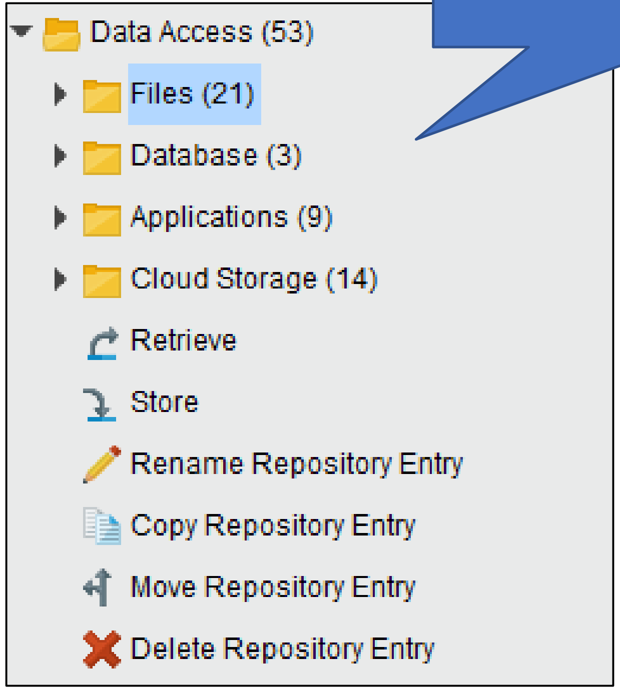
Files
- **CSV**
- **Excel**
- ARFF
- etc.

Database
- Oracle
- MySQL
- …

Application
- Twitter
- Salesforce

Web

Rapidminer provides various operators to access different data sources

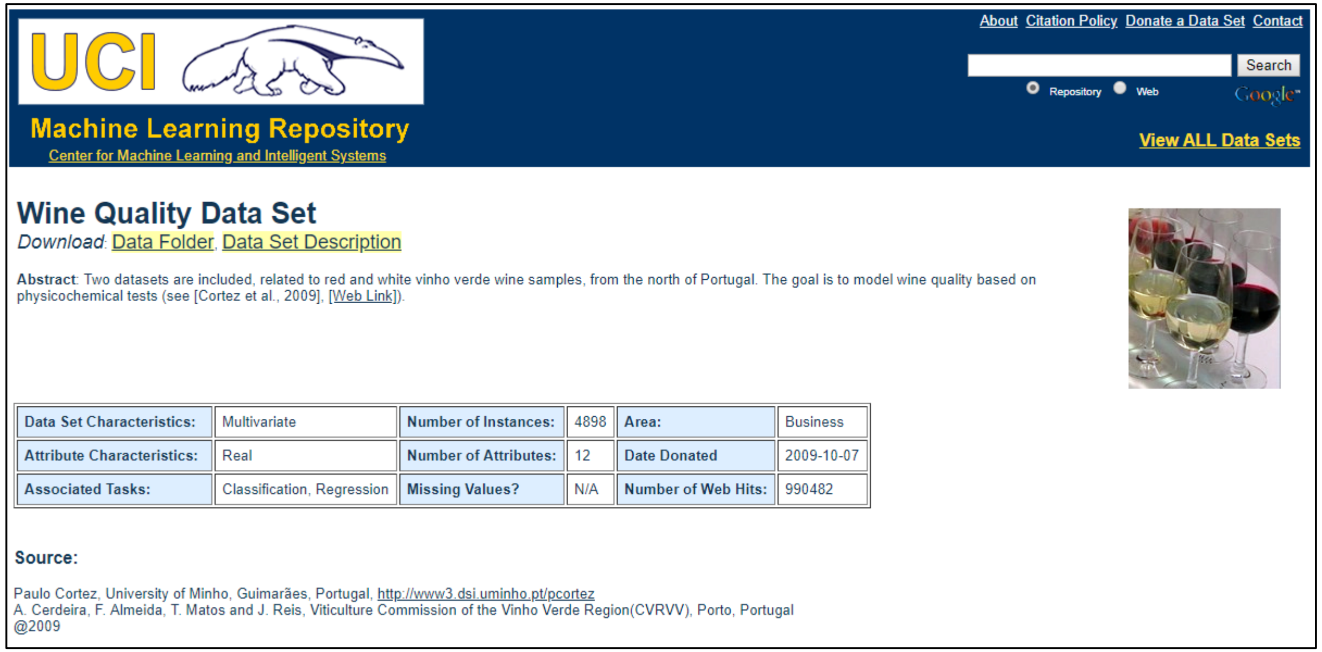


- Data Access (53)
  - Files (21)
  - Database (3)
  - Applications (9)
  - Cloud Storage (14)
  - Retrieve
  - Store
  - Rename Repository Entry
  - Copy Repository Entry
  - Move Repository Entry
  - Delete Repository Entry

Exercise: Working with Rapidminer to Load Dataset

Download Wine Datasets

- Go to https://archive.ics.uci.edu/ml/datasets/wine+quality


About Citation Policy Donate a Data Set Contact
Search
○ Repository ○ Web
Google
UCI
Machine Learning Repository
Center for Machine Learning and Intelligent Systems
View ALL Data Sets

Wine Quality Data Set

Download: Data Folder, Data Set Description

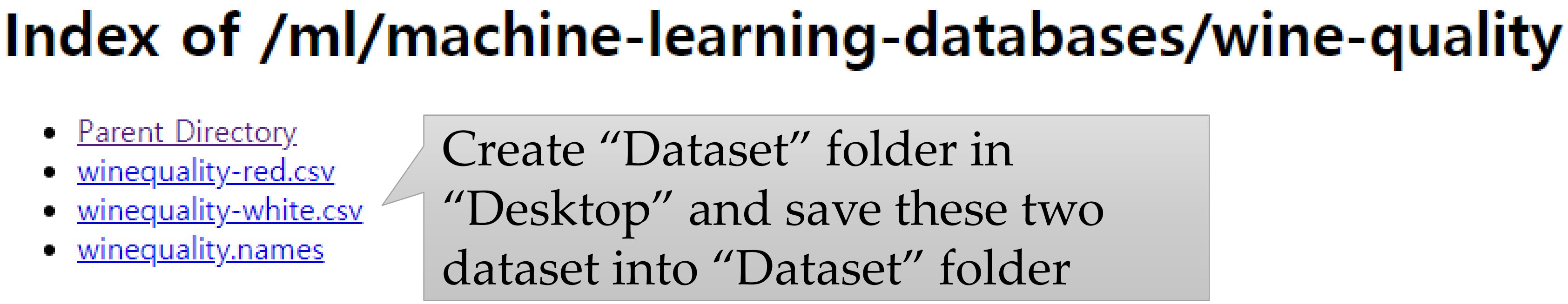
Abstract: Two datasets are included, related to red and white vinho verde wine samples, from the north of Portugal. The goal is to model wine quality based on physicochemical tests (see [Cortez et al., 2009], [Web Link]).

Data Set Characteristics:	Multivariate	Number of Instances:	4898	Area:	Business
Attribute Characteristics:	Real	Number of Attributes:	12	Date Donated	2009-10-07
Associated Tasks:	Classification, Regression	Missing Values?	N/A	Number of Web Hits:	990482

Source:

Paulo Cortez, University of Minho, Guimarães, Portugal, http://www3.dsi.uminho.pt/pcortez
A. Cerdeira, F. Almeida, T. Matos and J. Reis, Viticulture Commission of the Vinho Verde Region(CVRVV), Porto, Portugal
@2009

- Click “Data Folder” link



Index of /ml/machine-learning-databases/wine-quality

- Parent Directory
- winequality-red.csv
- winequality-white.csv
- winequality.names

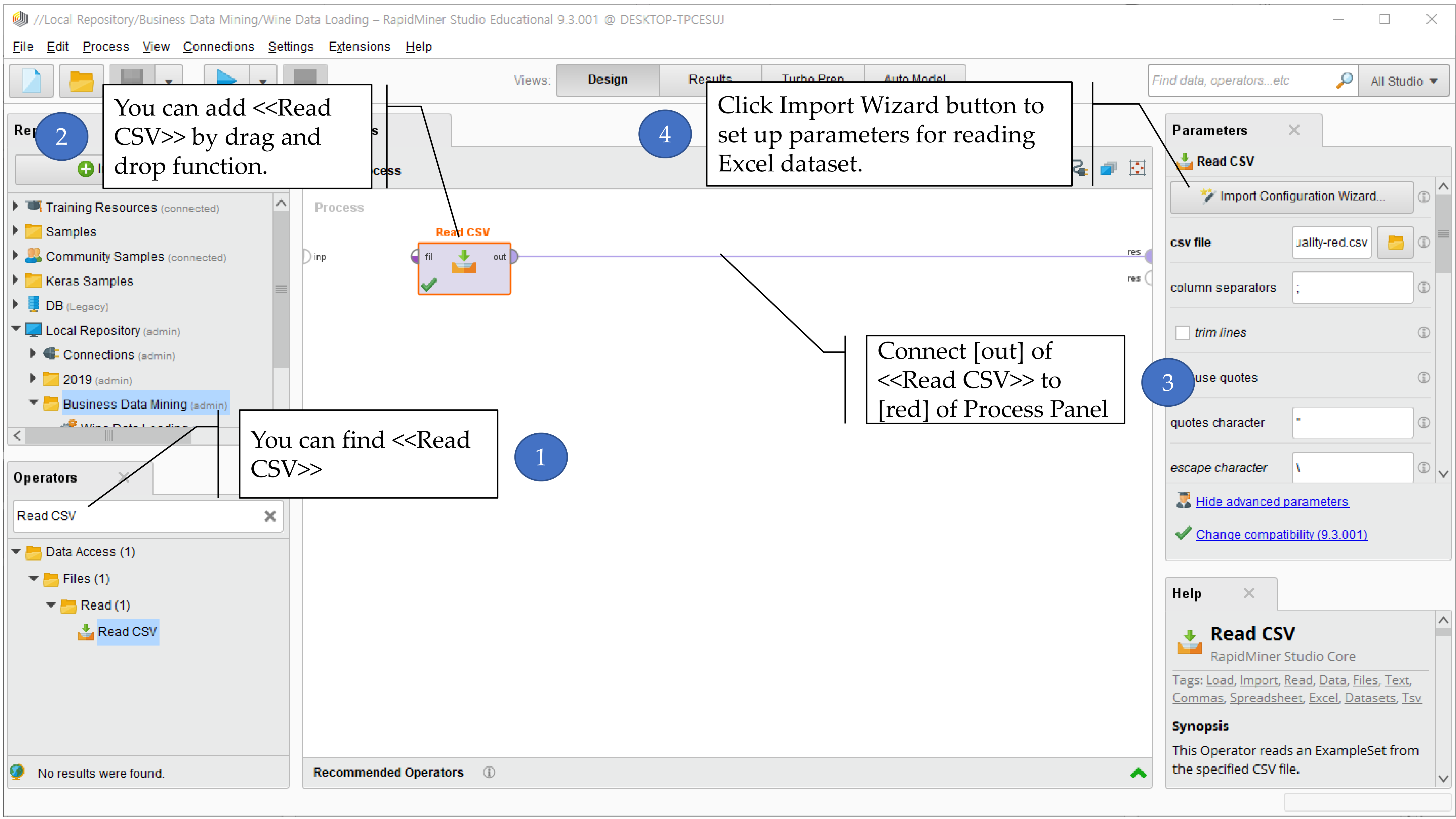
Create “Dataset” folder in “Desktop” and save these two dataset into “Dataset” folder

Exercise: Working with Rapidminer to Load Dataset

Load Dataset Using <<Read Excel>>

- **Tasks**
  - Read “winequality-red.csv”stored in CSV format

- **Steps**
  1. Find <<Read CSV>> operator from Operator Panel
  2. Add <<Read CSV>> to Process Panel
  3. Set up parameters of <<Read CSV>>
  4. Connect [exa] port of <<Read CSV>> to [res] port of Process Panel
  5. Execute the analysis process
  6. Save the process in to ‘Exercise’ repository



//Local Repository/Business Data Mining/Wine Data Loading – RapidMiner Studio Educational 9.3.001 @ DESKTOP-TPCESUJ
File Edit Process View Connections Settings Extensions Help
Views: Design Results Turbo Prep Auto Model
Find data, operators...etc
All Studio
Training Resources (connected)
Samples
Community Samples (connected)
Keras Samples
DB (Legacy)
Local Repository (admin)
Connections (admin)
2019 (admin)
Business Data Mining (admin)
Operators
Data Access (1)
Files (1)
Read (1)
Read CSV
No results were found.
Process
Read CSV
inp
fil
out
res
res
Recommended Operators
Parameters
Read CSV
Import Configuration Wizard...
Hide advanced parameters
Change compatibility (9.3.001)
Help

1 You can find <<Read CSV>>

2 You can add <<Read CSV>> by drag and drop function.

3 Connect [out] of <<Read CSV>> to [red] of Process Panel

4 Click Import Wizard button to set up parameters for reading Excel dataset.

Operators: Read CSV

csv file: uality-red.csv

column separators: ;

□ trim lines

use quotes

quotes character: "

escape character: \

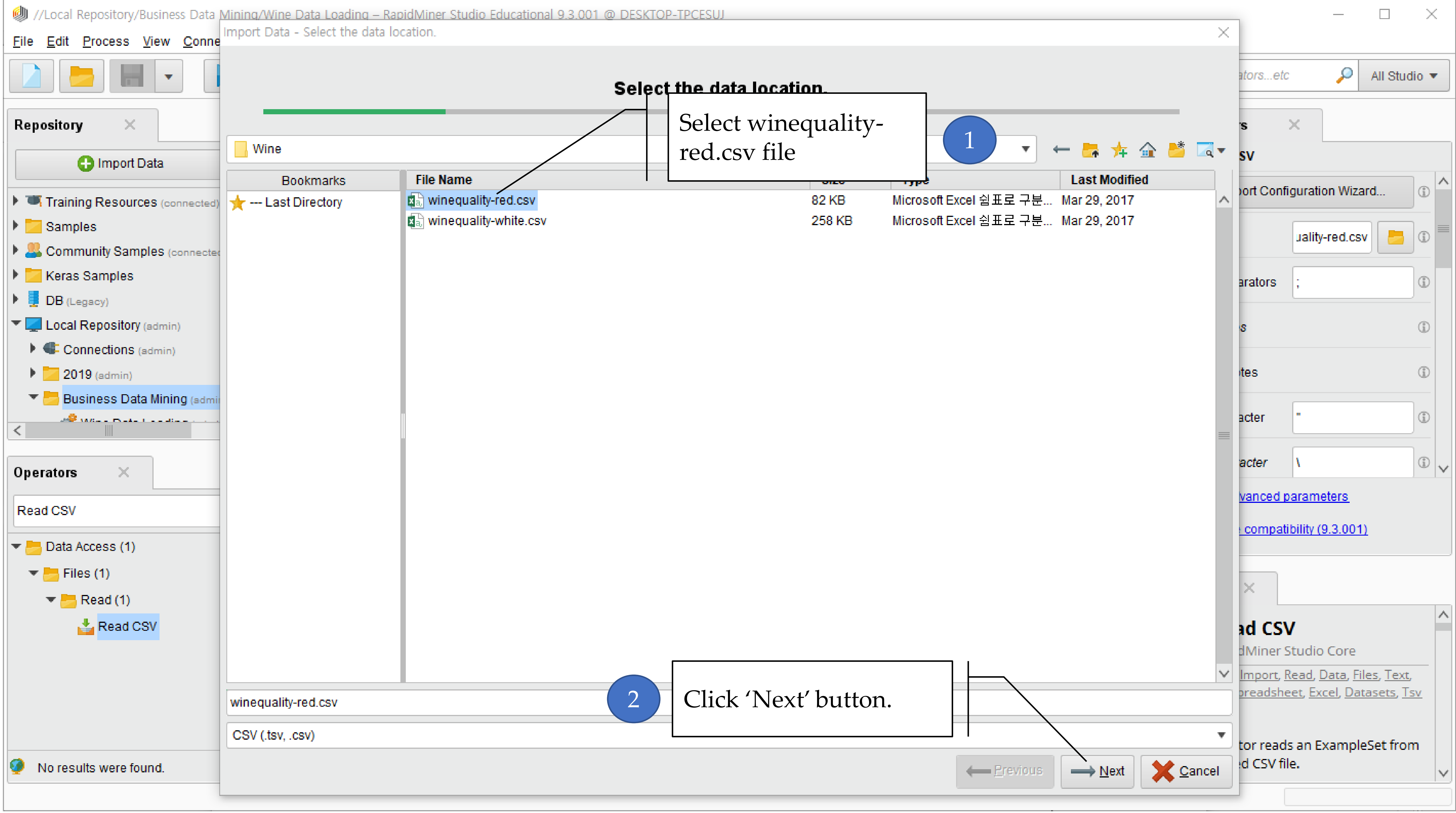
Read CSV

RapidMiner Studio Core

Tags: Load, Import, Read, Data, Files, Text, Commas, Spreadsheet, Excel, Datasets, Tsv

Synopsis

This Operator reads an ExampleSet from the specified CSV file.



Import Data - Select the data location.

Select the data location.

