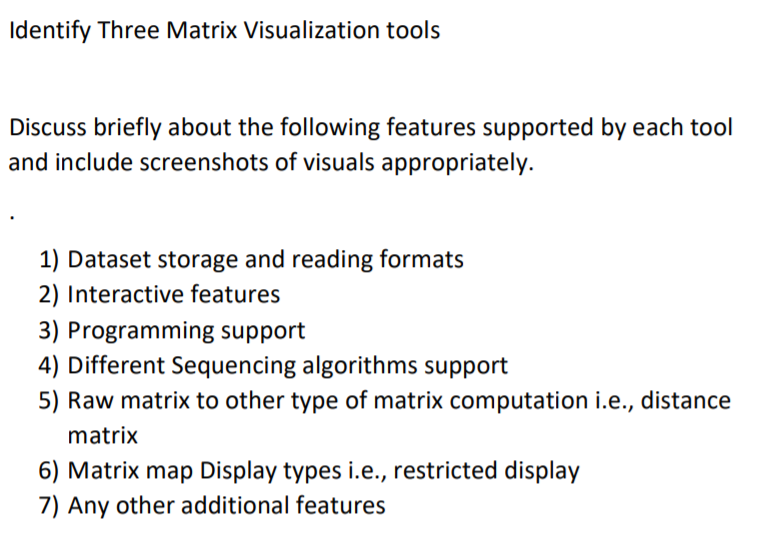
Data Visualization Digital Assignment 1

## Name: Om Ashish Mishra

## Registration Number: 16BCE0789

## Slot: C2

**The Question:**



**The Answer:**

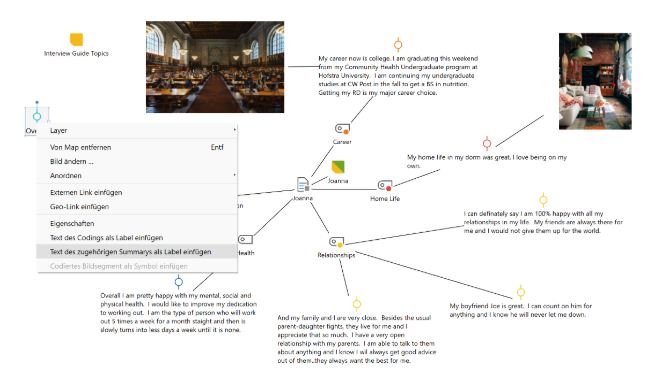
1. MAXQDA
2. Dataset storage and reading Format:

This can be in the form of documents and text and it can even store numbers.

1. Interactive features:

## MAXMaps – Create impressive Graphics

MAXMaps, the multi-functional graphics tool of MAXQDA, allows you to display your analytical process and its results in constant “live” connection to your data. Moreover, you may use the tool completely independently of your database in order to illustrate any given process, hypothesis or pattern in a graphic or diagram with an outstanding level of free creativity. Maps can also be a valuable tool for scientific explanation and can help to visualize complex relationships and theories.

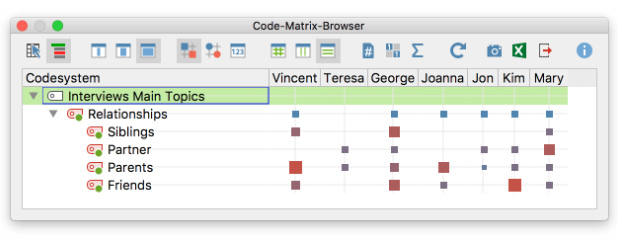


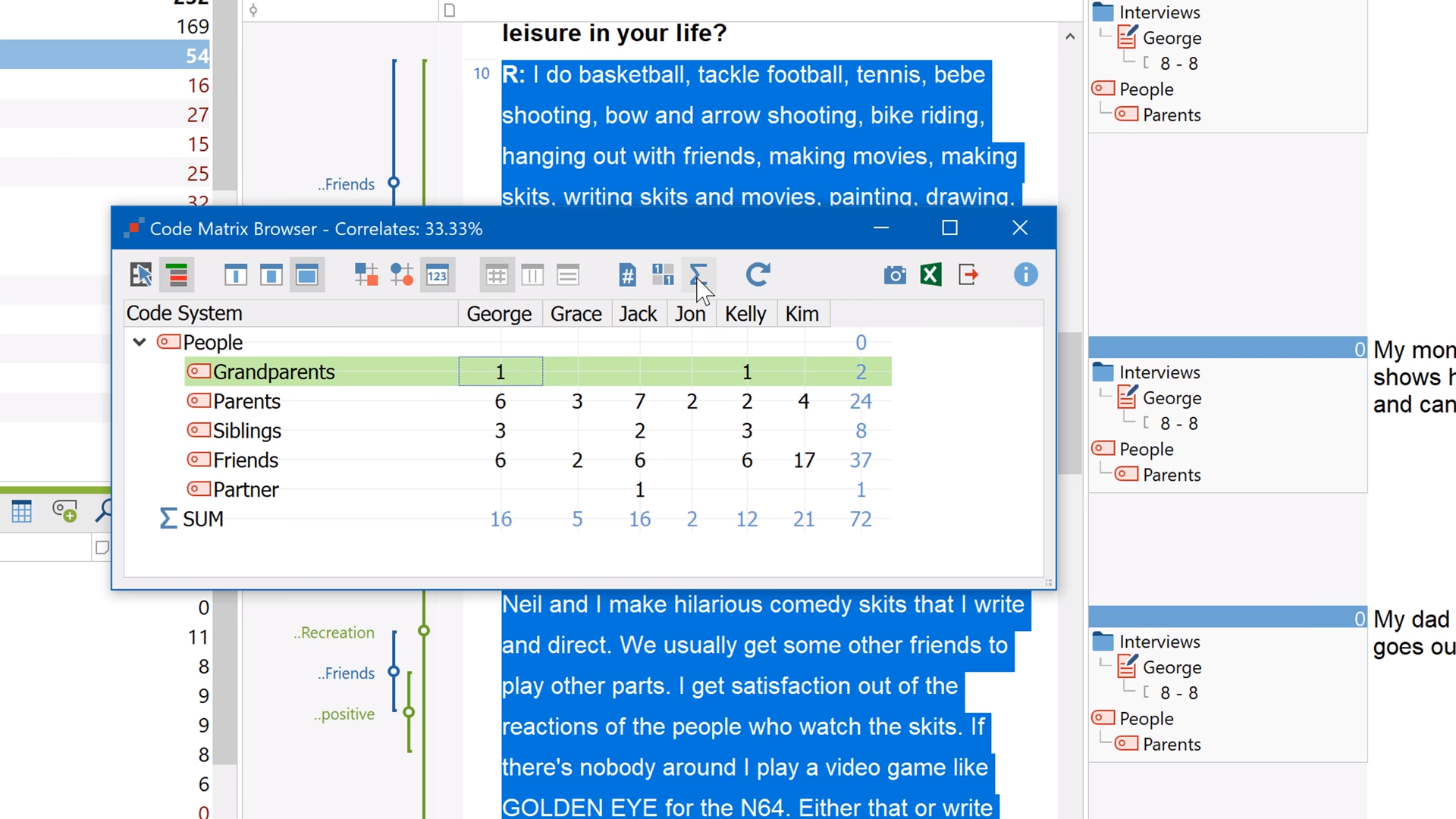
1. Programming Support

## Code Matrix Browser

The MAXQDA Code Matrix Browser offers you a completely new way of visualizing the distribution of codes throughout a selection or all of your texts. You will appreciate the easy and intuitive handling of the Code Matrix Browser.

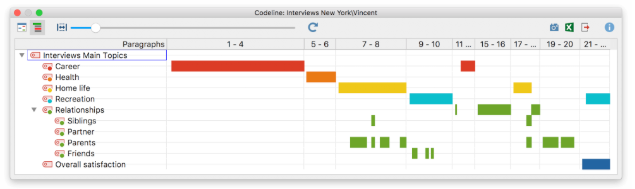
The matrix provides an overview of how many text segments from each text have been assigned a specific code and can be exported to a variety of formats.





## Codeline

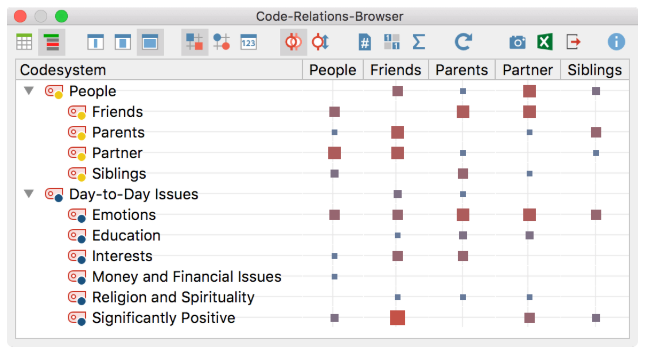
The MAXQDA Codeline offers a fast and comprehensive view of the codes assigned to a text to enhance your text analysis. The diagram is interactive – just click on a symbol in a cell of the matrix and the text passage will show up in your document window.



1. Different sequencing Algorithm Support

## Code Relations Browser

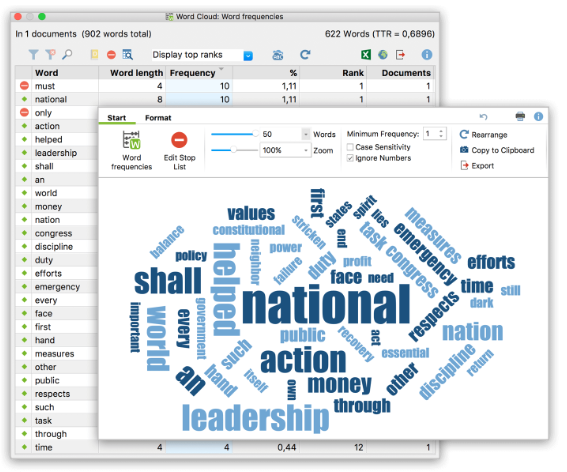
The Code Relations Browser shows the overlapping of your codes in a chosen document. This makes it easy to quickly identify possible connections between codes. The tool is also a good way to test the quality of your code system.



## Word Cloud

Use the MAXQDA Word Cloud to quickly see which words are used in your documents. The different font sizes tell you right away which words were used the most and which words were used less.

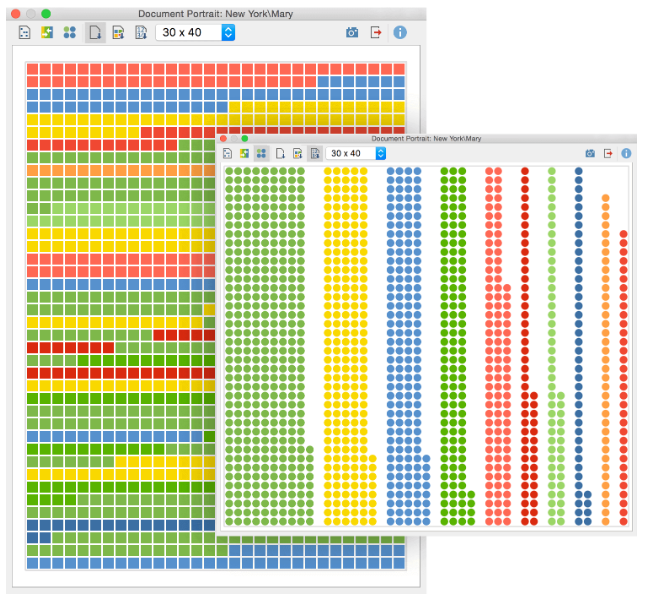
Word clouds offer you a simple way to visualize the most frequently used words in a document. Furthermore, MAXQDA puts the words in alphabetical order.



## Document Portrait

The Document Portrait displays any text as a „painting“ of either all or specific, selected codes assigned throughout this text. Choose colors for your code to assist you in your analysis, e.g. a special color for emotional factors (red) or economical factors (green).

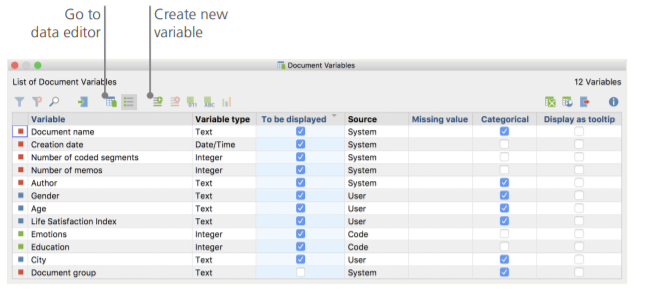
At one glance you will be able to tell which factors played an important role at what point of your interviews. You can of course change the assignment of colors at any time throughout your analytical process with a few clicks.



1. Raw Matrix Formation

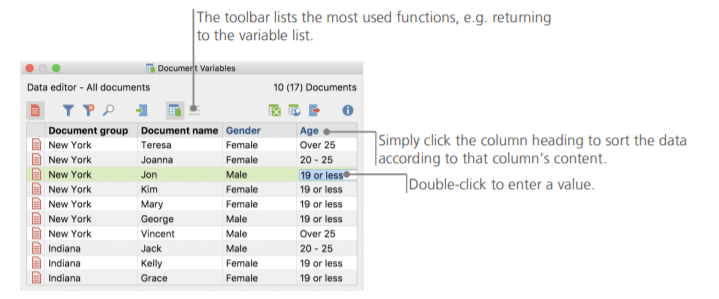
Defining document variables

One of the most important MAXQDA features is its support of Mixed Methods approaches. For every document in the Document System a whole set of attributes (that is, variables as used by statistics programs) can be managed. You can, for example, record personal data or background information for interviews as variables or variable values, and use them for your Mixed Methods analysis.



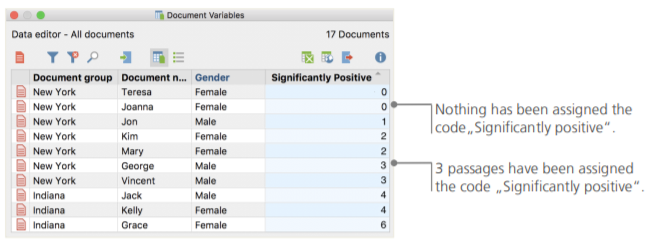
Entering variable values

To enter values for the Document variables, use the toolbar in the List of document variables and switch to the Data editor.



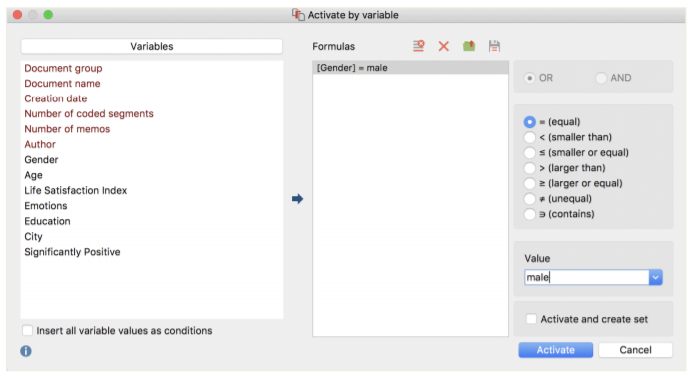
Transforming code frequencies into variables

The frequency of segments assigned to a certain code can be transformed into a variable, which will receive the code’s name. Right-click on the preferred code and choose Transform into a document variable in the context menu



Using document variables in your analysis

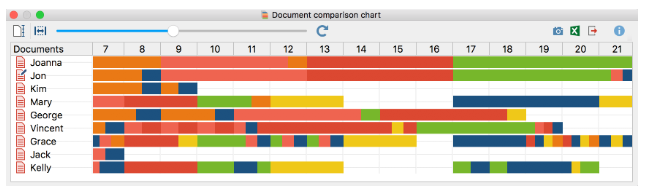
Quantitative and qualitative data can be connected in a variety of ways. The easiest way is to choose the document variables as selection criteria when you run a Coding Query. The relevant selection can be made with the function Activate by document variables in the “Mixed Methods” tab. First, choose a variable from the list on the left (in this case: Gender), press the arrow and then choose the variable value on the right (in this case: male). As soon as you click Activate, all documents with matching variable values will be activated in the Document System, in this example all interviews with male respondents.



1. Matrix Map Display Type

## Document Comparison Chart

A single mouse click presents a striking picture, comparing texts at a glance and letting you know which codes were assigned in which paragraphs. You are completely free to decide which codes and texts should be included in the comparison.

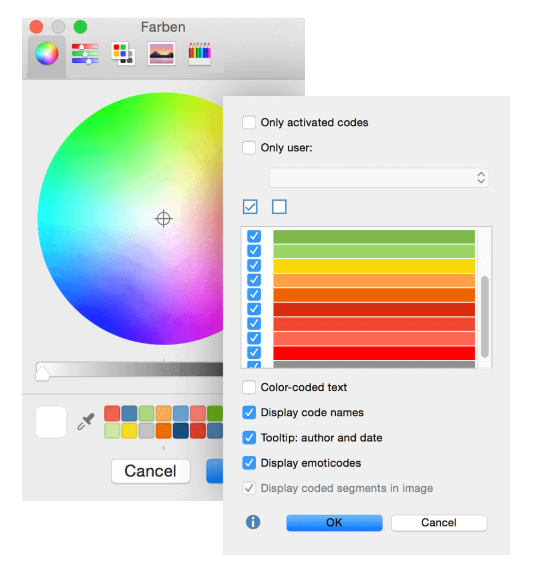


1. Additional Features

## Flexible color assignments

MAXQDA offers you more than 1,000,000 colors to use with your qualitative analysis. Assign them freely to codes, subcodes or texts.