Location: Wine

File Name	Size	Type	Last Modified
winequality-red.csv	82 KB	Microsoft Excel 쉼표로 구분...	Mar 29, 2017
winequality-white.csv	258 KB	Microsoft Excel 쉼표로 구분...	Mar 29, 2017

File name: winequality-red.csv

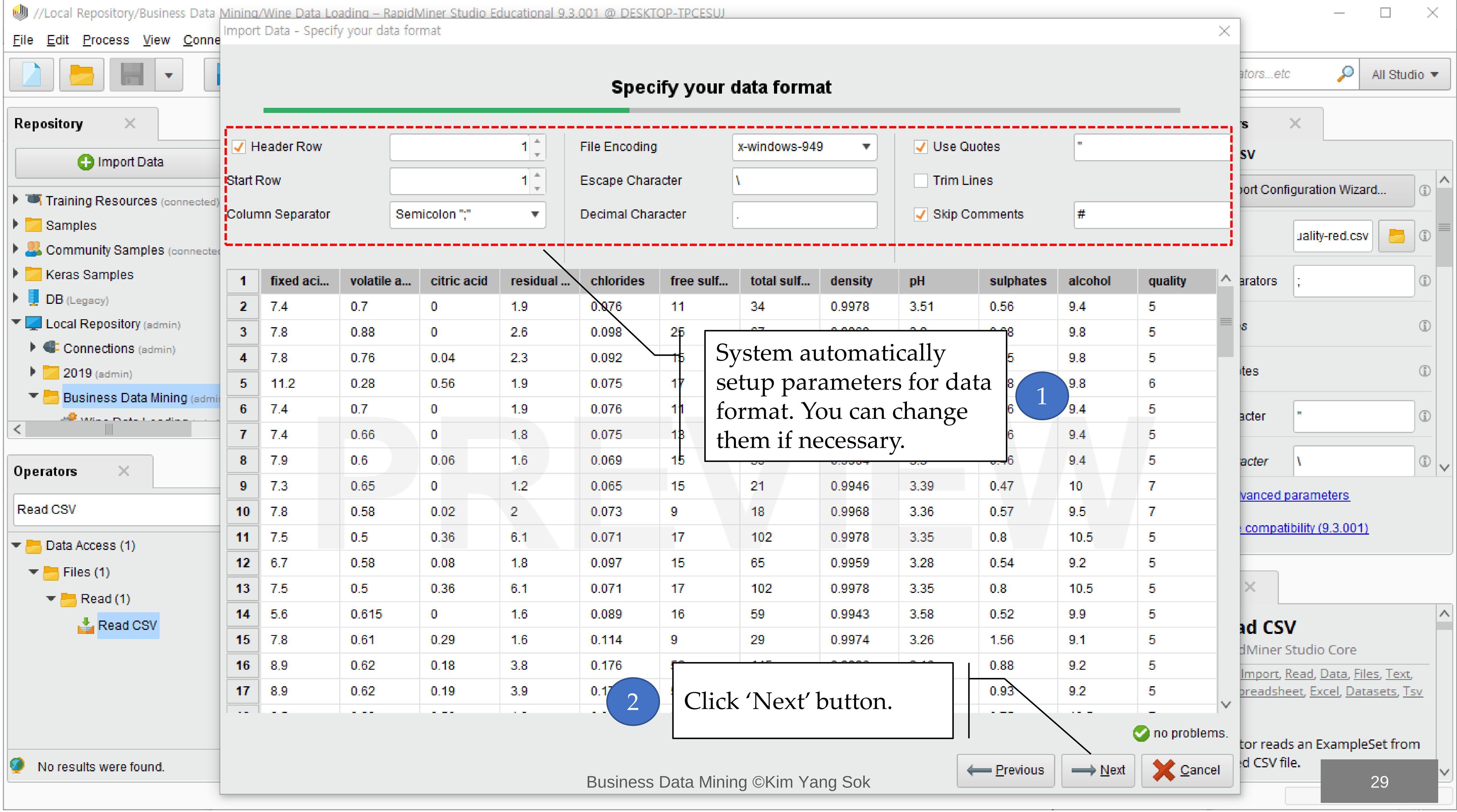
File type: CSV (.tsv, .csv)

Operators search: Read CSV

Read CSV parameters: file: uality-red.csv; column separators: ; ; quotes character: " ; escape character: \

1 Select winequality-red.csv file

2 Click 'Next' button.



//Local Repository/Business Data Mining/Wine Data Loading – RapidMiner Studio Educational 9.3.001 @ DESKTOP-TPCESUJ
File Edit Process View Conne
Repository
Import Data
Training Resources (connected)
Samples
Community Samples (connected)
Keras Samples
DB (Legacy)
Local Repository (admin)
Connections (admin)
2019 (admin)
Business Data Mining (admin)
Operators
Read CSV
Data Access (1)
Files (1)
Read (1)
Read CSV
No results were found.
All Studio
ort Configuration Wizard...
uality-red.csv
arators
vanced parameters
compatibility (9.3.001)
ad CSV
dMiner Studio Core
Import, Read, Data, Files, Text, preadsheet, Excel, Datasets, Tsv
tor reads an ExampleSet from d CSV file.
PREVIEW

Import Data - Specify your data format

Specify your data format

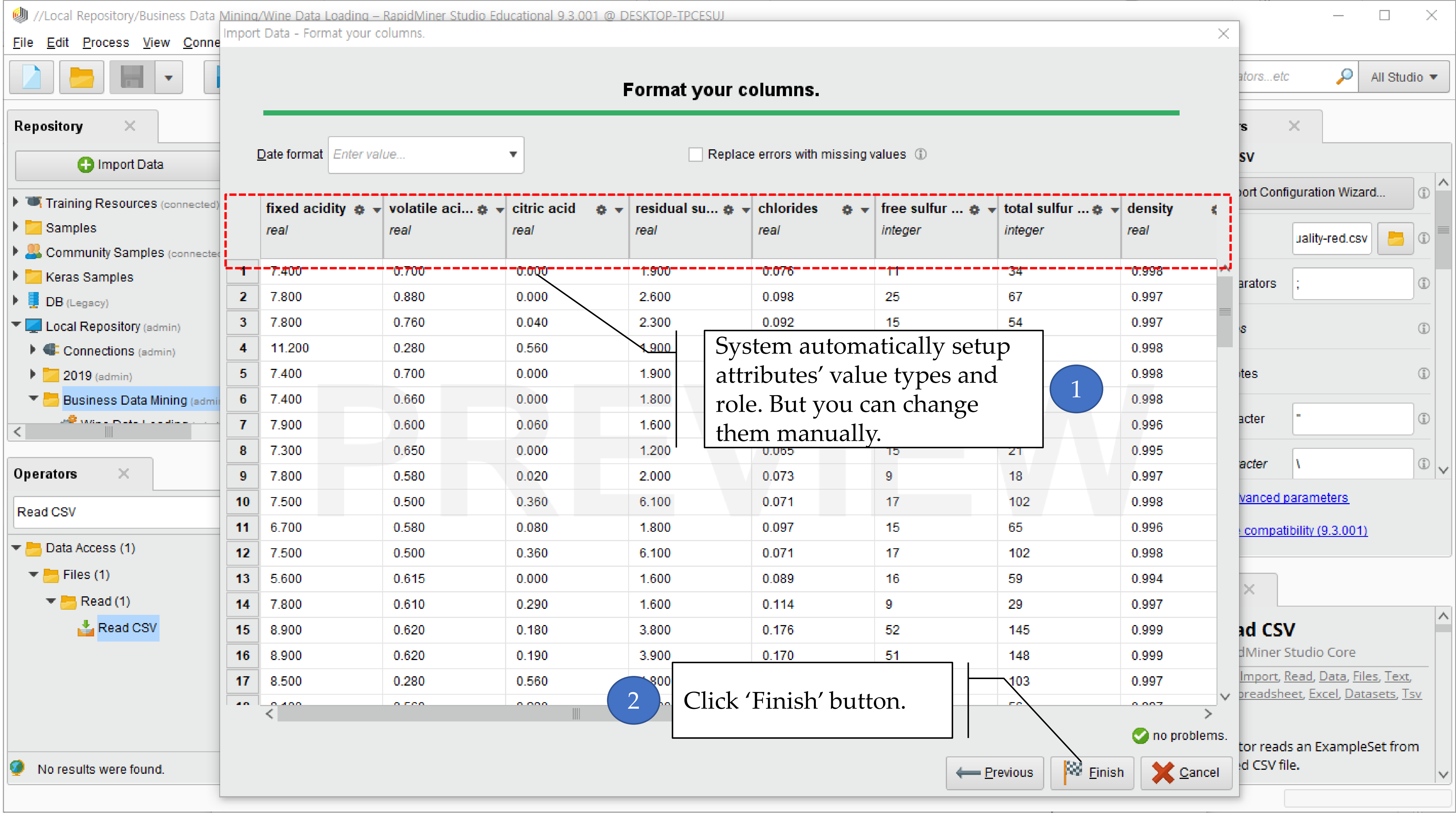
- ☑ Header Row: 1
- Start Row: 1
- Column Separator: Semicolon ";"
- File Encoding: x-windows-949
- Escape Character: \
- Decimal Character: .
- ☑ Use Quotes: "
- □ Trim Lines
- ☑ Skip Comments: #

1	fixed aci...	volatile a...	citric acid	residual ...	chlorides	free sulf...	total sulf...	density	pH	sulphates	alcohol	quality
2	7.4	0.7	0	1.9	0.076	11	34	0.9978	3.51	0.56	9.4	5
3	7.8	0.88	0	2.6	0.098	25	[illegible]	[illegible]	[illegible]	[illegible]	9.8	5
4	7.8	0.76	0.04	2.3	0.092	15	[illegible]	[illegible]	[illegible]	[illegible]	9.8	5
5	11.2	0.28	0.56	1.9	0.075	17	[illegible]	[illegible]	[illegible]	[illegible]	9.8	6
6	7.4	0.7	0	1.9	0.076	11	[illegible]	[illegible]	[illegible]	[illegible]	9.4	5
7	7.4	0.66	0	1.8	0.075	13	[illegible]	[illegible]	[illegible]	[illegible]	9.4	5
8	7.9	0.6	0.06	1.6	0.069	15	[illegible]	[illegible]	[illegible]	[illegible]	9.4	5
9	7.3	0.65	0	1.2	0.065	15	21	0.9946	3.39	0.47	10	7
10	7.8	0.58	0.02	2	0.073	9	18	0.9968	3.36	0.57	9.5	7
11	7.5	0.5	0.36	6.1	0.071	17	102	0.9978	3.35	0.8	10.5	5
12	6.7	0.58	0.08	1.8	0.097	15	65	0.9959	3.28	0.54	9.2	5
13	7.5	0.5	0.36	6.1	0.071	17	102	0.9978	3.35	0.8	10.5	5
14	5.6	0.615	0	1.6	0.089	16	59	0.9943	3.58	0.52	9.9	5
15	7.8	0.61	0.29	1.6	0.114	9	29	0.9974	3.26	1.56	9.1	5
16	8.9	0.62	0.18	3.8	0.176	[illegible]	[illegible]	[illegible]	[illegible]	0.88	9.2	5
17	8.9	0.62	0.19	3.9	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	0.93	9.2	5

no problems.

1 System automatically setup parameters for data format. You can change them if necessary.

2 Click 'Next' button.


//Local Repository/Business Data Mining/Wine Data Loading – RapidMiner Studio Educational 9.3.001 @ DESKTOP-TPCESUJ
File
Edit
Process
View
Conne
Repository
Import Data
Training Resources (connected)
Samples
Community Samples (connecte
Keras Samples
DB (Legacy)
Local Repository (admin)
Connections (admin)
2019 (admin)
Business Data Mining (admi
Operators
Read CSV
Data Access (1)
Files (1)
Read (1)
Read CSV
No results were found.
ators...etc
All Studio
ort Configuration Wizard...
uality-red.csv
arators
acter
acter
vanced parameters
compatibility (9.3.001)
ad CSV
dMiner Studio Core
Import, Read, Data, Files, Text,
preadsheet, Excel, Datasets, Tsv
tor reads an ExampleSet from
d CSV file.
Import Data - Format your columns.
PREVIEW

Format your columns.

Date format: Enter value...

□ Replace errors with missing values

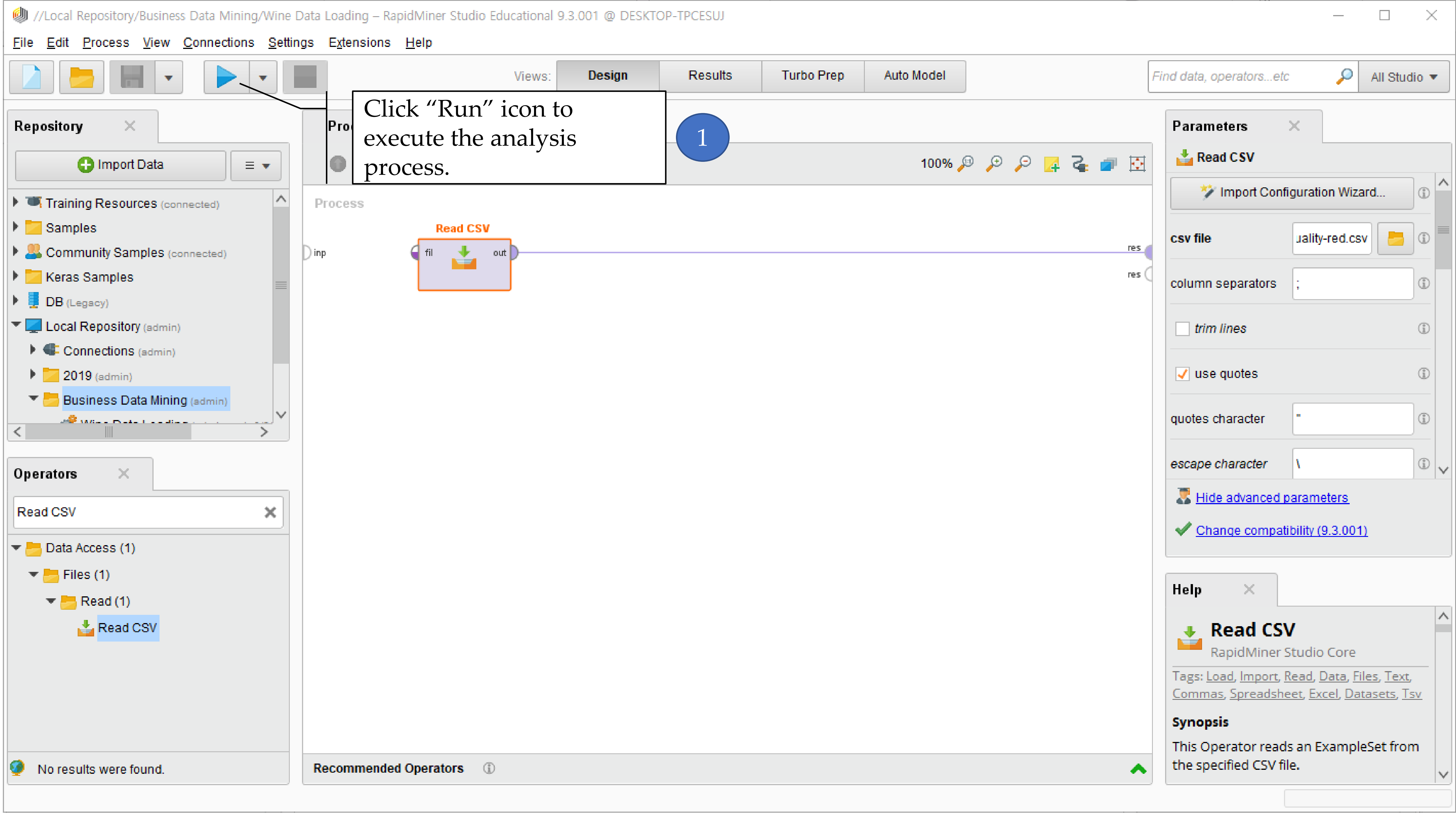
	fixed acidity real	volatile aci... real	citric acid real	residual su... real	chlorides real	free sulfur ... integer	total sulfur ... integer	density real
1	7.400	0.700	0.000	1.900	0.076	11	34	0.998
2	7.800	0.880	0.000	2.600	0.098	25	67	0.997
3	7.800	0.760	0.040	2.300	0.092	15	54	0.997
4	11.200	0.280	0.560	1.900				0.998
5	7.400	0.700	0.000	1.900				0.998
6	7.400	0.660	0.000	1.800				0.998
7	7.900	0.600	0.060	1.600				0.996
8	7.300	0.650	0.000	1.200	0.065	15	21	0.995
9	7.800	0.580	0.020	2.000	0.073	9	18	0.997
10	7.500	0.500	0.360	6.100	0.071	17	102	0.998
11	6.700	0.580	0.080	1.800	0.097	15	65	0.996
12	7.500	0.500	0.360	6.100	0.071	17	102	0.998
13	5.600	0.615	0.000	1.600	0.089	16	59	0.994
14	7.800	0.610	0.290	1.600	0.114	9	29	0.997
15	8.900	0.620	0.180	3.800	0.176	52	145	0.999
16	8.900	0.620	0.190	3.900	0.170	51	148	0.999
17	8.500	0.280	0.560	1.800			103	0.997

no problems.

Previous | Finish | Cancel

1 System automatically setup attributes' value types and role. But you can change them manually.

2 Click 'Finish' button.


//Local Repository/Business Data Mining/Wine Data Loading – RapidMiner Studio Educational 9.3.001 @ DESKTOP-TPCESUJ
File Edit Process View Connections Settings Extensions Help
Views: Design Results Turbo Prep Auto Model
Find data, operators...etc
All Studio
Repository
Import Data
Training Resources (connected)
Samples
Community Samples (connected)
Keras Samples
DB (Legacy)
Local Repository (admin)
Connections (admin)
2019 (admin)
Business Data Mining (admin)
Operators
Read CSV
Data Access (1)
Files (1)
Read (1)
Read CSV
Process
100%
Read CSV
inp
fil
out
res
res
Recommended Operators
Parameters
Read CSV
Import Configuration Wizard...
Hide advanced parameters
Change compatibility (9.3.001)
Help

1

Click "Run" icon to execute the analysis process.

No results were found.

csv file: uality-red.csv

column separators: ;

□ *trim lines*

☑ use quotes

quotes character: "

escape character: \

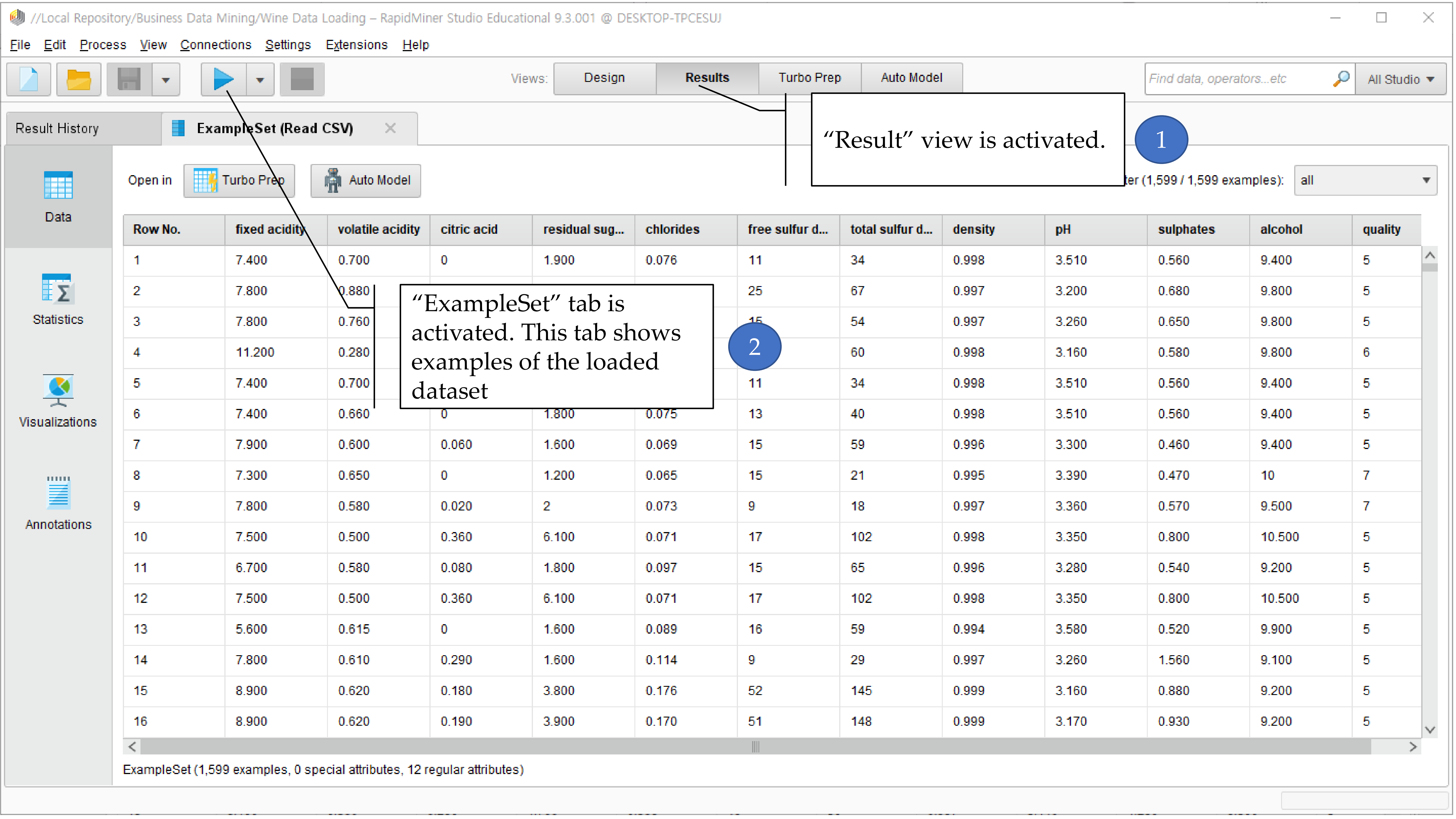
Read CSV

RapidMiner Studio Core

Tags: Load, Import, Read, Data, Files, Text, Commas, Spreadsheet, Excel, Datasets, Tsv

Synopsis

This Operator reads an ExampleSet from the specified CSV file.

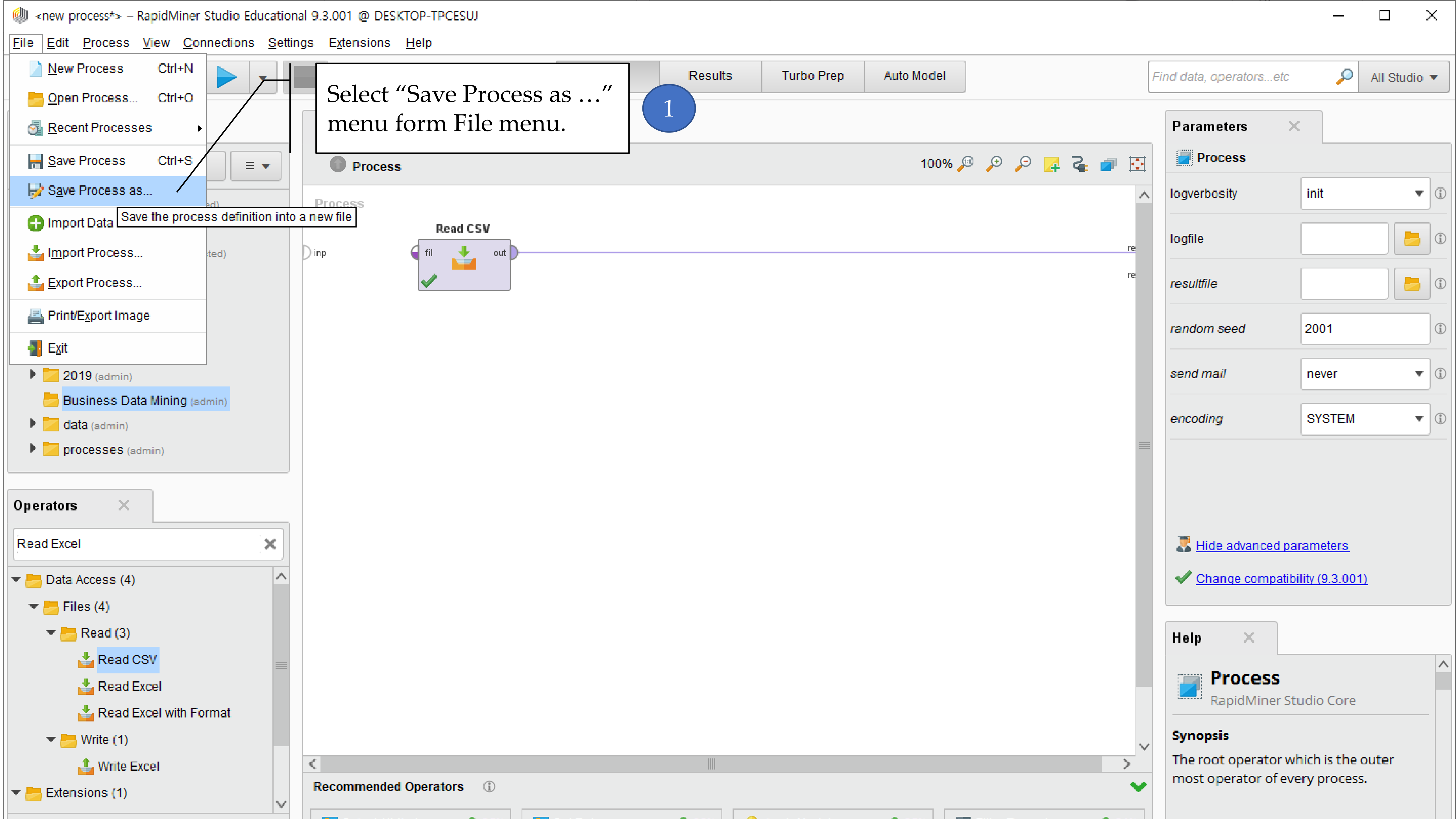


“Result” view is activated.

“ExampleSet” tab is activated. This tab shows examples of the loaded dataset

Row No.	fixed acidity	volatile acidity	citric acid	residual sug...	chlorides	free sulfur d...	total sulfur d...	density	pH	sulphates	alcohol	quality
1	7.400	0.700	0	1.900	0.076	11	34	0.998	3.510	0.560	9.400	5
2	7.800	0.880				25	67	0.997	3.200	0.680	9.800	5
3	7.800	0.760				15	54	0.997	3.260	0.650	9.800	5
4	11.200	0.280					60	0.998	3.160	0.580	9.800	6
5	7.400	0.700				11	34	0.998	3.510	0.560	9.400	5
6	7.400	0.660	0	1.800	0.075	13	40	0.998	3.510	0.560	9.400	5
7	7.900	0.600	0.060	1.600	0.069	15	59	0.996	3.300	0.460	9.400	5
8	7.300	0.650	0	1.200	0.065	15	21	0.995	3.390	0.470	10	7
9	7.800	0.580	0.020	2	0.073	9	18	0.997	3.360	0.570	9.500	7
10	7.500	0.500	0.360	6.100	0.071	17	102	0.998	3.350	0.800	10.500	5
11	6.700	0.580	0.080	1.800	0.097	15	65	0.996	3.280	0.540	9.200	5
12	7.500	0.500	0.360	6.100	0.071	17	102	0.998	3.350	0.800	10.500	5
13	5.600	0.615	0	1.600	0.089	16	59	0.994	3.580	0.520	9.900	5
14	7.800	0.610	0.290	1.600	0.114	9	29	0.997	3.260	1.560	9.100	5
15	8.900	0.620	0.180	3.800	0.176	52	145	0.999	3.160	0.880	9.200	5
16	8.900	0.620	0.190	3.900	0.170	51	148	0.999	3.170	0.930	9.200	5

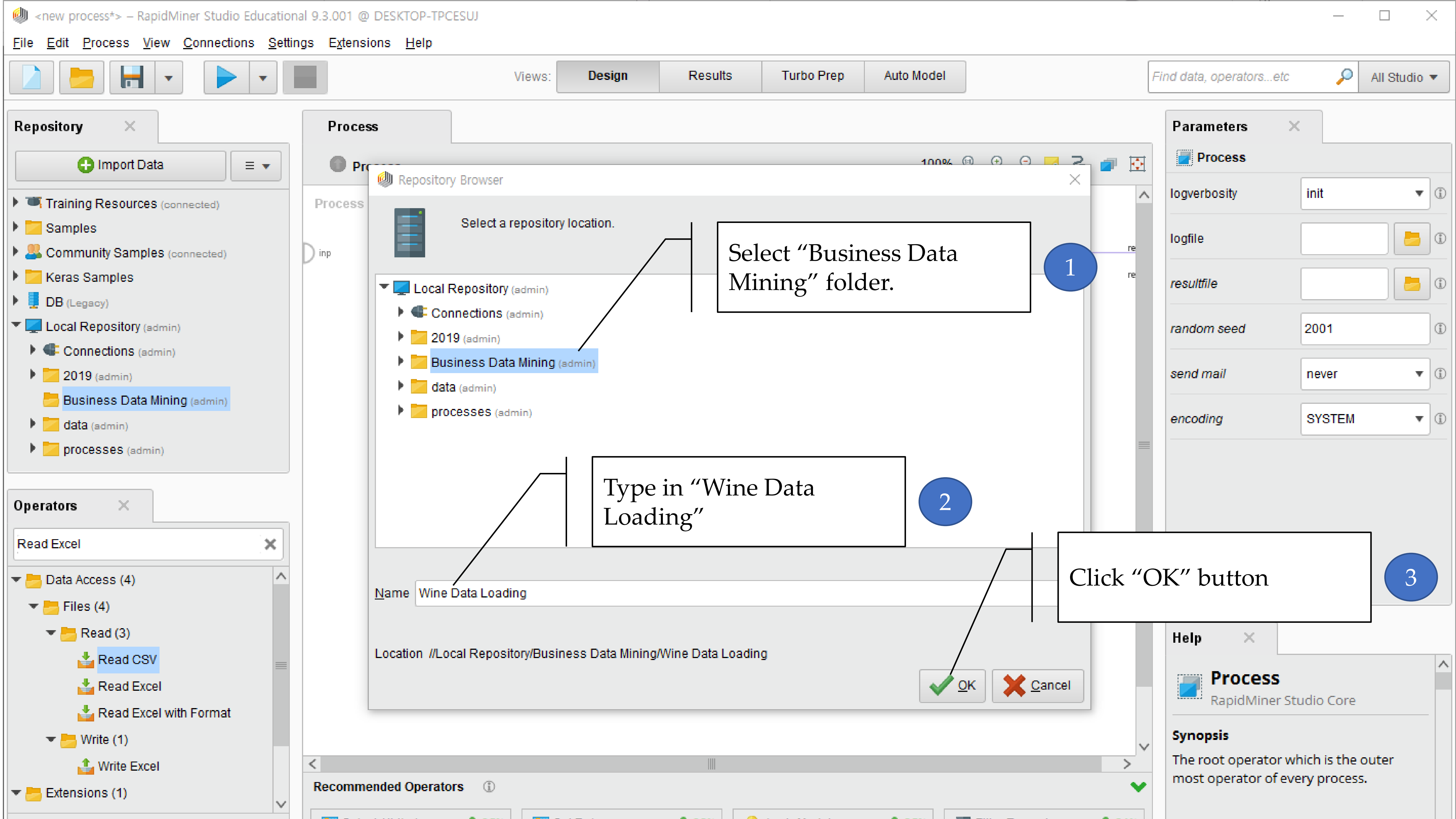
ExampleSet (1,599 examples, 0 special attributes, 12 regular attributes)



Select “Save Process as …” menu form File menu.

logverbosity	init
logfile	
resultfile	
random seed	2001
send mail	never
encoding	SYSTEM

The root operator which is the outer most operator of every process.



Repository Browser

Select a repository location.

Select “Business Data Mining” folder. 1

Type in “Wine Data Loading” 2

Click “OK” button 3

Name Wine Data Loading

Location //Local Repository/Business Data Mining/Wine Data Loading

logverbosity init

logfile

resultfile

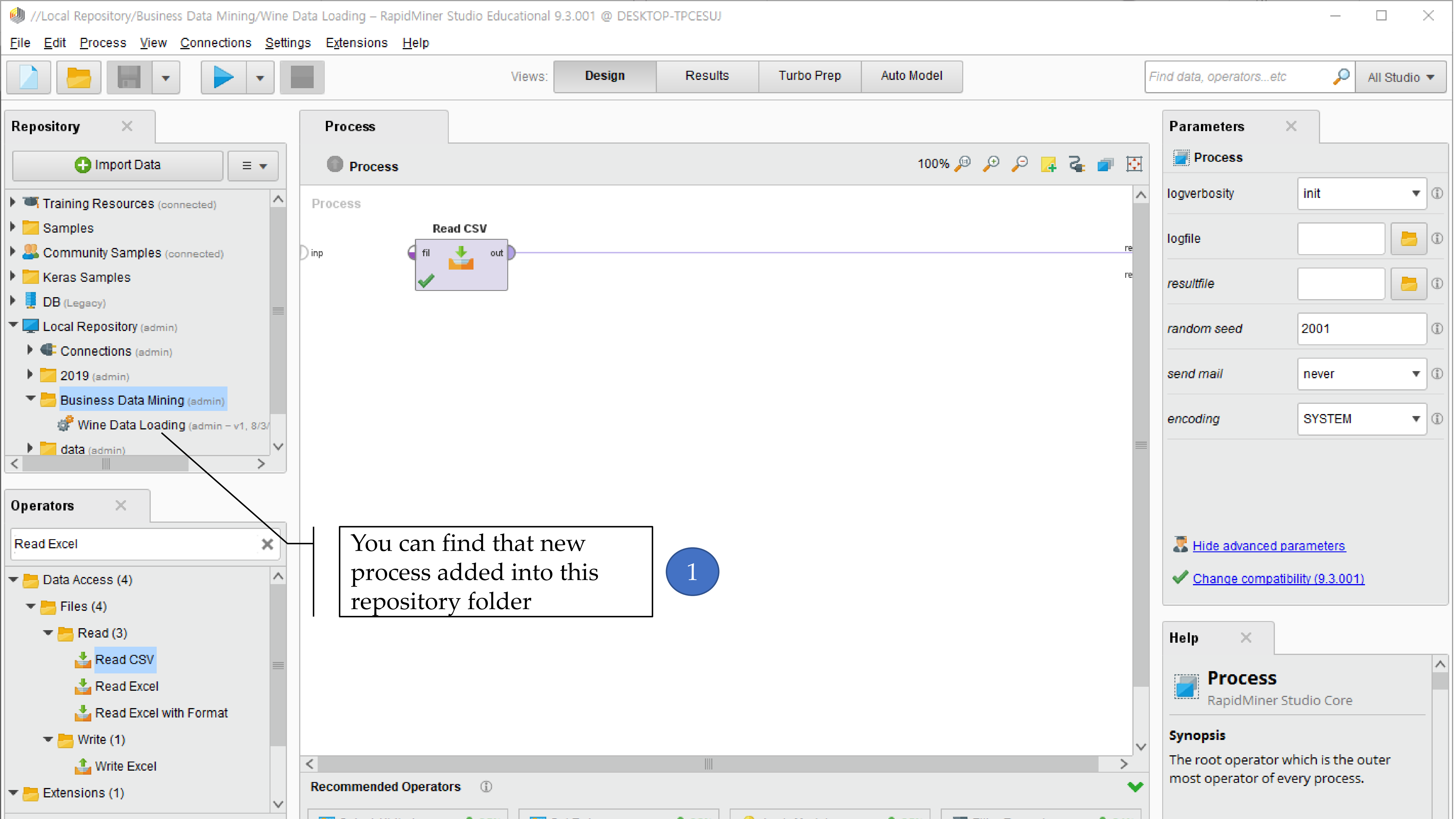
random seed 2001

send mail never

encoding SYSTEM

Synopsis

The root operator which is the outer most operator of every process.



logverbosity: init

logfile:

resultfile:

random seed: 2001

send mail: never

encoding: SYSTEM

Synopsis

The root operator which is the outer most operator of every process.