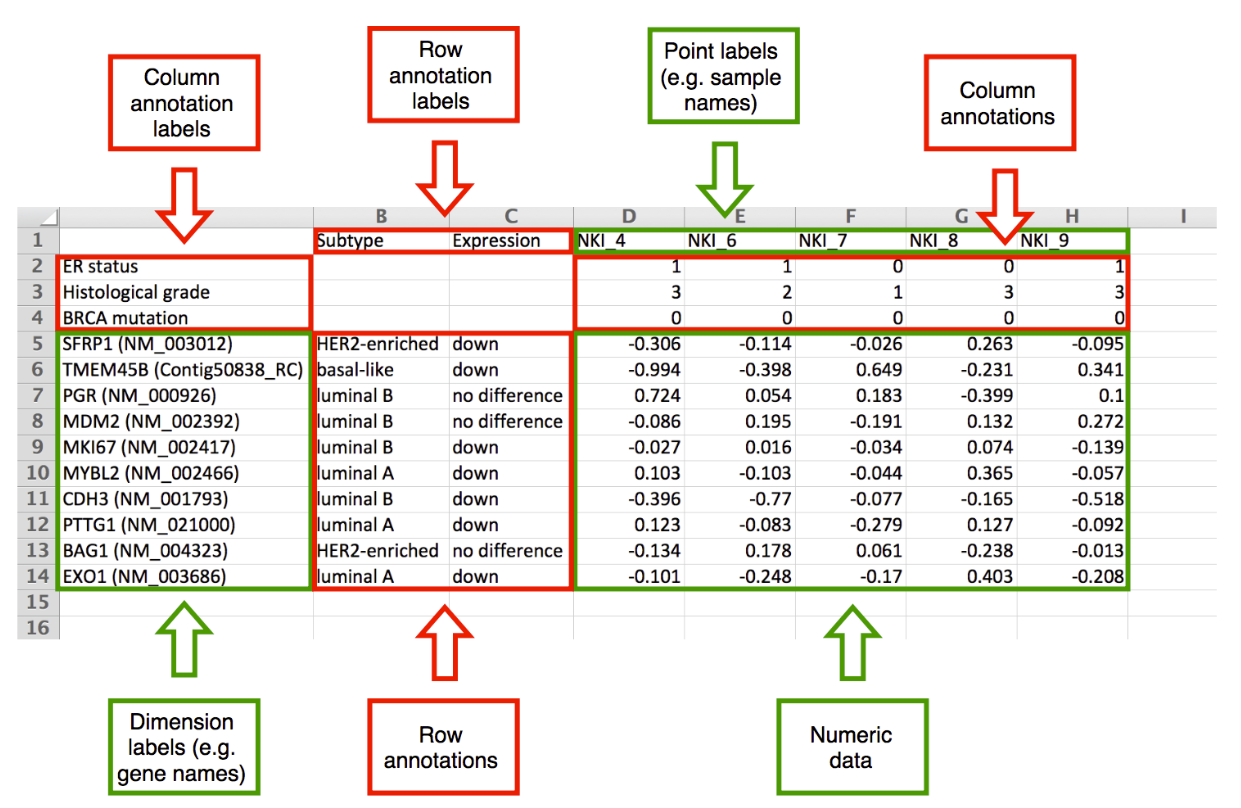
Color assignments are user-controlled and can be used functionally, not just merely as an aesthetic accessory. You can set any of your color stripes to invisible in order to work as easily and organized as possible (e.g. you can hide codings from another team member or from a specific code).



1. ClustVis
2. Dataset storage and reading Format:

##### Data import

We aimed for a simple input data format. The numeric data matrix is situated in the bottom right corner, dimensions presented in rows and points in columns. Row labels and annotations are left from the matrix, column labels and annotations are above the matrix. Annotation labels are in the first row and column, respectively. Format of the input file is shown on the image below. Annotations are optional, data sets without annotations can be uploaded as well (on the example image, omitting rows 2-4 and/or columns B and C). When taking data from spreadsheet program (e.g. MS Excel), you can copy-paste the data to 'Paste data' box or export the data as delimited text file (ending with .csv or .tab) and then upload this file to ClustVis. Uploading Excel native files directly (.xls or .xlsx) doesn't work.



In addition, it is possible to load settings that you have saved earlier (including data, drop-down settings etc.) or import data from [MEM](http://biit.cs.ut.ee/mem/) which has a collection of public datasets from [ArrayExpress](http://www.ebi.ac.uk/arrayexpress/" \t "_blank). The latter case is described more closely in the next section. It is also possible to use one of the testing datasets to get an overview of the web tool.

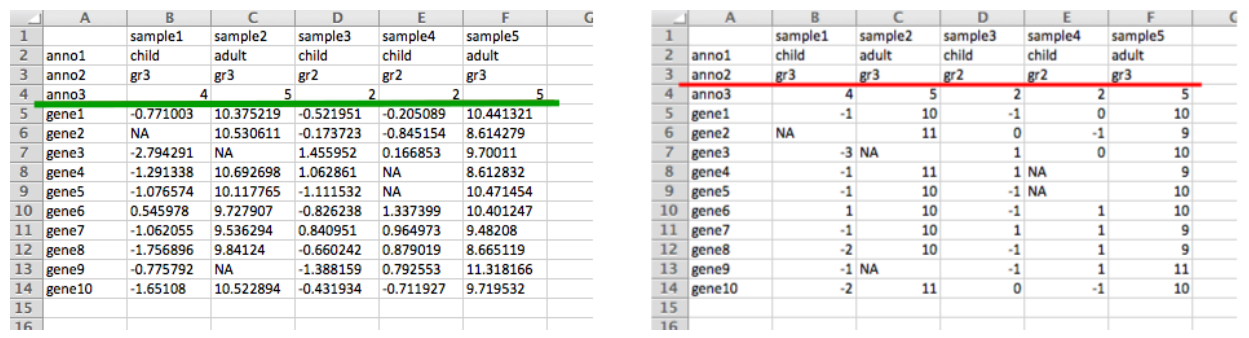
If your dataset is not uploaded correctly (no rows are shown), please check the following:

* Make sure file is chosen for upload or pasted to the text box.
* Make sure all rows have equal number of columns. In case of doubt, it is safer to open the data in a spreadsheet program and copy-paste from there rather than choosing a file for upload.
* Make sure there are no duplicate row or column names.
* Make sure all missing values are marked with NA or empty cell.
* If automatic detection of the delimiter is not working correctly, try to set it manually (uncheck the 'detect delimiter' checkbox).
* If automatic detection of the column and row annotations is not working correctly, try to set it manually (uncheck the 'detect column and row annotations' checkbox).

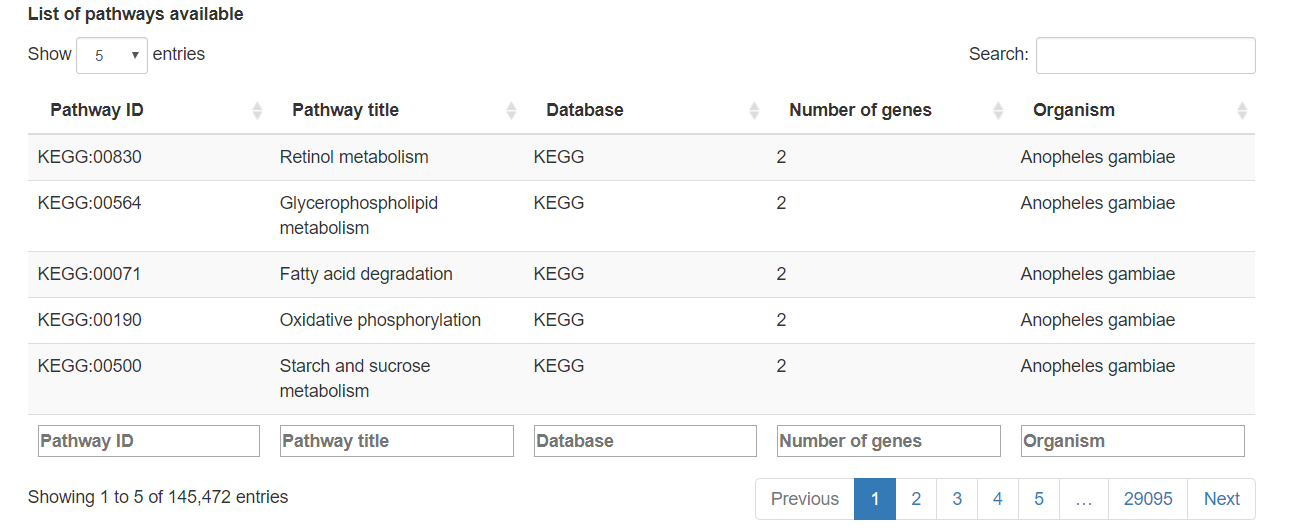
For user-uploaded datasets, ClustVis automatically detects both delimiter and number of annotation rows from the data by default. To find delimiter, it counts for each possible delimiter (comma, tabulator, semicolon) how many times it appears on each row. We use the heuristic where minimum is taken over all rows and the delimiter with the greatest score is chosen as the right one.