You can find that new process added into this repository folder

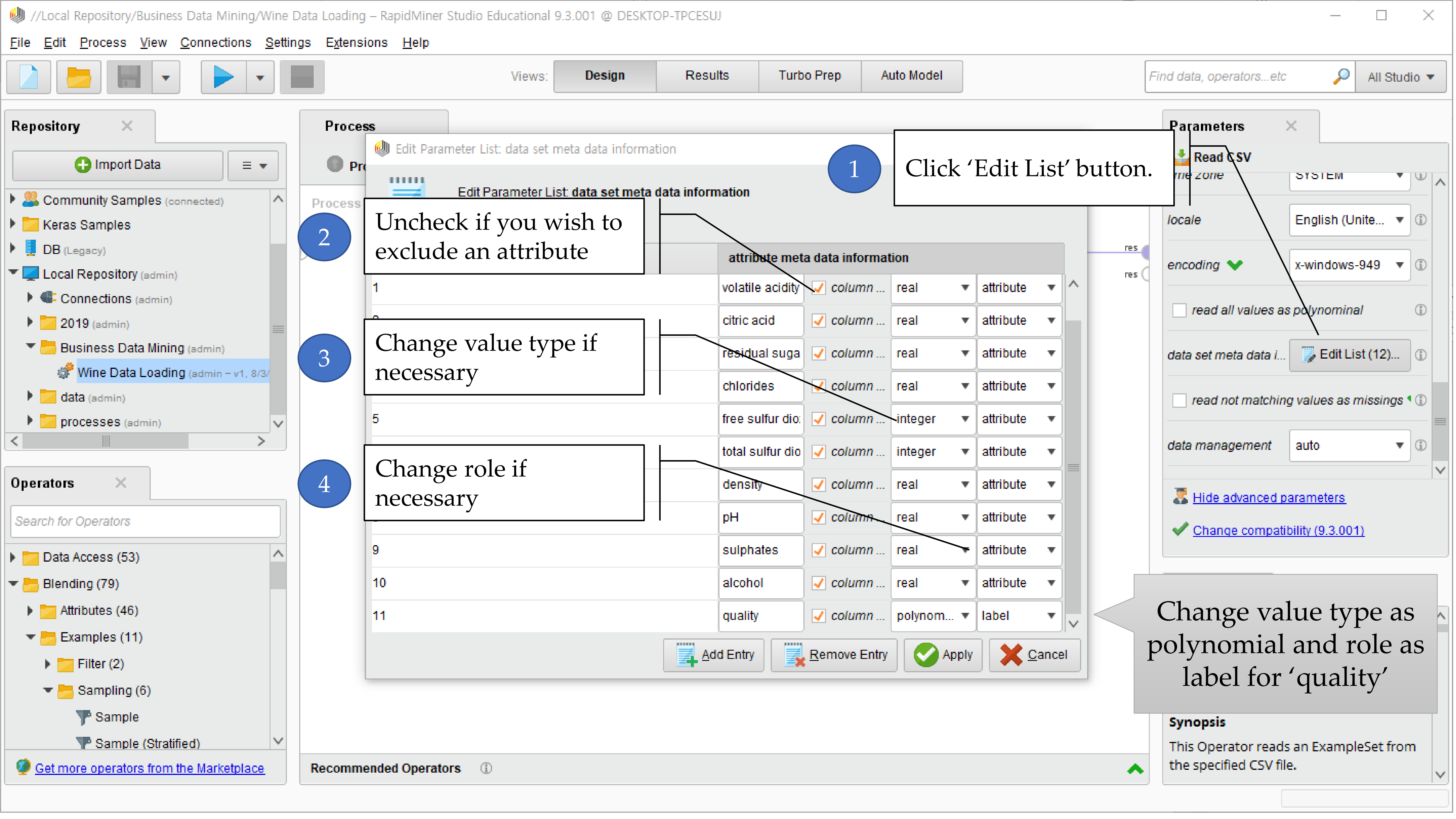
Exercise 2:
Changing Meta Data Information

Exercise: Changing Meta Data Information

Task & Process

- **Tasks**
  - Modify meta data information after loading dataset.

- **Steps**
  1. Select <<Read CSV>> operator
  2. Click ‘Edit List’ button beside ‘data set meta data information’ parameter
  3. Change meta data information


//Local Repository/Business Data Mining/Wine Data Loading – RapidMiner Studio Educational 9.3.001 @ DESKTOP-TPCESUJ
File Edit Process View Connections Settings Extensions Help
Views: Design Results Turbo Prep Auto Model
Find data, operators...etc
All Studio
Repository
Import Data
Community Samples (connected)
Keras Samples
DB (Legacy)
Local Repository (admin)
Connections (admin)
2019 (admin)
Business Data Mining (admin)
Wine Data Loading (admin – v1, 8/3/
data (admin)
processes (admin)
Operators
Search for Operators
Data Access (53)
Blending (79)
Attributes (46)
Examples (11)
Filter (2)
Sampling (6)
Sample
Sample (Stratified)
Get more operators from the Marketplace
Process
Recommended Operators
Parameters
Read CSV
Hide advanced parameters
Change compatibility (9.3.001)
Add Entry
Remove Entry
Apply
Cancel

1 Click 'Edit List' button.

2 Uncheck if you wish to exclude an attribute

3 Change value type if necessary

4 Change role if necessary

Change value type as polynomial and role as label for 'quality'

Edit Parameter List: data set meta data information

Edit Parameter List: **data set meta data information**

	attribute meta data information			
1	volatile acidity	☑ column ...	real	attribute
	citric acid	☑ column ...	real	attribute
	residual suga	☑ column ...	real	attribute
	chlorides	☑ column ...	real	attribute
5	free sulfur dio:	☑ column ...	integer	attribute
	total sulfur dio	☑ column ...	integer	attribute
	density	☑ column ...	real	attribute
	pH	☑ column ...	real	attribute
9	sulphates	☑ column ...	real	attribute
10	alcohol	☑ column ...	real	attribute
11	quality	☑ column ...	polynom...	label

Read CSV parameters:

- time zone: SYSTEM
- locale: English (Unite...
- encoding: x-windows-949
- □ read all values as polynominal
- data set meta data i...: Edit List (12)...
- □ read not matching values as missings
- data management: auto

Synopsis

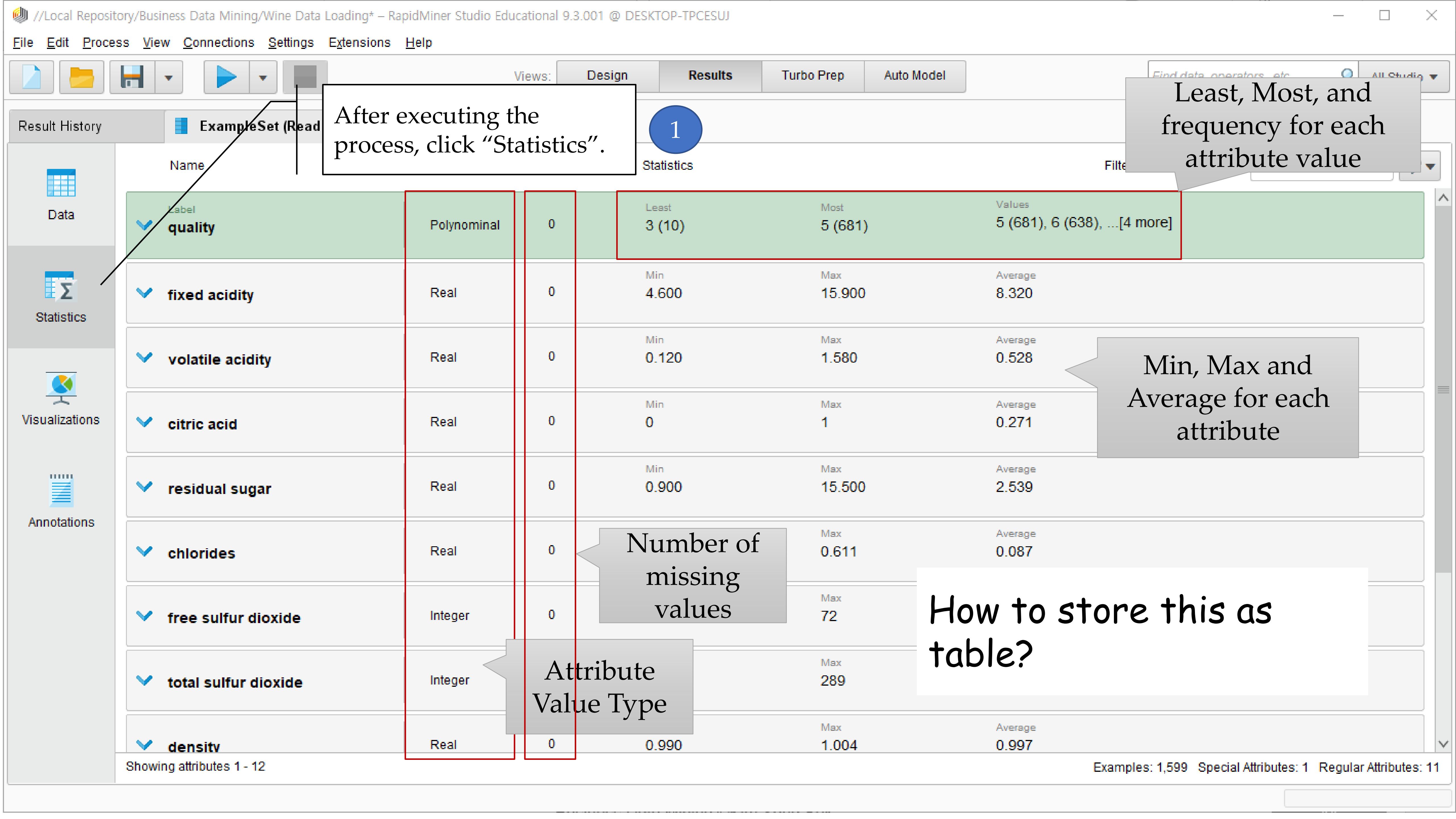
This Operator reads an ExampleSet from the specified CSV file.

Exercise 3: Exploring Data with Rapidminer

Exercise: Exploring Data with Rapidminer

Obtain descriptive statistics

- **You can understand data by reviewing descriptive statistics, which include the followings:**
- **Central value**
  - Numerical value
    - Average(Mean)
    - Median
    - Mode
  - Categorical
    - Least / Most Frequent
- **Dispersion**
  - Range
  - Min & Max
  - Variance
  - Standard deviation


//Local Repository/Business Data Mining/Wine Data Loading* – RapidMiner Studio Educational 9.3.001 @ DESKTOP-TPCESUJ
File Edit Process View Connections Settings Extensions Help
Views: Design Results Turbo Prep Auto Model
Find data, operators...etc
All Studio
Result History
ExampleSet (Read
Data
Statistics
Visualizations
Annotations
Filte
1

After executing the process, click “Statistics”.

Least, Most, and frequency for each attribute value

Min, Max and Average for each attribute

Number of missing values

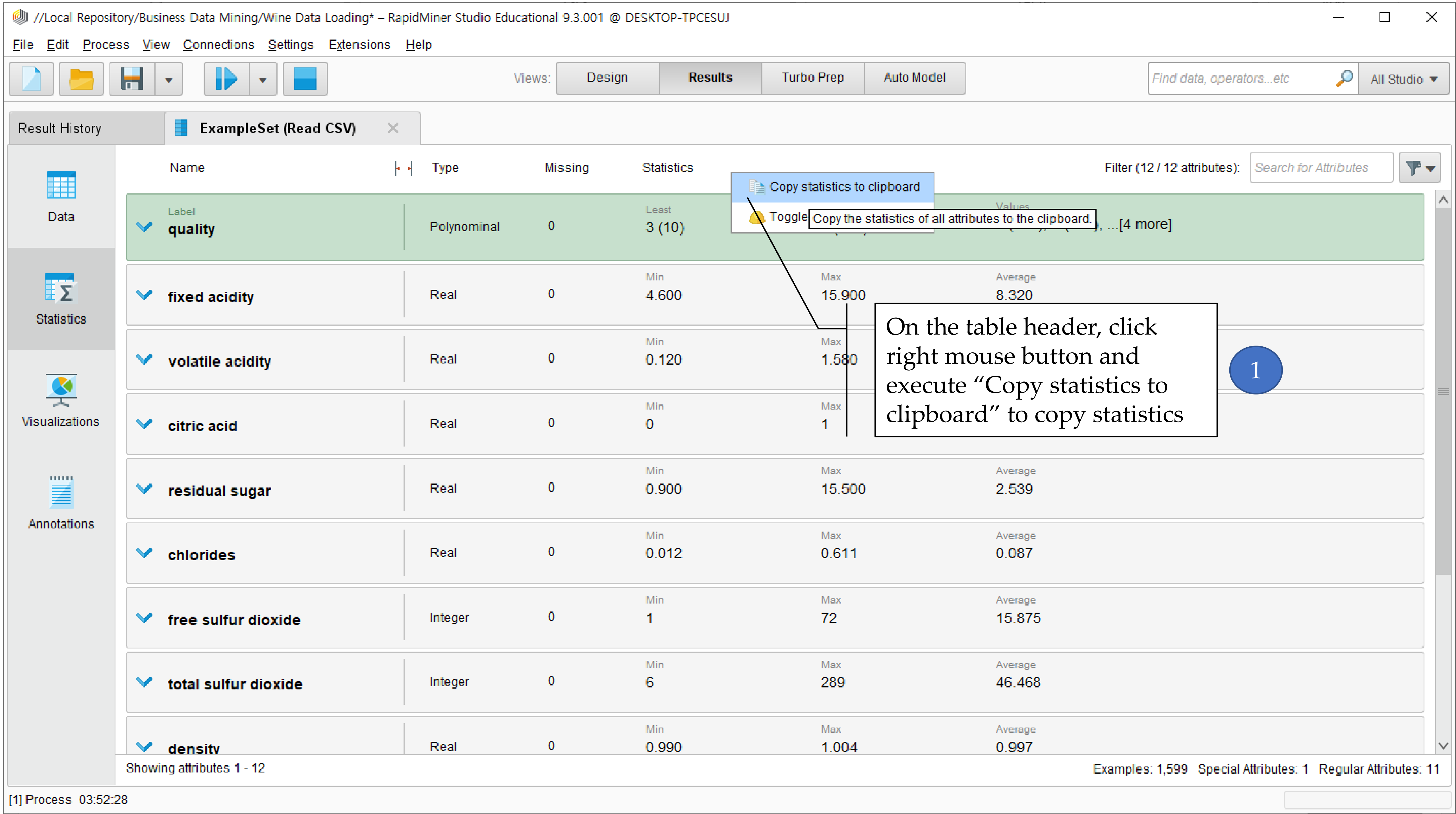
Attribute Value Type

Name			Statistics		
Label quality	Polynominal	0	Least 3 (10)	Most 5 (681)	Values 5 (681), 6 (638), ...[4 more]
fixed acidity	Real	0	Min 4.600	Max 15.900	Average 8.320
volatile acidity	Real	0	Min 0.120	Max 1.580	Average 0.528
citric acid	Real	0	Min 0	Max 1	Average 0.271
residual sugar	Real	0	Min 0.900	Max 15.500	Average 2.539
chlorides	Real	0		Max 0.611	Average 0.087
free sulfur dioxide	Integer	0		Max 72	
total sulfur dioxide	Integer			Max 289	
density	Real	0	0.990	Max 1.004	Average 0.997

Showing attributes 1 - 12

Examples: 1,599 Special Attributes: 1 Regular Attributes: 11

How to store this as table?