When finding number of annotation rows and columns, the largest possible numeric matrix from bottom right corner is found. If the matrix contains non-integers (i.e. fractional numbers) and there are some left columns or top rows that contain integers only, these first rows and/or columns are considered integer-valued annotations (e.g. some grade coded with integers: 1, 2, 3).

The situation is depicted on the following images where there are three annotation lines and the green and red line show automatic detection. On the left, the numeric matrix contains non-integer values and the last annotation row is detected correctly. On the right, the numeric matrix contains only integers and the last annotation row cannot be detected automatically.



It will look like:



1. Interactive features:

##### Interactivity of the plots

With the recent updates of the tool, interactive mode was added to both PCA plot and heatmap which allows to click and hover over specific areas of the plot. This mode is available when going to PCA or heatmap tab and choosing 'change plot labels', 'add interactivity'. Interactivity is still experimental, we are hoping to get feedback from users. It is not made the default option because plots are slower to render. It is recommended to first set other options in non-interactive mode and then switch to interactive mode as the last step. For larger datasets where an interactive plot would take too long to render, it is automatically switched to non-interactive mode and a warning message is shown.

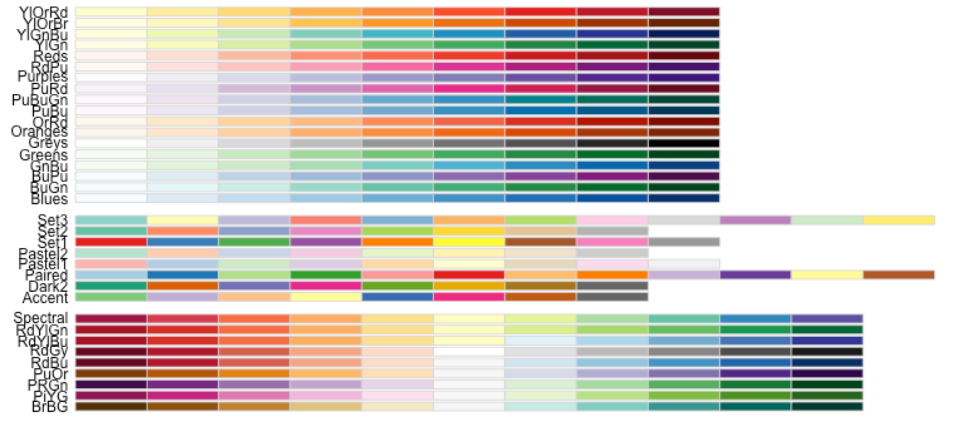
Interactive mode includes the following additional options:

* Hover over a point on the PCA plot to see additional information.
* Click on a point on the PCA plot to see values from one column on a separate jitter plot. If 'plot\_link' appears among column annotations, the points are linked to an external resource instead.
* Hover over a row name, column name or cell on the heatmap to see additional information.
* Click on a row name, column name or cell on the heatmap to see values from one row, column or cell on a separate jitter plot.
* Hover over a point on the jitter plot to see additional information.
* Click on a point on the jitter plot to go to an external resource. It is only available if there is a row or column annotation called 'plot\_link'.
* Click on a column name on the jitter plot to see a reduced jitter plot with values from only this column.

The default tooltips can be overridden by having a row or column annotation called 'plot\_tooltip'. Below the jitter plot, a table is shown that includes all data on the plot and, in addition, information about missing values. Row and column IDs in the table are changed into links if 'plot\_link' appears among annotations.

##### PCA and heatmap

These are the main tabs, allowing you to generate and customize PCA plot and heatmap. Each individual setting is described more precisely with a tooltip that appears if you hover over with the mouse. To download an image, you can use one of the buttons above the plot. The following color palettes from ColorBrewer are available:



Only up to eight color groups are allowed on the PCA plot because human eye cannot distinguish more colors easily. In this case, shapes should be enough for separating the groups. The following shapes are used:



1. Programming Support

##### Data pre-processing

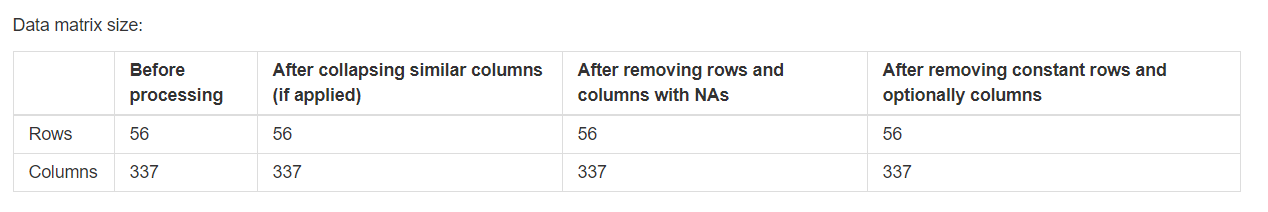
On this tab, you can choose the method that is used for PCA. This method is also used for imputing missing values to the heatmap and it also determines, for example, whether values on the heatmap are centered or not. Number of components returned depends on the dimensions of the input data matrix. If there are more observations (n) than dimensions (d) then d principal components are calculated. Otherwise, the number of principal components is n.

In general, SVD (Singular Value Decomposition) is the standard PCA method that is most often used. Imputation means that if there are any missing values in the dataset, they are predicted and filled iteratively using other values in the dataset during SVD calculation. Other PCA methods are further strategies to cope with the missing value problem.

Unit variance is the most common scaling method. This means that all variables are scaled so that they will be equally important (variance = 1) when finding the components. As a result, a difference of 1 means that the values are one standard deviation away from each other, or from the average of the row if rows are centered.

If the variables already have a common scale and you want to keep their original variability (i.e. variables where the variability is larger should contribute more to the components), you can apply 'no scaling'.

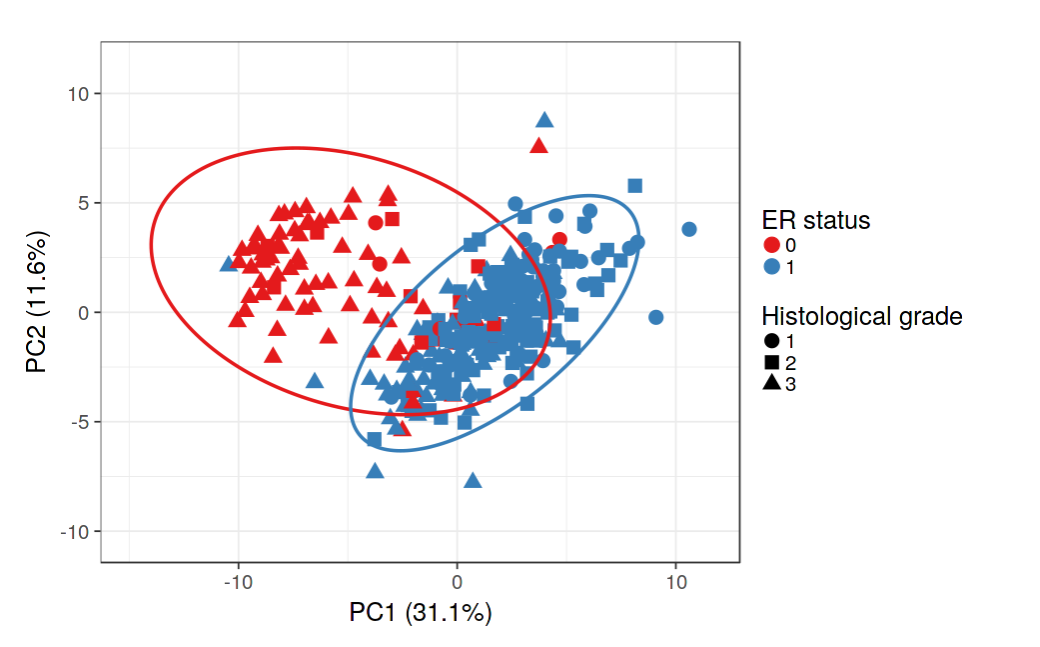
When exactly to use some other PCA method (Nipals, probabilistic) or scaling (Pareto, vector scaling) is a matter of testing with each specific dataset. Interpreting the results of PCA and heatmap plots is quite subjective and needs further validation using other methods.



1. Different sequencing Algorithm Support

PCA

Unit variance scaling is applied to rows; SVD with imputation is used to calculate principal components. X and Y axis show principal component 1 and principal component 2 that explain 31.1% and 11.6% of the total variance, respectively. Prediction ellipses are such that with probability 0.95, a new observation from the same group will fall inside the ellipse. N = 337 data points.