//Local Repository/Business Data Mining/Wine Data Loading* – RapidMiner Studio Educational 9.3.001 @ DESKTOP-TPCESUJ

Copy the statistics of all attributes to the clipboard.

Name	Type	Missing	Statistics		
Label quality	Polynominal	0	Least 3 (10)		Values ..., ...[4 more]
fixed acidity	Real	0	Min 4.600	Max 15.900	Average 8.320
volatile acidity	Real	0	Min 0.120	Max 1.580	
citric acid	Real	0	Min 0	Max 1	
residual sugar	Real	0	Min 0.900	Max 15.500	Average 2.539
chlorides	Real	0	Min 0.012	Max 0.611	Average 0.087
free sulfur dioxide	Integer	0	Min 1	Max 72	Average 15.875
total sulfur dioxide	Integer	0	Min 6	Max 289	Average 46.468
density	Real	0	Min 0.990	Max 1.004	Average 0.997

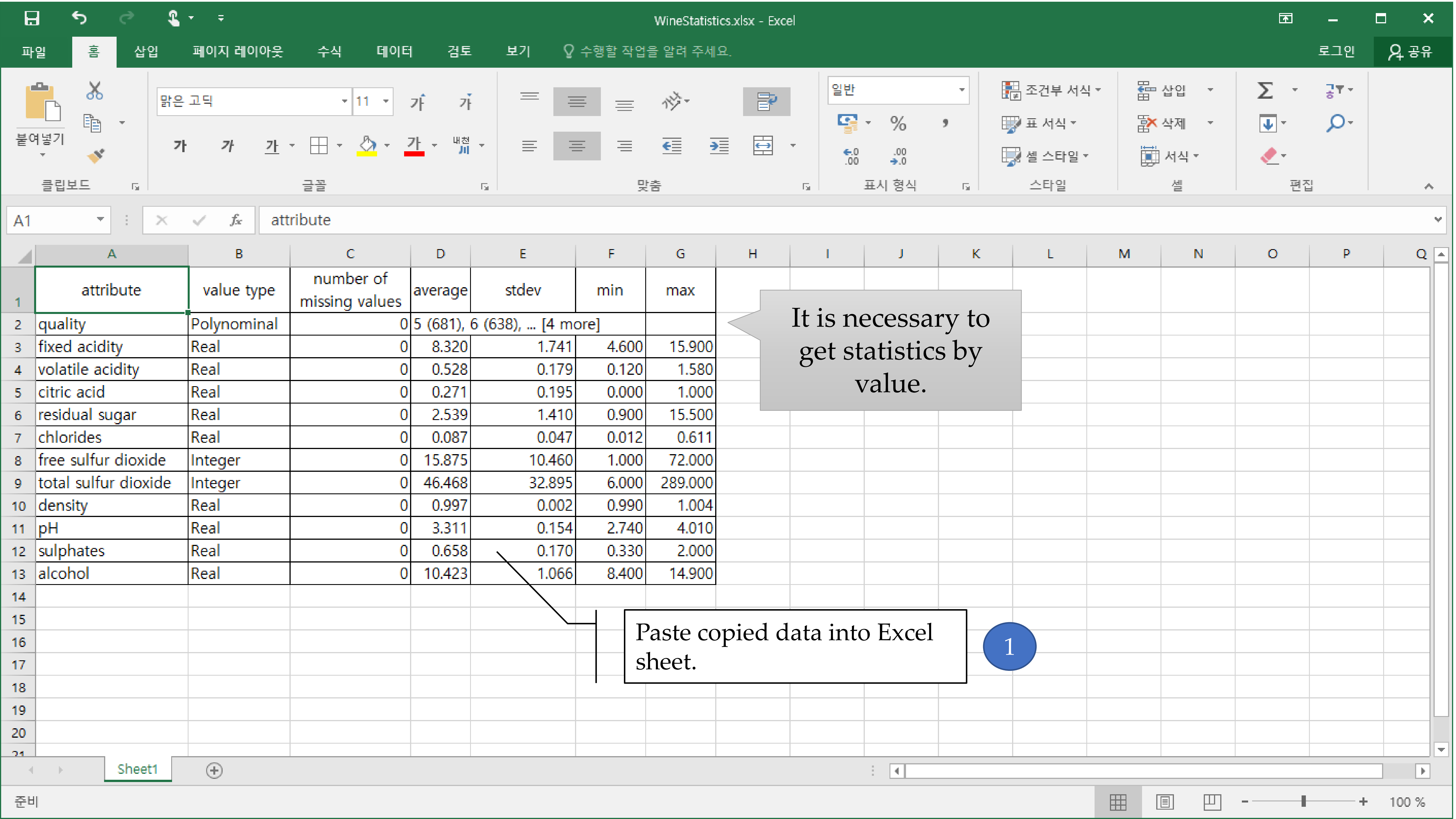
Showing attributes 1 - 12

Examples: 1,599 Special Attributes: 1 Regular Attributes: 11

[1] Process 03:52:28

On the table header, click right mouse button and execute “Copy statistics to clipboard” to copy statistics

1

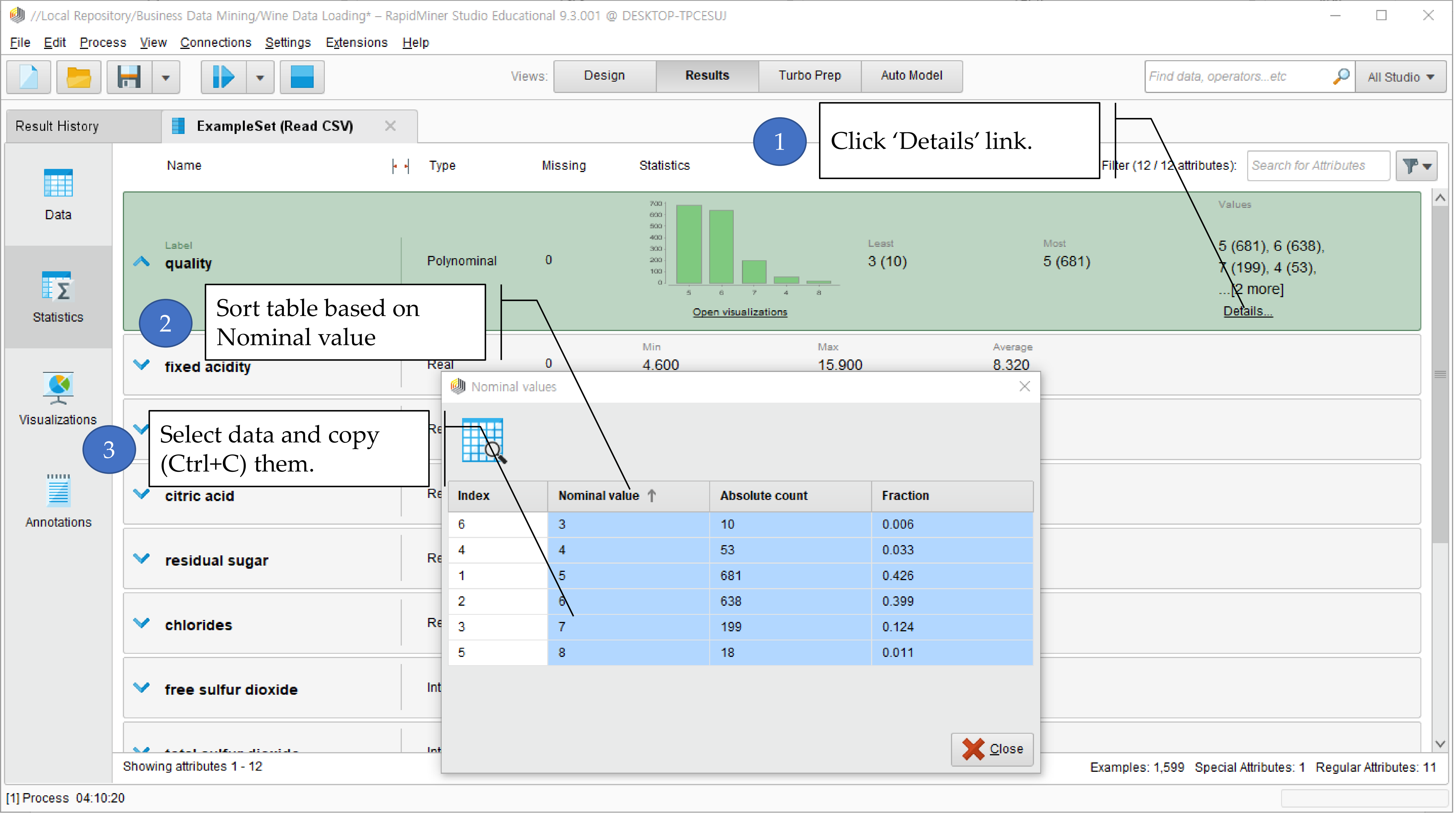


attribute	value type	number of missing values	average	stdev	min	max
quality	Polynominal	0	5 (681), 6 (638), ... [4 more]			
fixed acidity	Real	0	8.320	1.741	4.600	15.900
volatile acidity	Real	0	0.528	0.179	0.120	1.580
citric acid	Real	0	0.271	0.195	0.000	1.000
residual sugar	Real	0	2.539	1.410	0.900	15.500
chlorides	Real	0	0.087	0.047	0.012	0.611
free sulfur dioxide	Integer	0	15.875	10.460	1.000	72.000
total sulfur dioxide	Integer	0	46.468	32.895	6.000	289.000
density	Real	0	0.997	0.002	0.990	1.004
pH	Real	0	3.311	0.154	2.740	4.010
sulphates	Real	0	0.658	0.170	0.330	2.000
alcohol	Real	0	10.423	1.066	8.400	14.900

It is necessary to get statistics by value.

Paste copied data into Excel sheet.

1

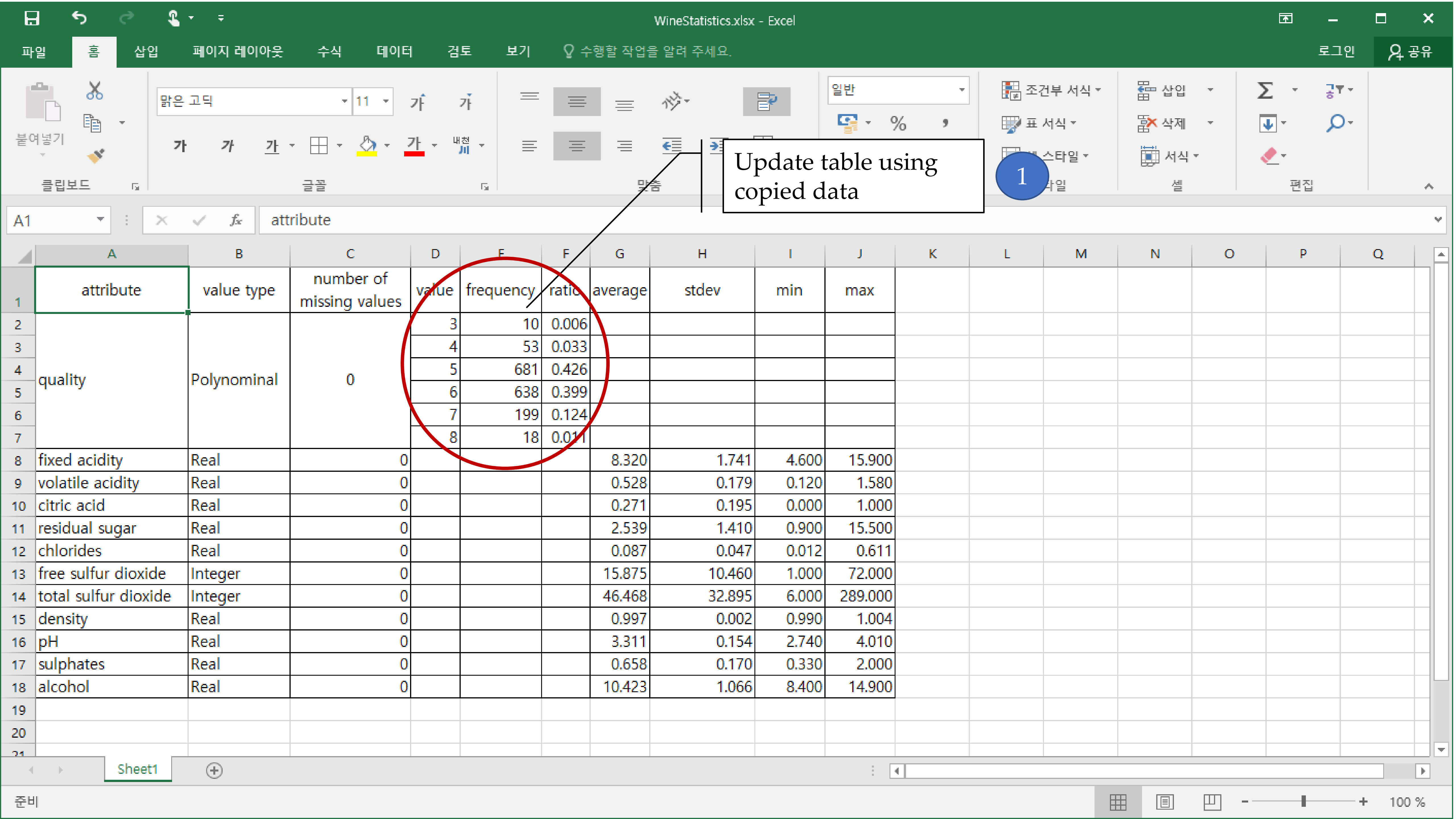


1 Click ‘Details’ link.

2 Sort table based on Nominal value

3 Select data and copy (Ctrl+C) them.

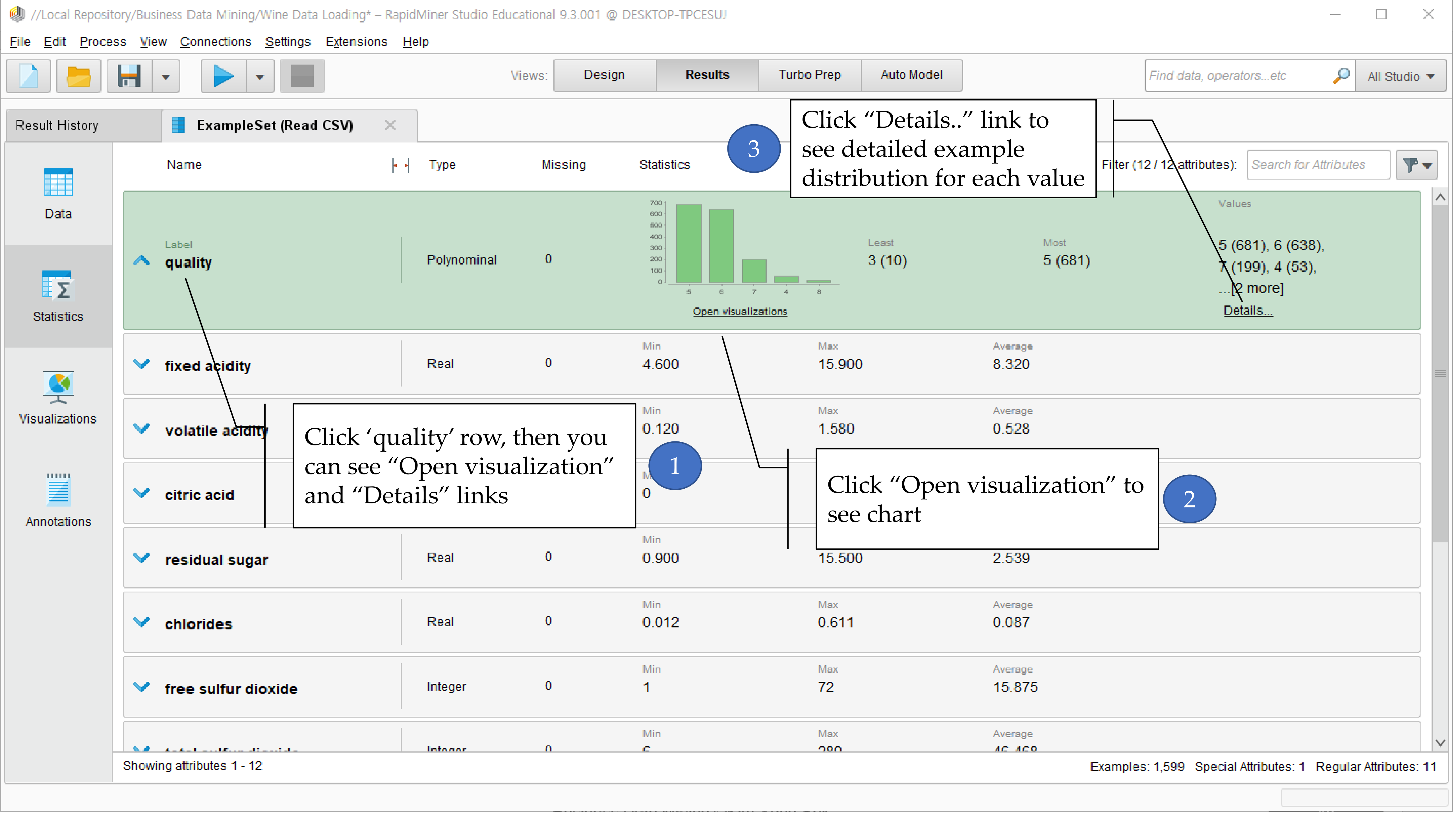
Index	Nominal value ↑	Absolute count	Fraction
6	3	10	0.006
4	4	53	0.033
1	5	681	0.426
2	6	638	0.399
3	7	199	0.124
5	8	18	0.011