Heat Map

Rows are centered; unit variance scaling is applied to rows. Imputation is used for missing value estimation. Both rows and columns are clustered using correlation distance and average linkage. 56 rows, 337 columns.

Heatmap shows a data matrix where coloring gives an

overview of the numeric differences. In ClustVis, hierar-

chical clustering can be optionally applied to dimensions

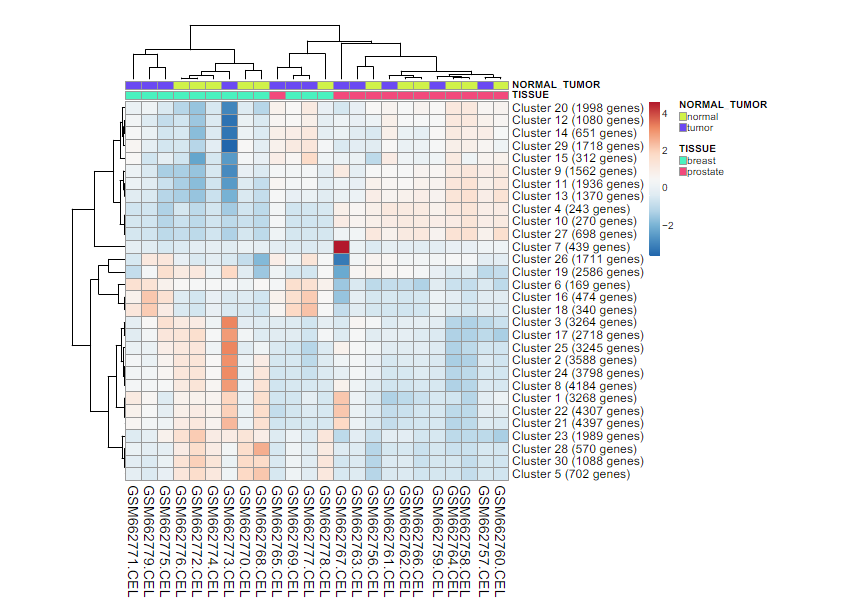
and/or observations. Users can choose which clustering

Heatmap shows a data matrix where coloring gives an

overview of the numeric differences. In ClustVis, hierar-

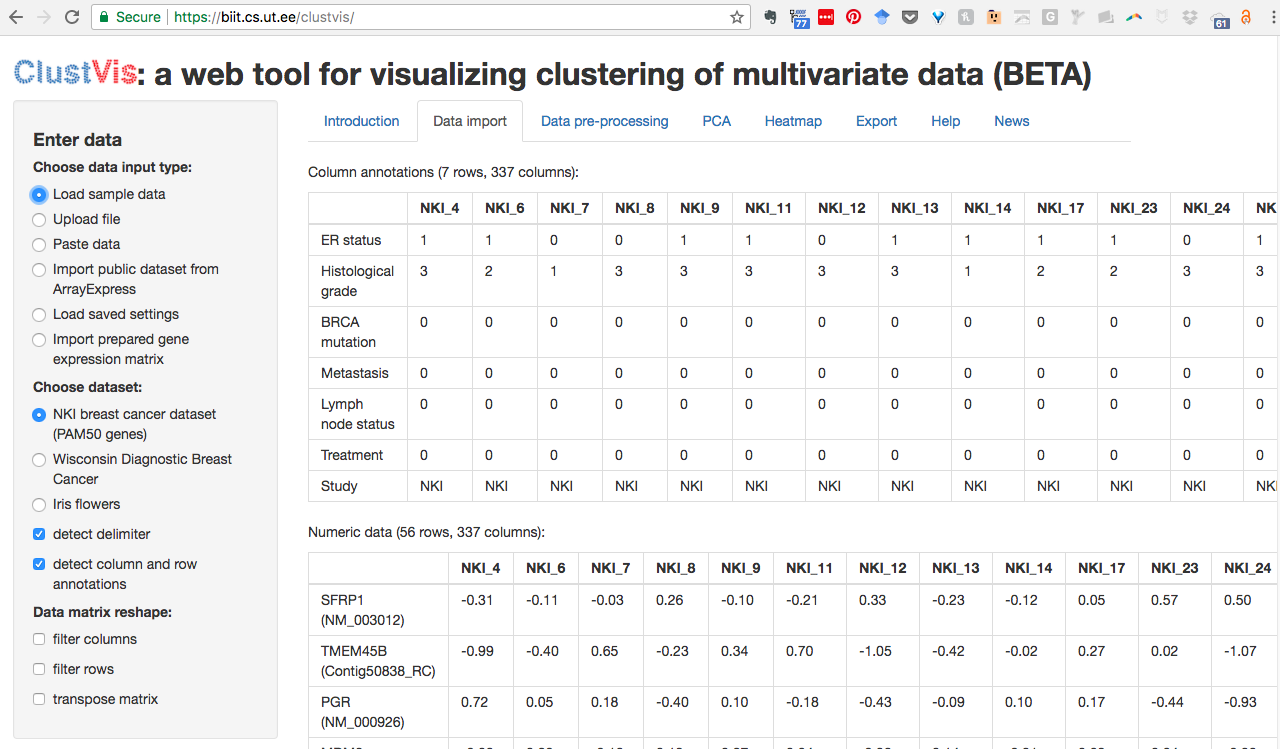
chical clustering can be optionally applied to dimensions

and/or observations. Users can choose which clustering



1. Raw Matrix Formation

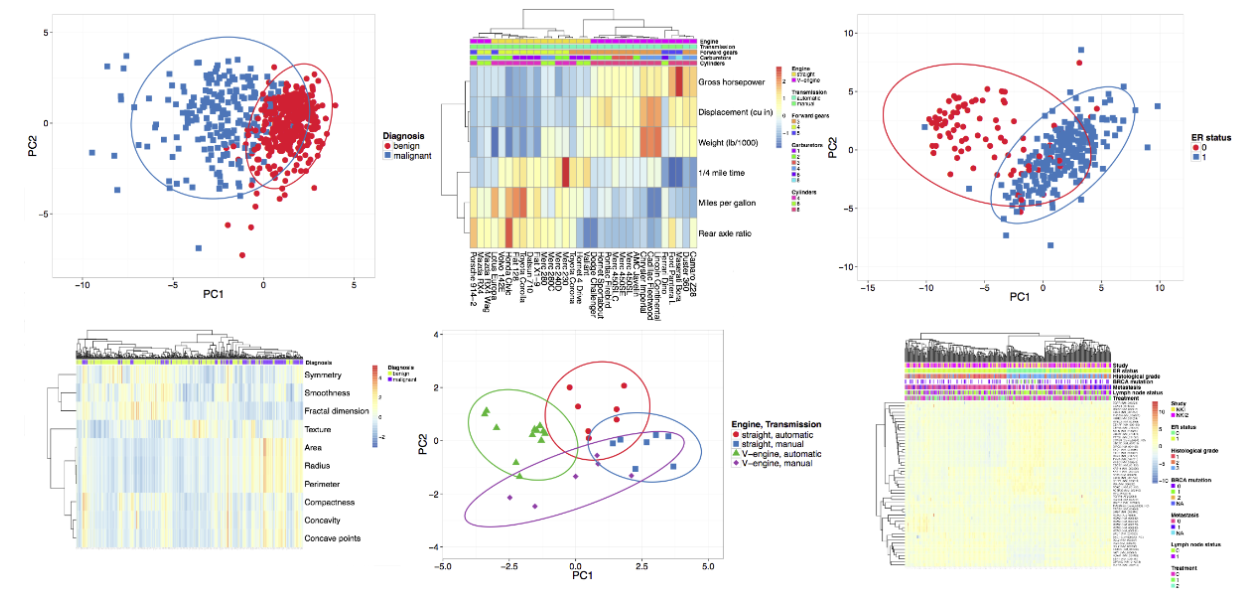
The Matrix formation by importing dataset into the ClustVis



1. Matrix Map Display Type

This web tool allows users to upload their own data and easily create Principal Component Analysis (PCA) plots and heatmaps. Data can be uploaded as a file or by copy-pasteing it to the text box. Data format is shown under "Help" tab.

The tool is inspired by discussions in [PREDECT](http://predect.eu/) project and borrows some code from [BoxPlotR](http://boxplot.tyerslab.com/" \t "_blank). Several [R](http://www.r-project.org/) packages are used internally, including shiny, ggplot2, pheatmap, gridSVG, RColorBrewer, FactoMineR, pcaMethods, gProfileR, shinyBS, shinyjs and others. It is developed in [BIIT](http://biit.cs.ut.ee/) Research Group. The source code of ClustVis is available in [GitHub](https://github.com/taunometsalu/ClustVis).