Update table using copied data

1

attribute	value type	number of missing values	value	frequency	ratio	average	stdev	min	max
quality	Polynominal	0	3	10	0.006				
			4	53	0.033				
			5	681	0.426				
			6	638	0.399				
			7	199	0.124				
			8	18	0.011				
fixed acidity	Real	0				8.320	1.741	4.600	15.900
volatile acidity	Real	0				0.528	0.179	0.120	1.580
citric acid	Real	0				0.271	0.195	0.000	1.000
residual sugar	Real	0				2.539	1.410	0.900	15.500
chlorides	Real	0				0.087	0.047	0.012	0.611
free sulfur dioxide	Integer	0				15.875	10.460	1.000	72.000
total sulfur dioxide	Integer	0				46.468	32.895	6.000	289.000
density	Real	0				0.997	0.002	0.990	1.004
pH	Real	0				3.311	0.154	2.740	4.010
sulphates	Real	0				0.658	0.170	0.330	2.000
alcohol	Real	0				10.423	1.066	8.400	14.900



Name	Type	Missing	Statistics		
Label quality	Polynominal	0	Least 3 (10)	Most 5 (681)	Values 5 (681), 6 (638), 7 (199), 4 (53), ...[2 more] Details...
fixed acidity	Real	0	Min 4.600	Max 15.900	Average 8.320
volatile acidity			Min 0.120	Max 1.580	Average 0.528
citric acid			M 0		
residual sugar	Real	0	Min 0.900	15.500	2.539
chlorides	Real	0	Min 0.012	Max 0.611	Average 0.087
free sulfur dioxide	Integer	0	Min 1	Max 72	Average 15.875
total sulfur dioxide	Integer	0	Min 6	Max 289	Average 46.468

Showing attributes 1 - 12

Examples: 1,599 Special Attributes: 1 Regular Attributes: 11

1 Click 'quality' row, then you can see "Open visualization" and "Details" links

2 Click "Open visualization" to see chart

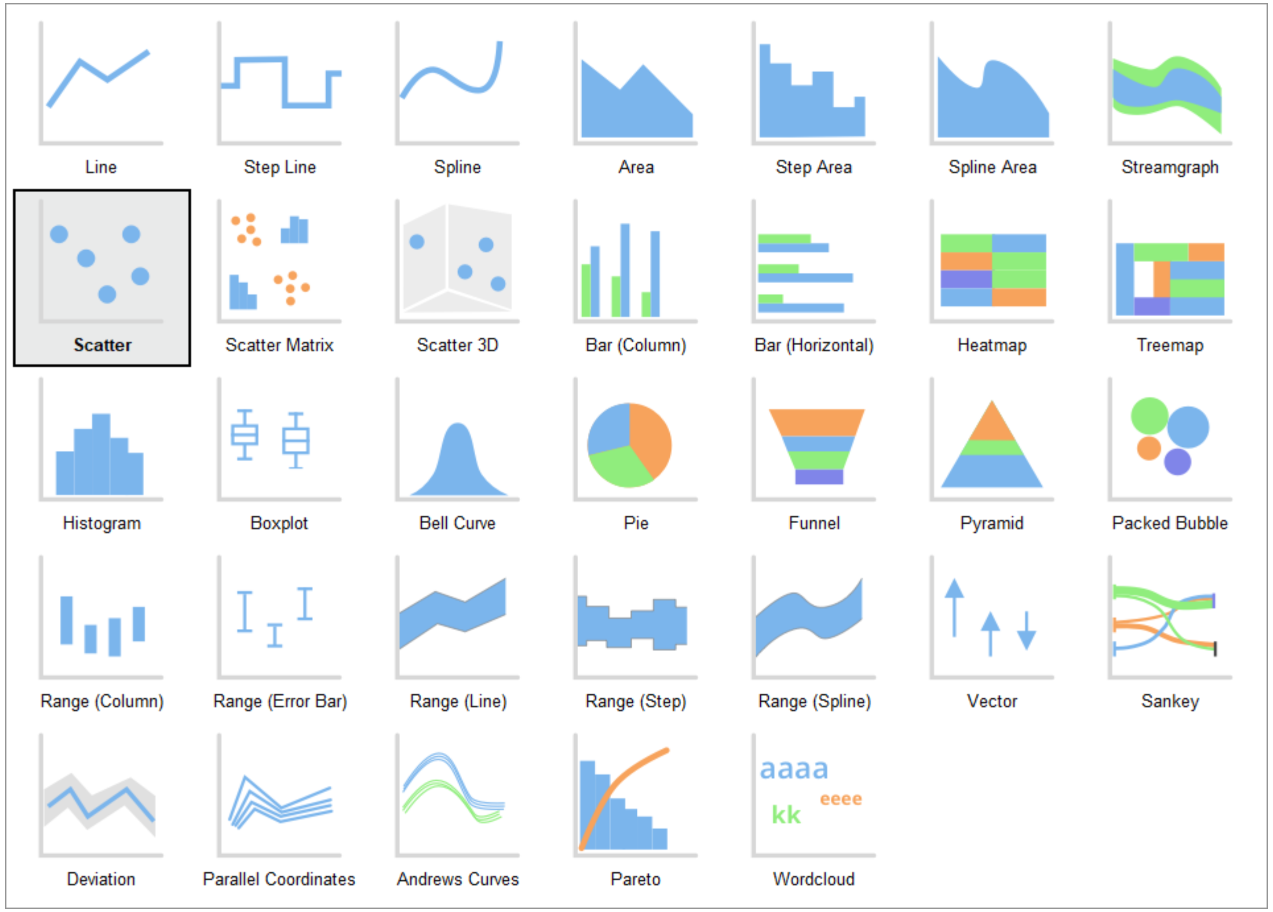
3 Click "Details.." link to see detailed example distribution for each value

Exercise: Exploring Data with Rapidminer

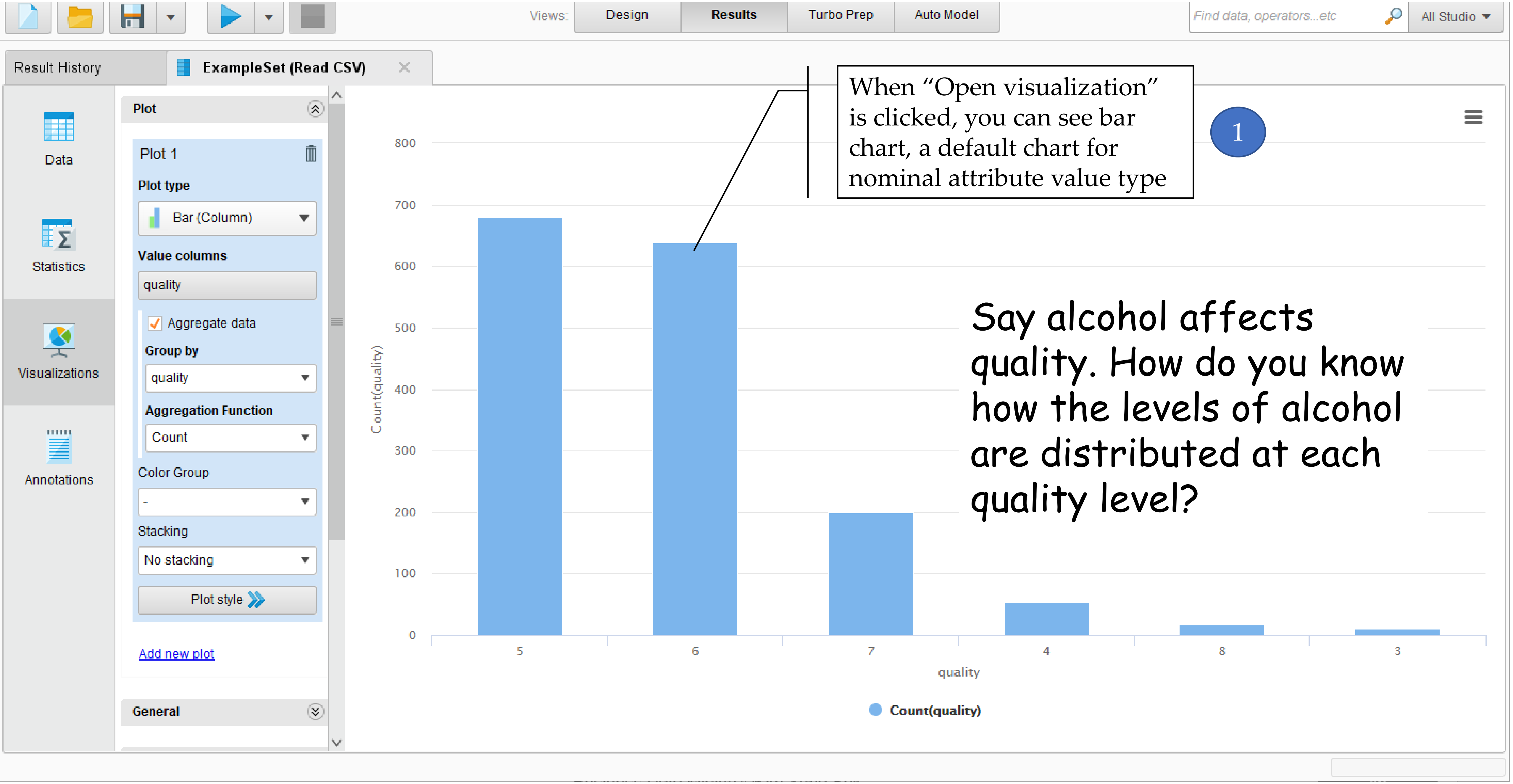
Visualize Data with Charts

- **You can use charts to easily identify patterns within data.**
- **Many charts are available, but we only focus on the following four charts**

	Single Attribute	**Multiple Attributes**
Categorical Values	Bar Charts	Stacked Bar chart
Numerical Values	Histogram	Scatter Plot



Visualize single nominal attribute with a bar chart



Plot type: Bar (Column)

Value columns: quality

☑ Aggregate data

Group by: quality

Aggregation Function: Count

Color Group: -

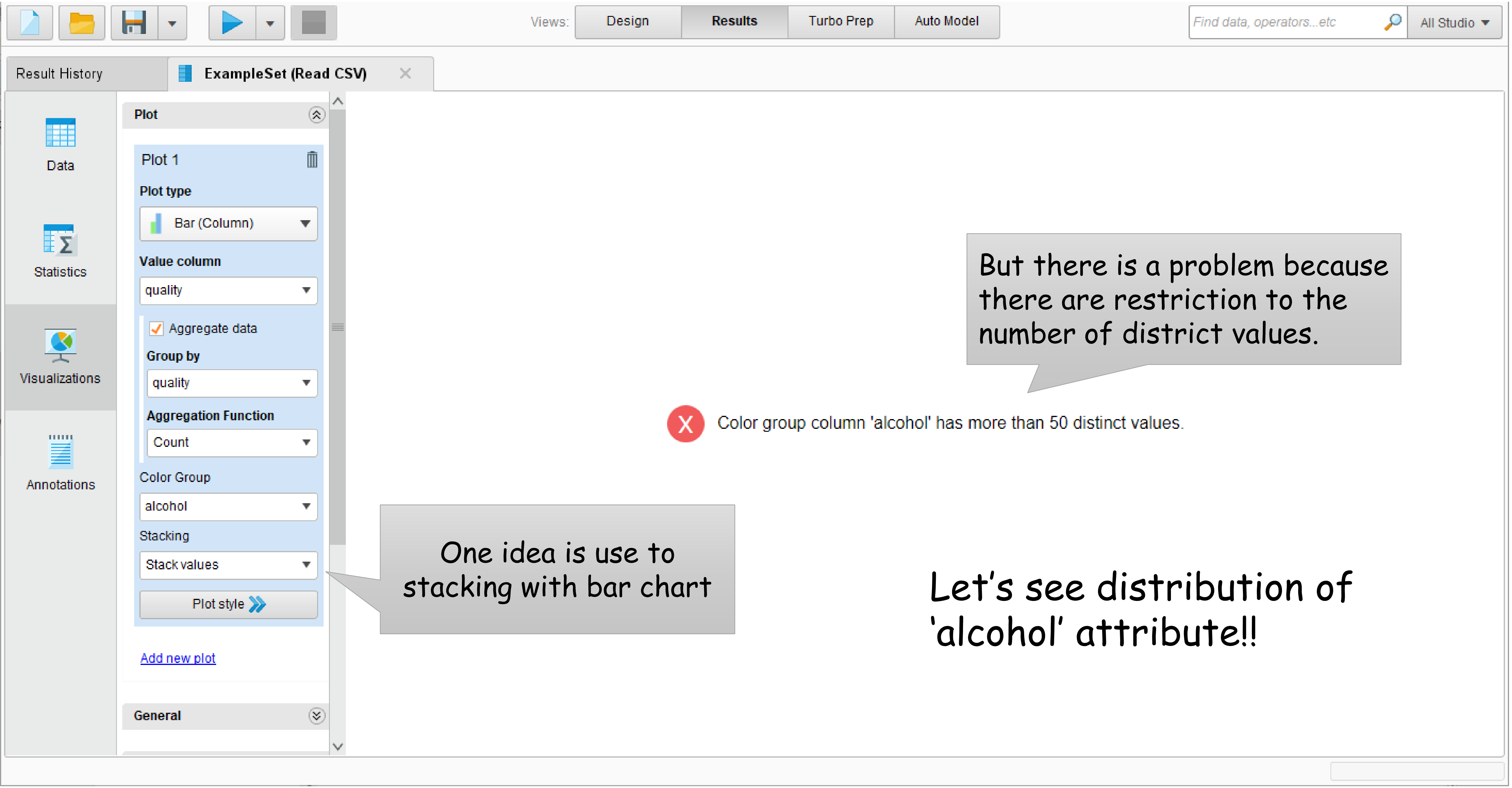
Stacking: No stacking

Plot style

When "Open visualization" is clicked, you can see bar chart, a default chart for nominal attribute value type

Say alcohol affects quality. How do you know how the levels of alcohol are distributed at each quality level?

Visualize single nominal attribute with a bar chart



Plot type: Bar (Column)

Value column: quality

☑ Aggregate data

Group by: quality

Aggregation Function: Count

Color Group: alcohol

Stacking: Stack values

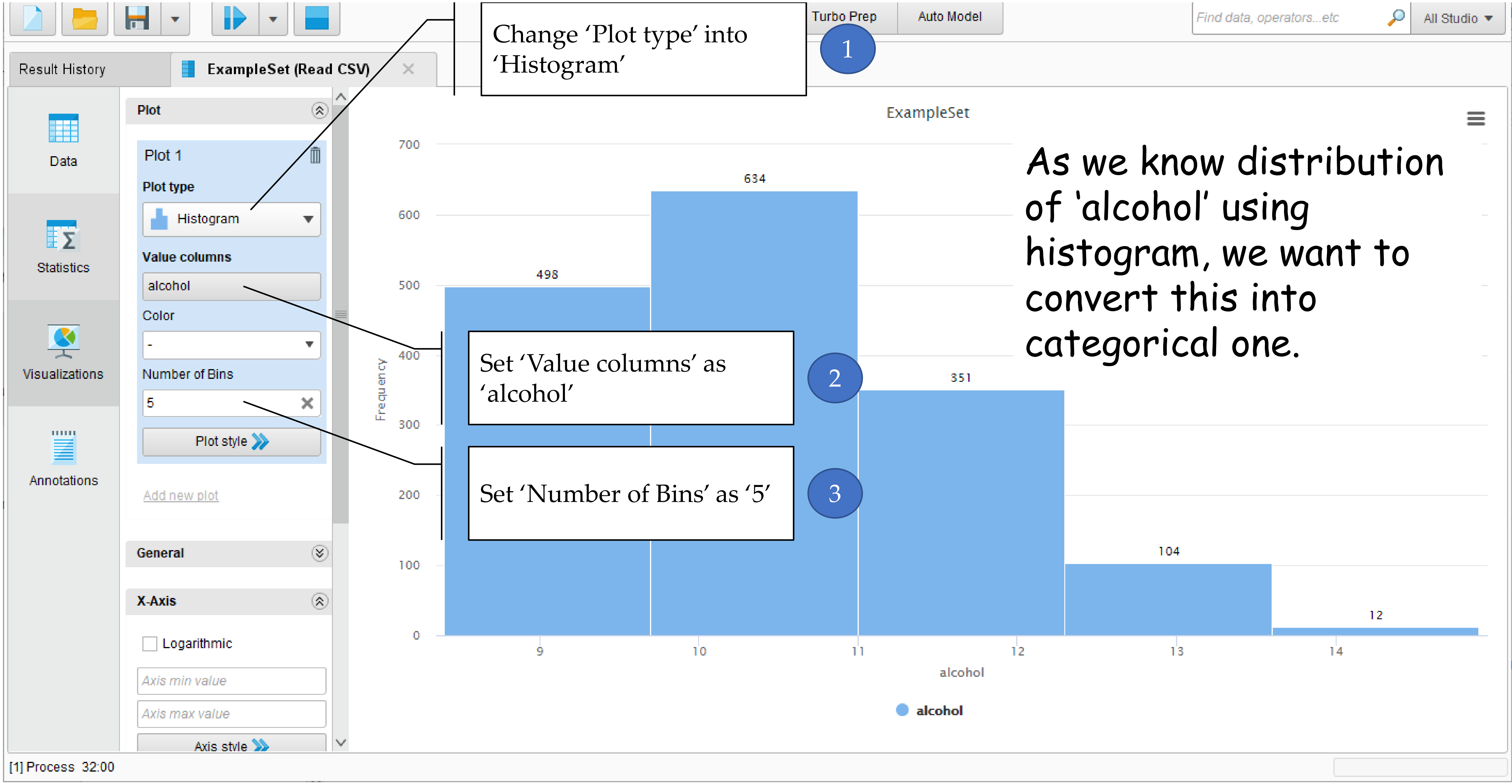
Color group column 'alcohol' has more than 50 distinct values.