1. Additional Features

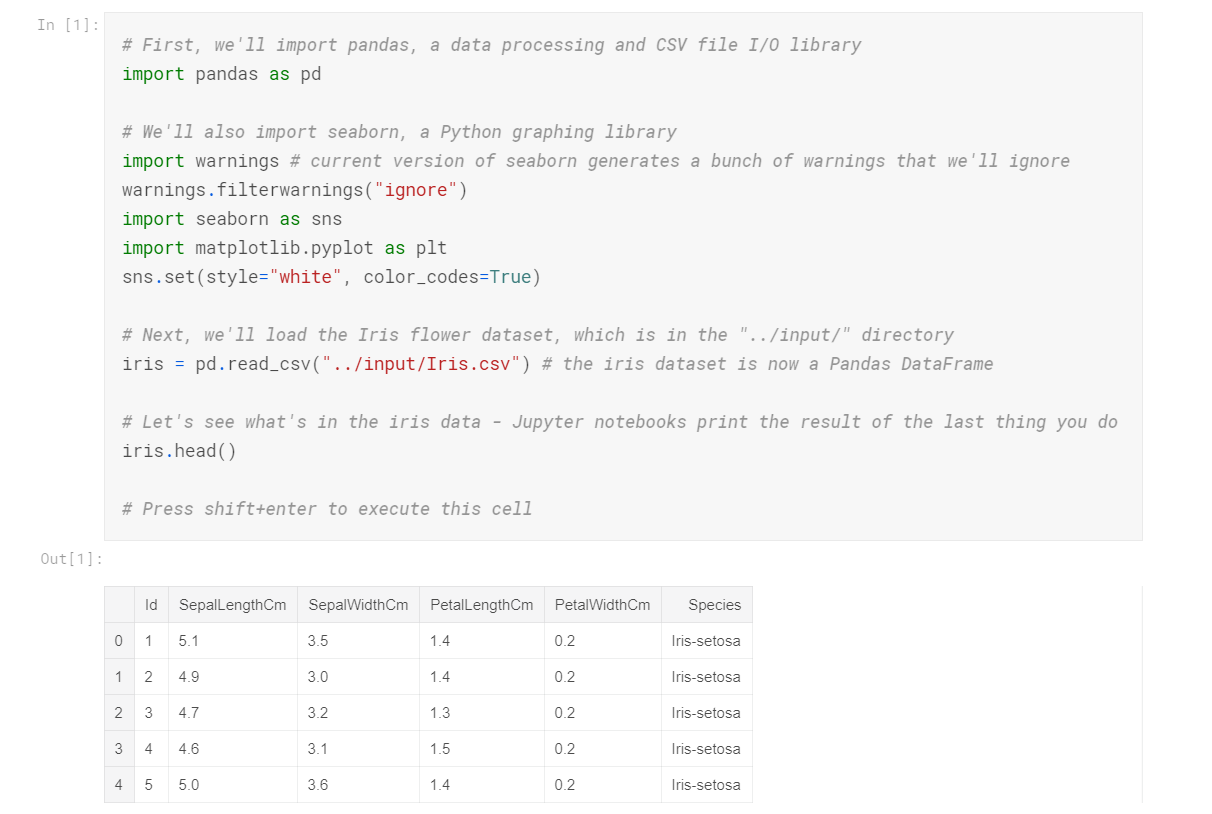
### **Export of analysis results**

Both PCA plot and heatmap can be downloaded using one of the three formats (PDF, EPS or SVG). It is also possible to download the initial data in case a public data set was used that was not available to the user before. Intermediate results, including processed data, PCA scores, PCA loadings and variance explained by the PCA components can be downloaded as well.

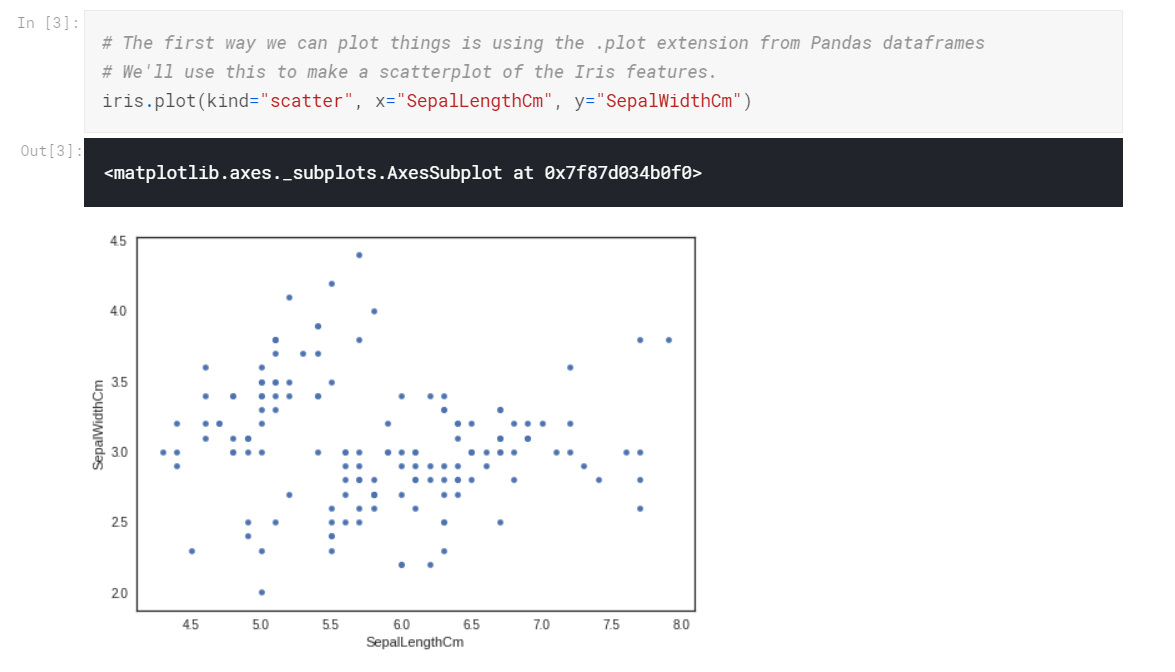
Both PCA and heatmap show an example caption above the plot that describes the data processing and visualization using full sentences. This makes it more convenient to describe the plot in an article written by a user of ClustVis.

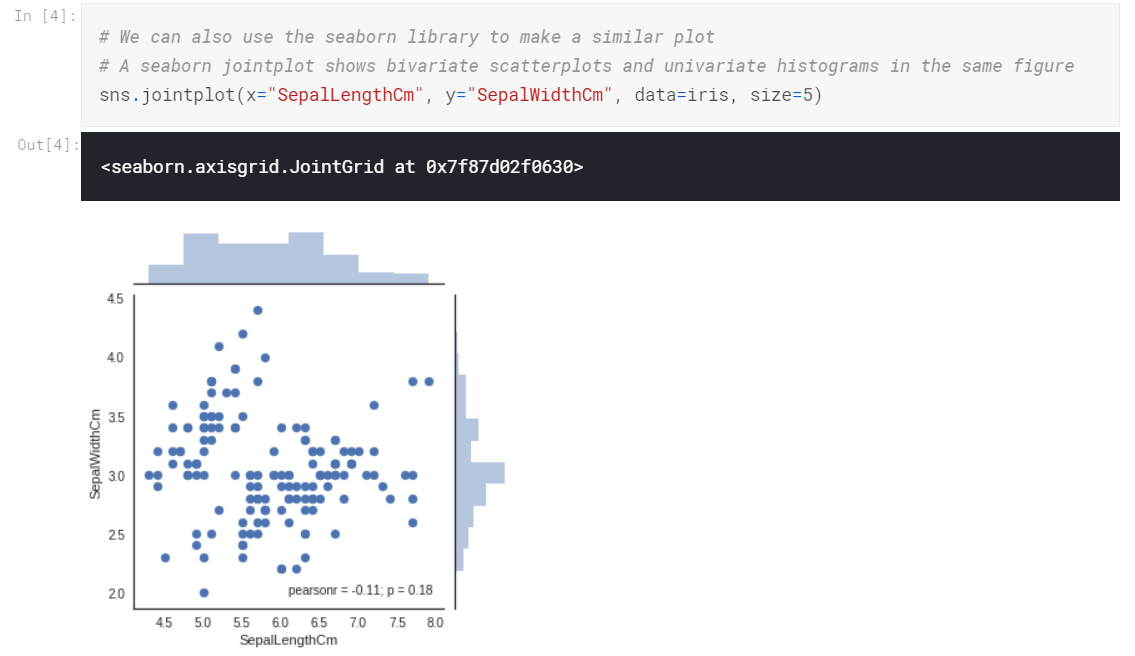
Due to the large number of settings available there is also an option to save them and generate a link that loads pre-saved settings. Subsequently, the link can be sent to collaborators to show current findings or keep it to return to the same data set later.