But there is a problem because there are restriction to the number of district values.

One idea is use to stacking with bar chart

Let’s see distribution of ‘alcohol’ attribute!!

Visualize single numerical attribute with a histogram



1. Change 'Plot type' into 'Histogram'
2. Set 'Value columns' as 'alcohol'
3. Set 'Number of Bins' as '5'

As we know distribution of 'alcohol' using histogram, we want to convert this into categorical one.

Result History | ExampleSet (Read CSV)

Plot

Plot 1

Plot type: Histogram

Value columns: alcohol

Color: -

Number of Bins: 5

General

X-Axis

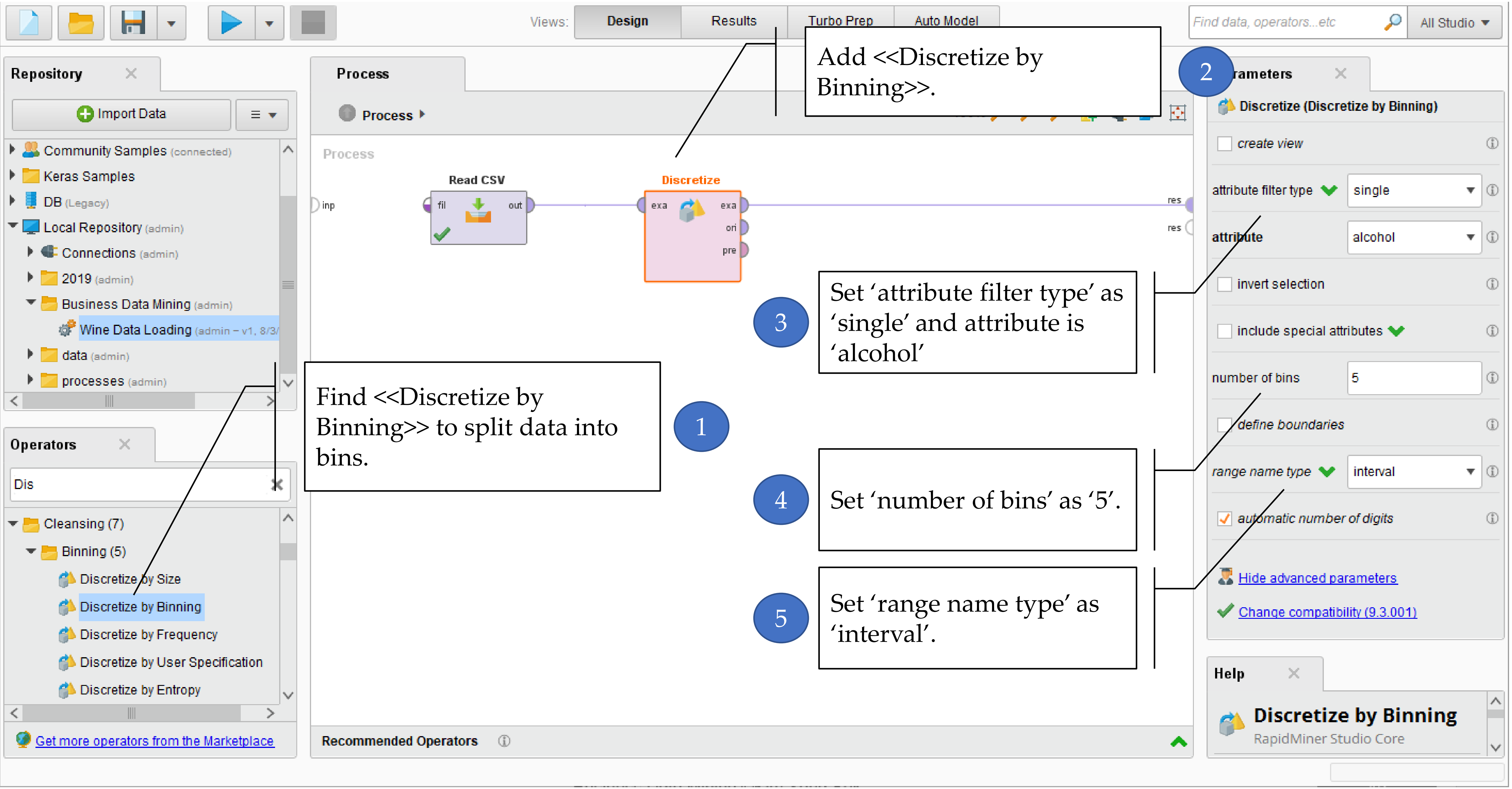
□ Logarithmic

Axis min value

Axis max value

[1] Process 32:00

Use <<Discretize by Binning>> to generate categorized attribute.



1. Find <<Discretize by Binning>> to split data into bins.

2. Add <<Discretize by Binning>>.

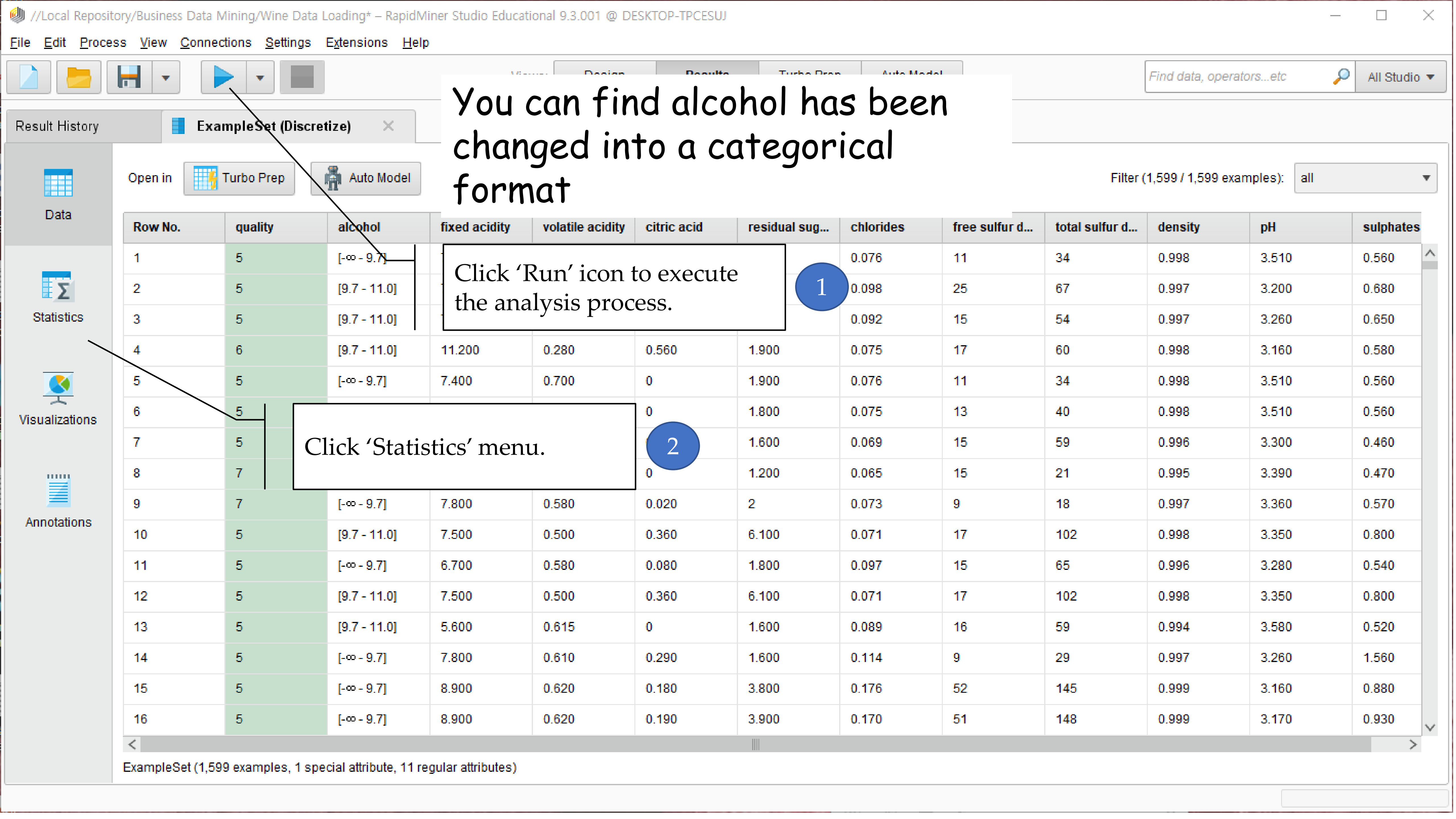
3. Set 'attribute filter type' as 'single' and attribute is 'alcohol'

4. Set 'number of bins' as '5'.

5. Set 'range name type' as 'interval'.

Parameters – Discretize (Discretize by Binning)

- □ create view
- attribute filter type: single
- attribute: alcohol
- □ invert selection
- □ include special attributes
- number of bins: 5
- □ define boundaries
- range name type: interval
- ☑ automatic number of digits


//Local Repository/Business Data Mining/Wine Data Loading* – RapidMiner Studio Educational 9.3.001 @ DESKTOP-TPCESUJ
File
Edit
Process
View
Connections
Settings
Extensions
Help
Find data, operators...etc
All Studio
Result History
ExampleSet (Discretize)
Data
Statistics
Visualizations
Annotations
Open in
Turbo Prep
Auto Model
Filter (1,599 / 1,599 examples):
all
1
2

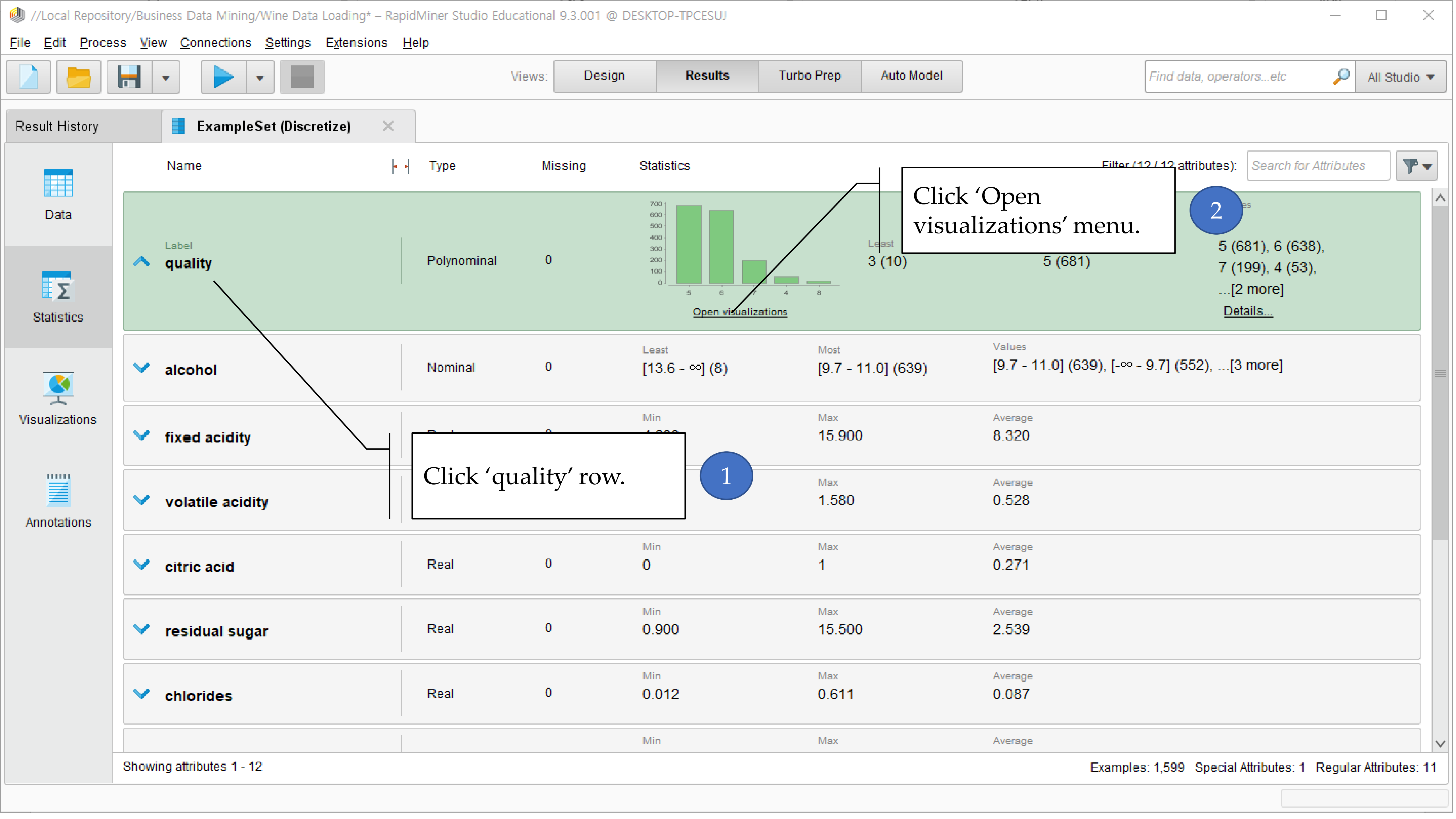
You can find alcohol has been changed into a categorical format

Click 'Run' icon to execute the analysis process.

Click 'Statistics' menu.

Row No.	quality	alcohol	fixed acidity	volatile acidity	citric acid	residual sug...	chlorides	free sulfur d...	total sulfur d...	density	pH	sulphates
1	5	[-∞ - 9.7]					0.076	11	34	0.998	3.510	0.560
2	5	[9.7 - 11.0]					0.098	25	67	0.997	3.200	0.680
3	5	[9.7 - 11.0]					0.092	15	54	0.997	3.260	0.650
4	6	[9.7 - 11.0]	11.200	0.280	0.560	1.900	0.075	17	60	0.998	3.160	0.580
5	5	[-∞ - 9.7]	7.400	0.700	0	1.900	0.076	11	34	0.998	3.510	0.560
6	5				0	1.800	0.075	13	40	0.998	3.510	0.560
7	5					1.600	0.069	15	59	0.996	3.300	0.460
8	7				0	1.200	0.065	15	21	0.995	3.390	0.470
9	7	[-∞ - 9.7]	7.800	0.580	0.020	2	0.073	9	18	0.997	3.360	0.570
10	5	[9.7 - 11.0]	7.500	0.500	0.360	6.100	0.071	17	102	0.998	3.350	0.800
11	5	[-∞ - 9.7]	6.700	0.580	0.080	1.800	0.097	15	65	0.996	3.280	0.540
12	5	[9.7 - 11.0]	7.500	0.500	0.360	6.100	0.071	17	102	0.998	3.350	0.800
13	5	[9.7 - 11.0]	5.600	0.615	0	1.600	0.089	16	59	0.994	3.580	0.520
14	5	[-∞ - 9.7]	7.800	0.610	0.290	1.600	0.114	9	29	0.997	3.260	1.560
15	5	[-∞ - 9.7]	8.900	0.620	0.180	3.800	0.176	52	145	0.999	3.160	0.880
16	5	[-∞ - 9.7]	8.900	0.620	0.190	3.900	0.170	51	148	0.999	3.170	0.930

ExampleSet (1,599 examples, 1 special attribute, 11 regular attributes)


//Local Repository/Business Data Mining/Wine Data Loading* – RapidMiner Studio Educational 9.3.001 @ DESKTOP-TPCESUJ
File Edit Process View Connections Settings Extensions Help
Views: Design Results Turbo Prep Auto Model
Find data, operators...etc
All Studio
Result History
ExampleSet (Discretize)
Data
Statistics
Visualizations
Annotations
Filter (12 / 12 attributes):
Search for Attributes
700
600
500
400
300
200
100
0
5
6
7
4
8
Open visualizations
Details...

Name	Type	Missing	Statistics		
Label quality	Polynominal	0	Least 3 (10)	5 (681)	5 (681), 6 (638), 7 (199), 4 (53), ...[2 more]
alcohol	Nominal	0	Least [13.6 - ∞] (8)	Most [9.7 - 11.0] (639)	Values [9.7 - 11.0] (639), [-∞ - 9.7] (552), ...[3 more]
fixed acidity			Min	Max 15.900	Average 8.320
volatile acidity				Max 1.580	Average 0.528
citric acid	Real	0	Min 0	Max 1	Average 0.271
residual sugar	Real	0	Min 0.900	Max 15.500	Average 2.539
chlorides	Real	0	Min 0.012	Max 0.611	Average 0.087
			Min	Max	Average

Showing attributes 1 - 12

Examples: 1,599 Special Attributes: 1 Regular Attributes: 11

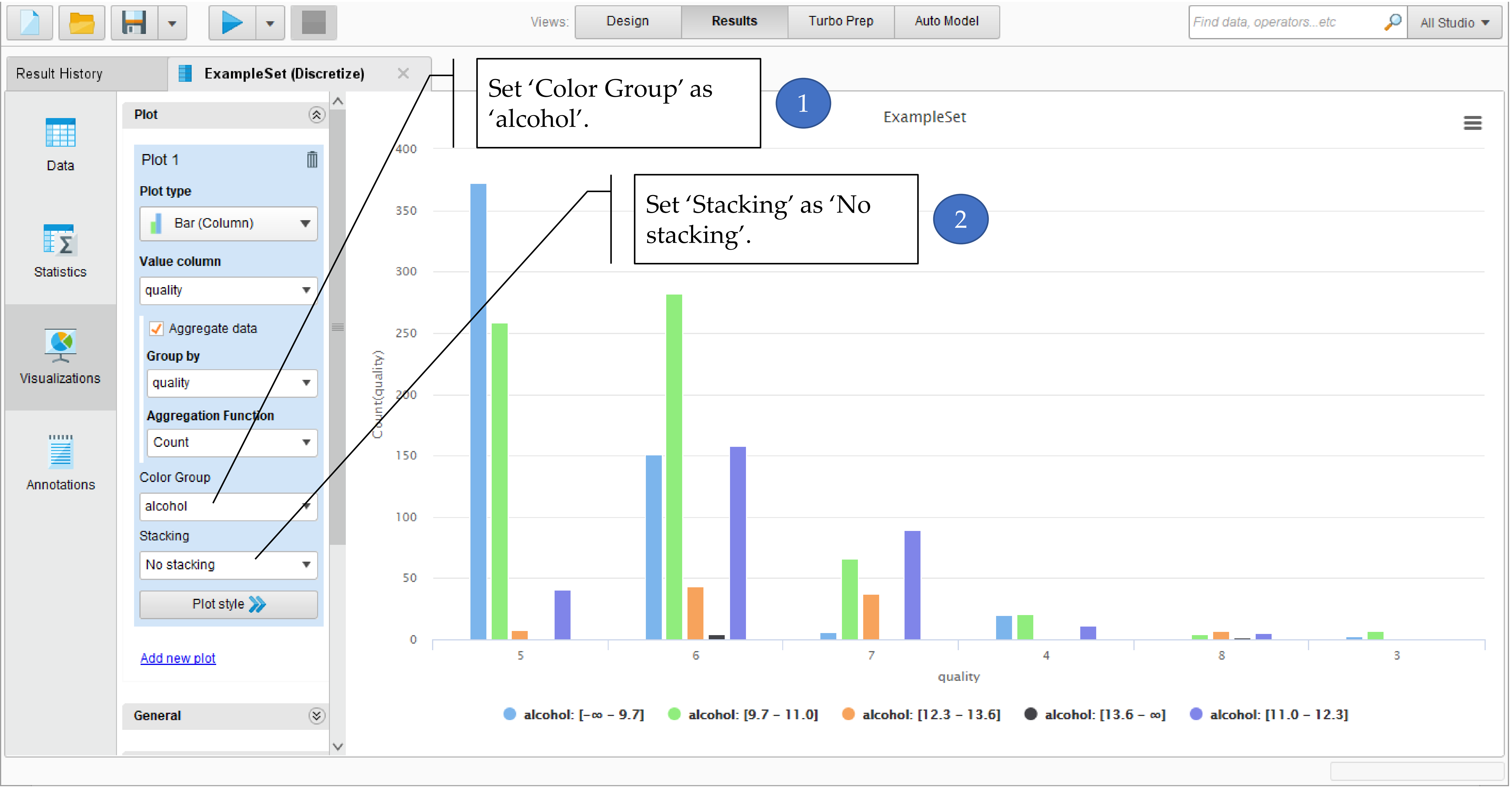
Click ‘Open visualizations’ menu.

2

Click ‘quality’ row.

1

Visualize multiple nominal attributes with a stacked bar chart



1 Set ‘Color Group’ as ‘alcohol’.

2 Set ‘Stacking’ as ‘No stacking’.

Plot 1

Plot type: Bar (Column)

Value column: quality

☑ Aggregate data

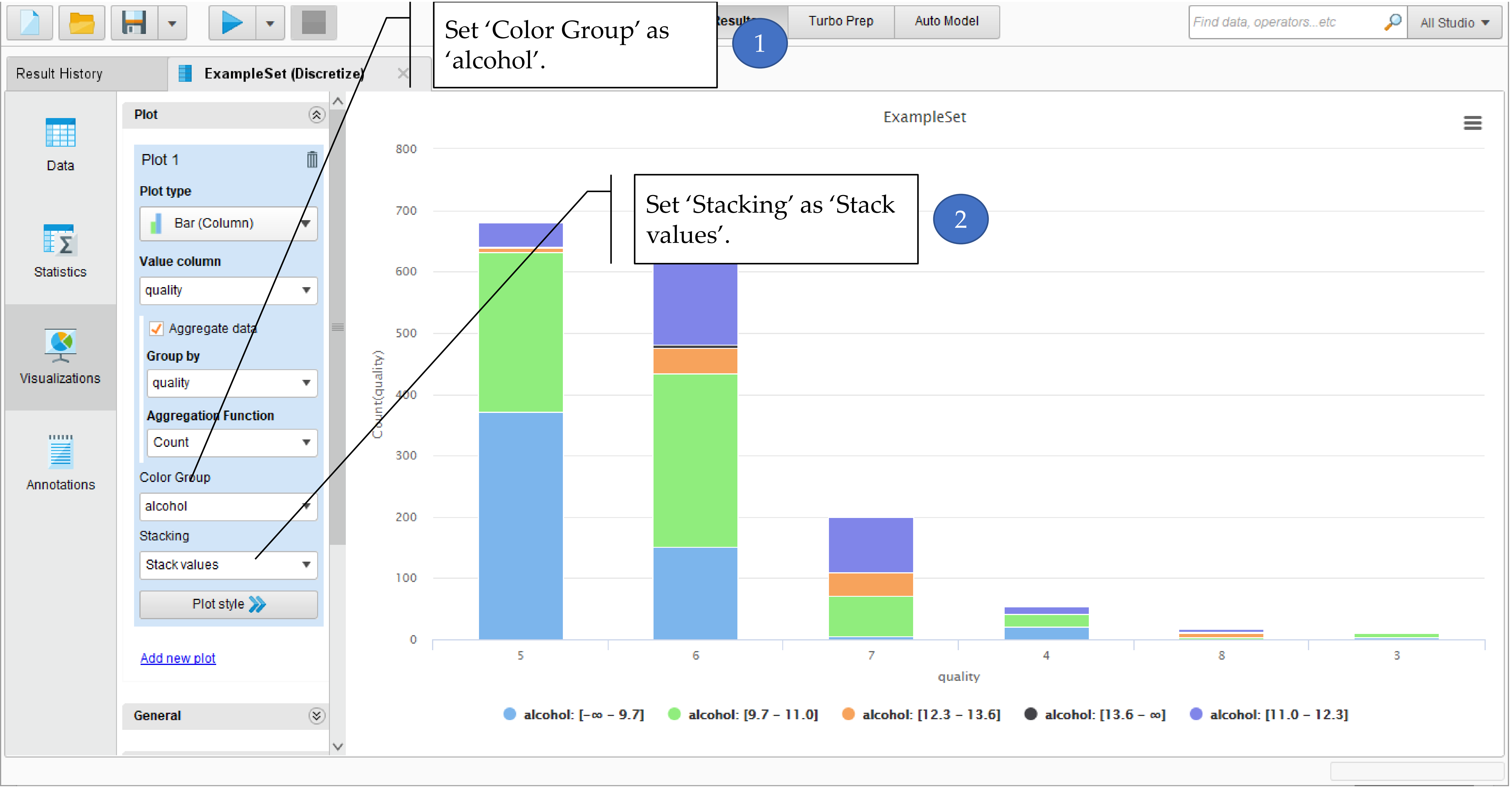
Group by: quality

Aggregation Function: Count

Color Group: alcohol

Stacking: No stacking

Visualize multiple nominal attributes with a stacked bar chart



1 Set ‘Color Group’ as ‘alcohol’.

2 Set ‘Stacking’ as ‘Stack values’.

Plot

Plot 1

Plot type: Bar (Column)

Value column: quality

☑ Aggregate data

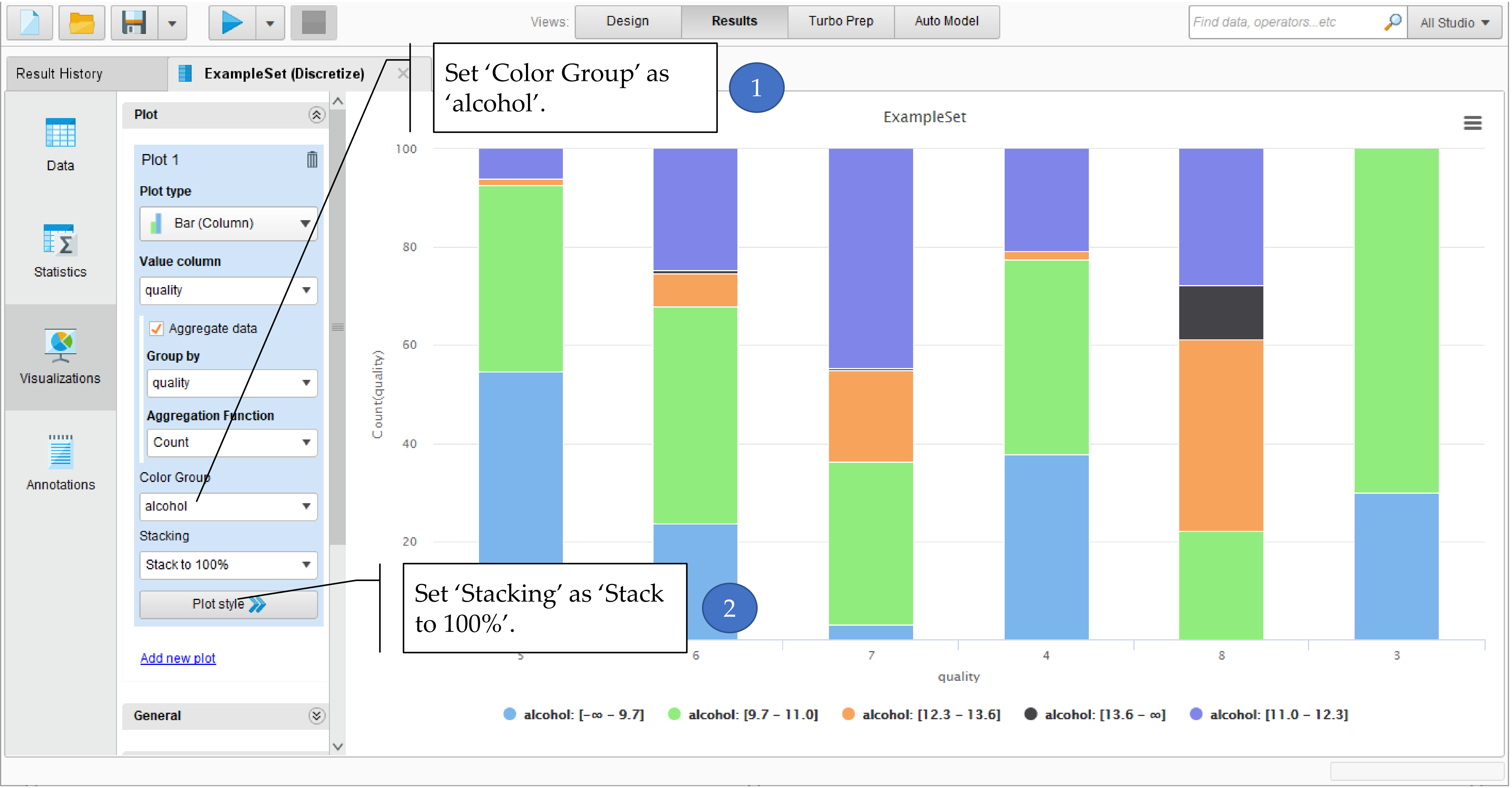
Group by: quality

Aggregation Function: Count

Color Group: alcohol

Stacking: Stack values

Visualize multiple nominal attributes with a stacked bar chart



Set 'Color Group' as 'alcohol'. 1

Set 'Stacking' as 'Stack to 100%'. 2

Plot 1

Plot type: Bar (Column)

Value column: quality

☑ Aggregate data

Group by: quality

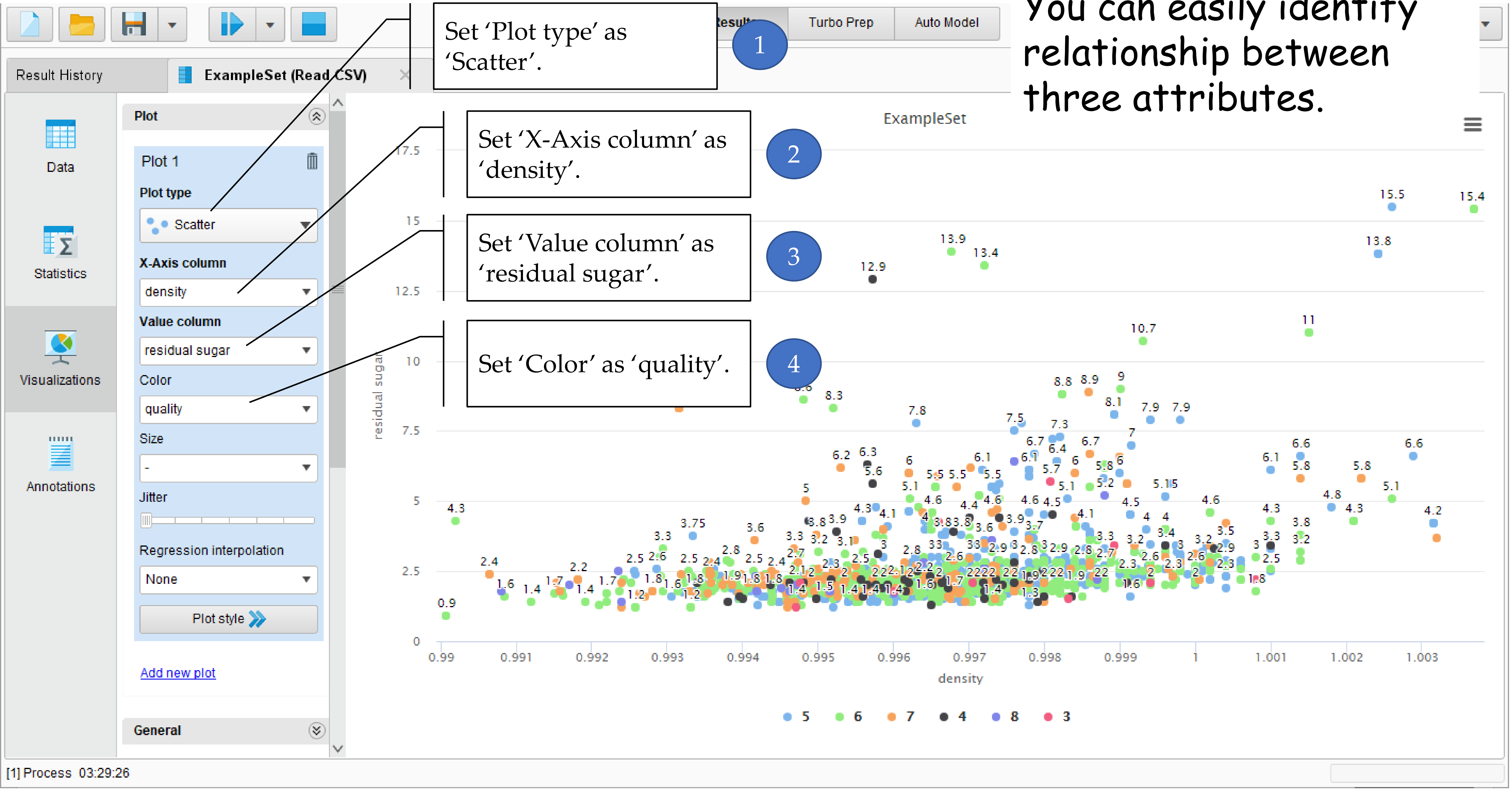
Aggregation Function: Count

Color Group: alcohol

Stacking: Stack to 100%

Visualize three attributes with a scatter plot

You can easily identify relationship between three attributes.



1. Set 'Plot type' as 'Scatter'.
2. Set 'X-Axis column' as 'density'.
3. Set 'Value column' as 'residual sugar'.
4. Set 'Color' as 'quality'.

Result History | ExampleSet (Read CSV)

Plot

Plot 1

Plot type: Scatter

X-Axis column: density

Value column: residual sugar

Color: quality

Size: -

Jitter

Regression interpolation: None

Plot style

Add new plot

General

[1] Process 03:29:26

Conclusion

Conclusion

- **In Data Understanding phase, it is necessary to gather, describe and explore data and verify data quality.**

- **Through exercises, we learn how to load dataset and how to explore dataset with Rapidminer**

- **Now you need to have knowledge on data to be analyzed and go to next phase of CRISP-DM.**

QUESTIONS?