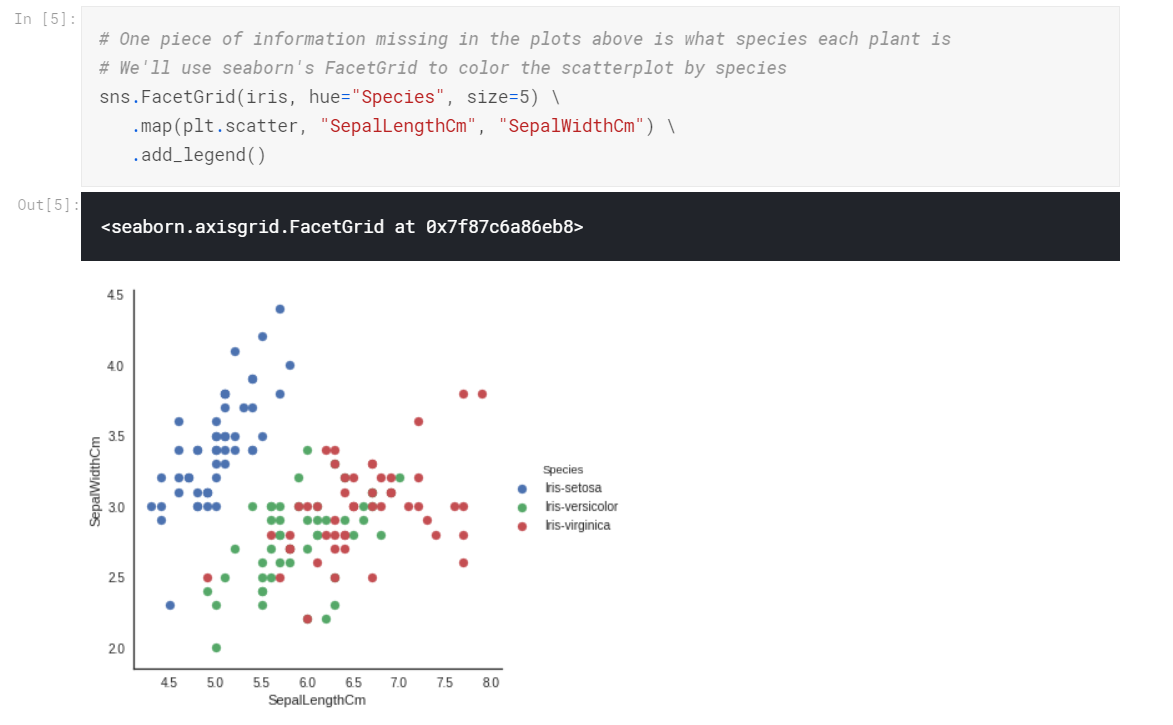
1. Python
2. Dataset storage and reading Format:

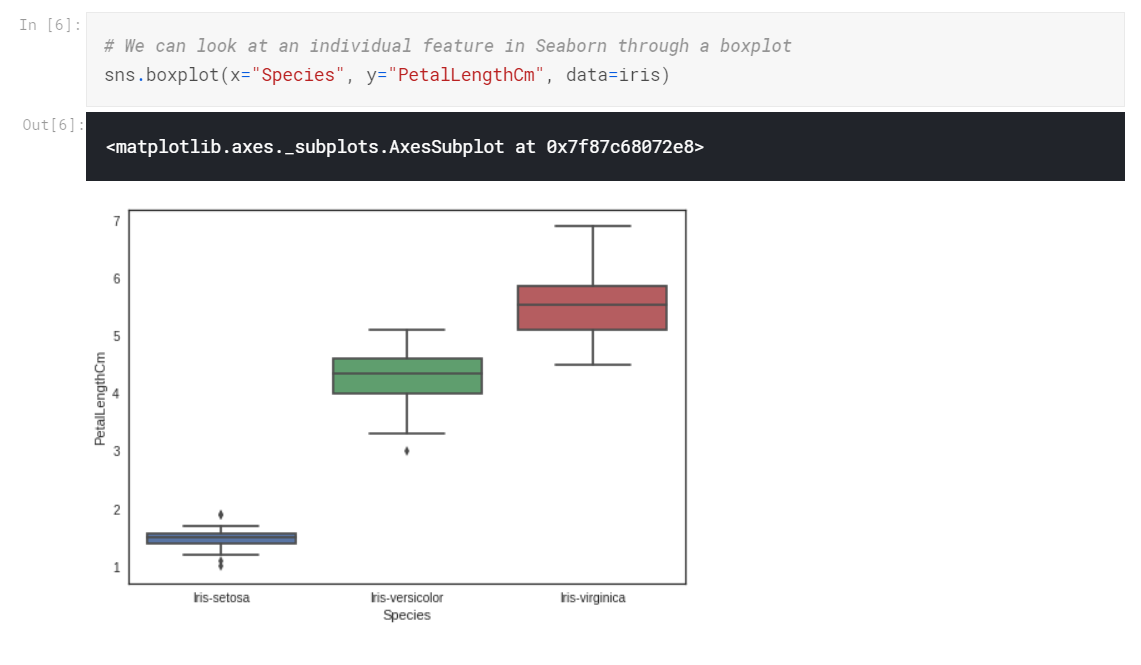


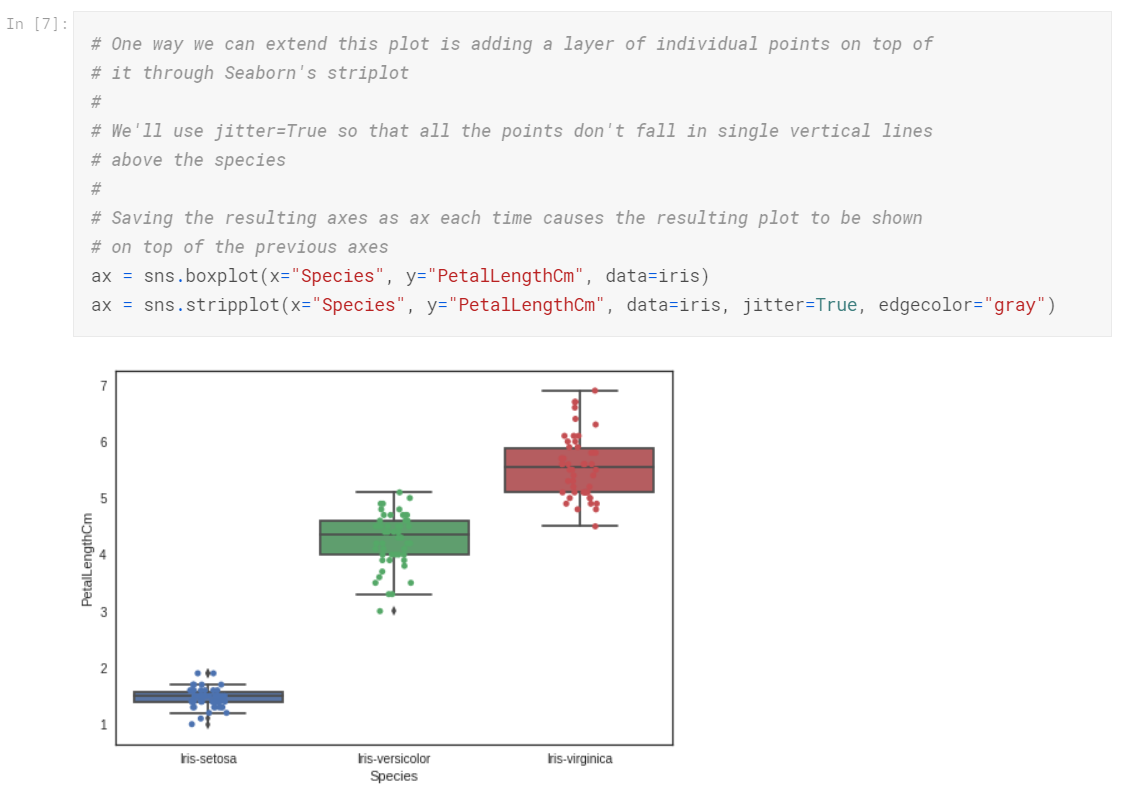
1. Interactive features:



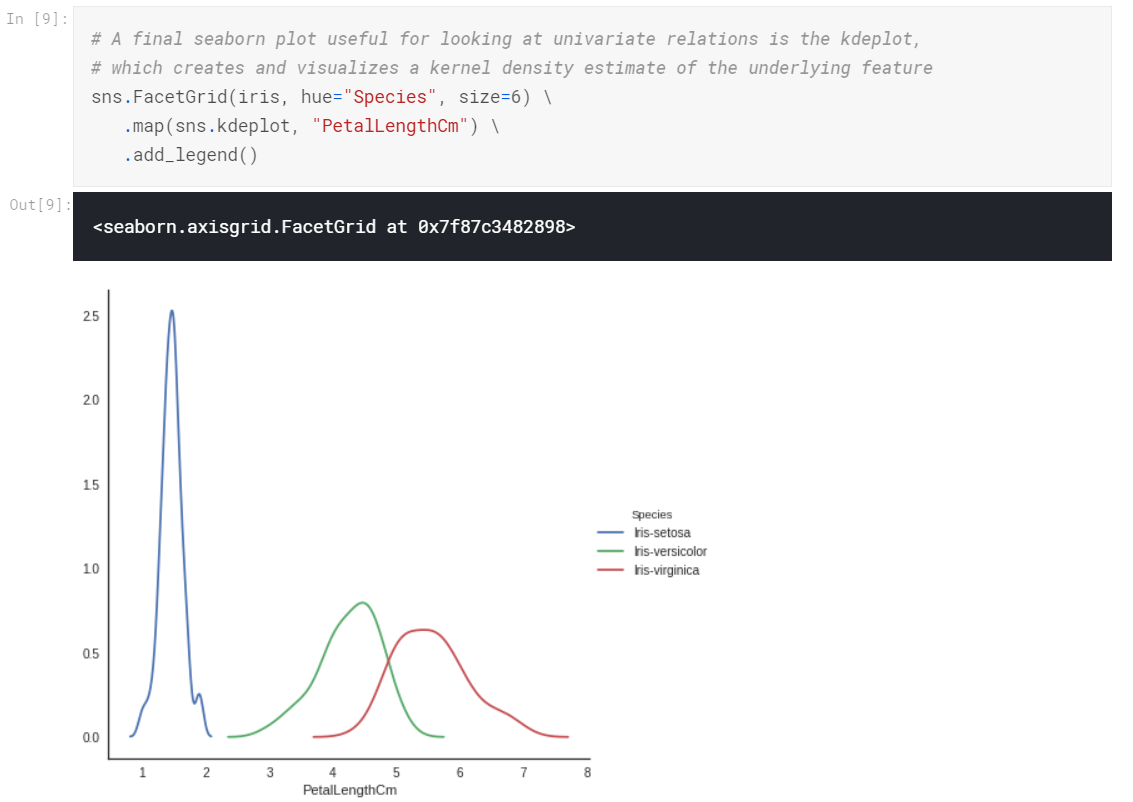






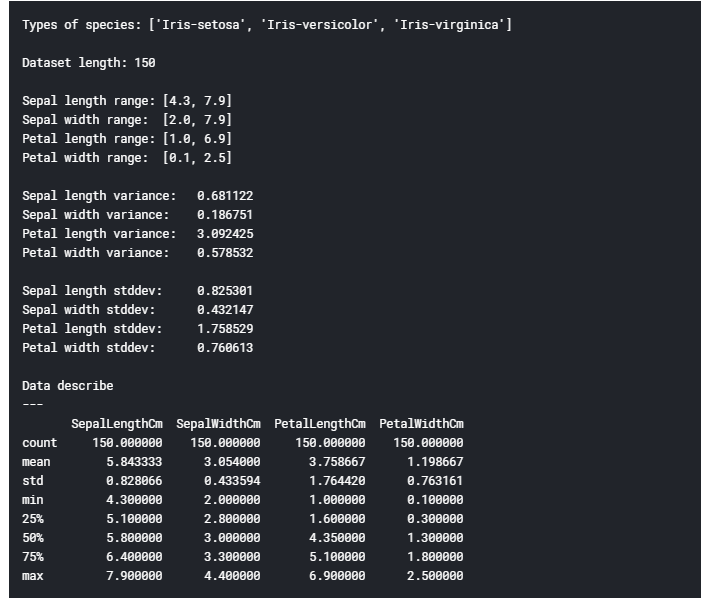




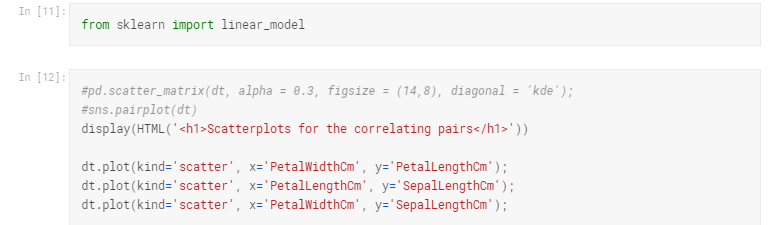


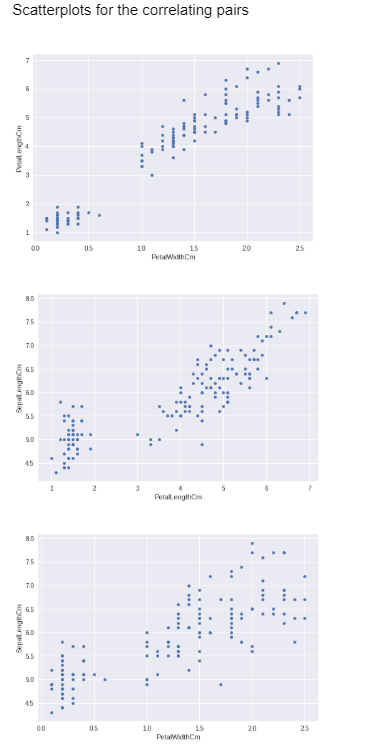
1. Programming Support



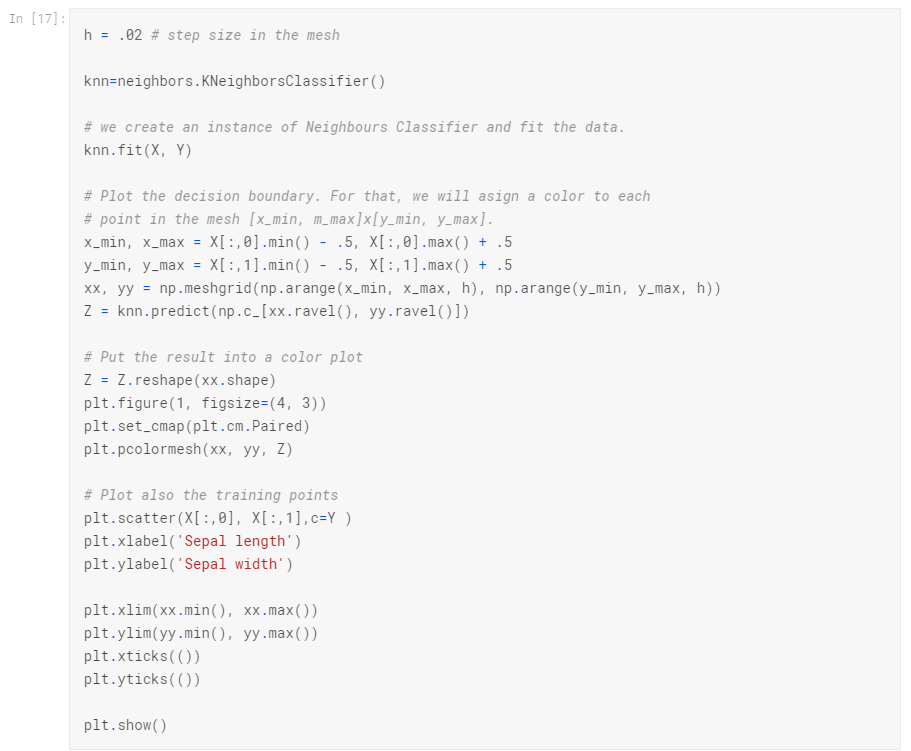


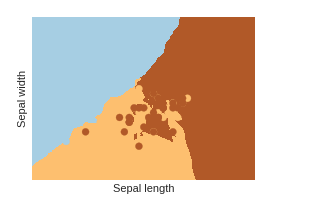
1. Different sequencing Algorithm Support
2. Clustering





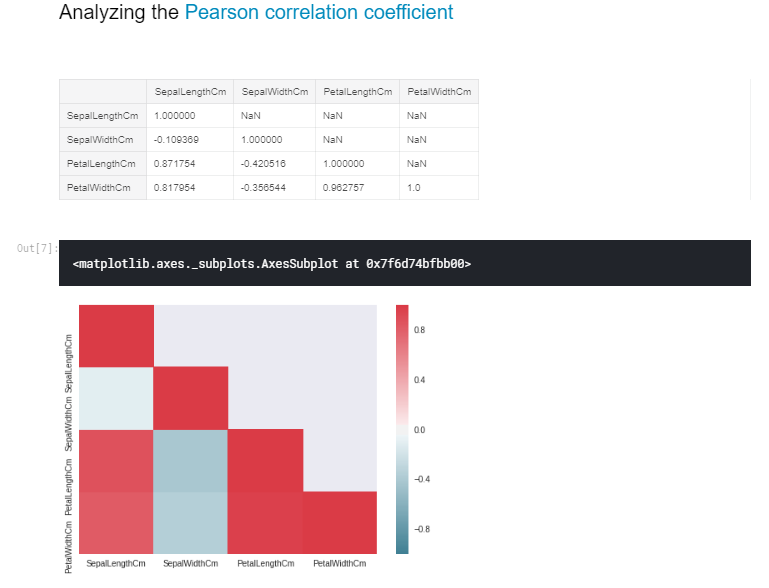
1. KNN Algorithms:





1. Raw Matrix Formation





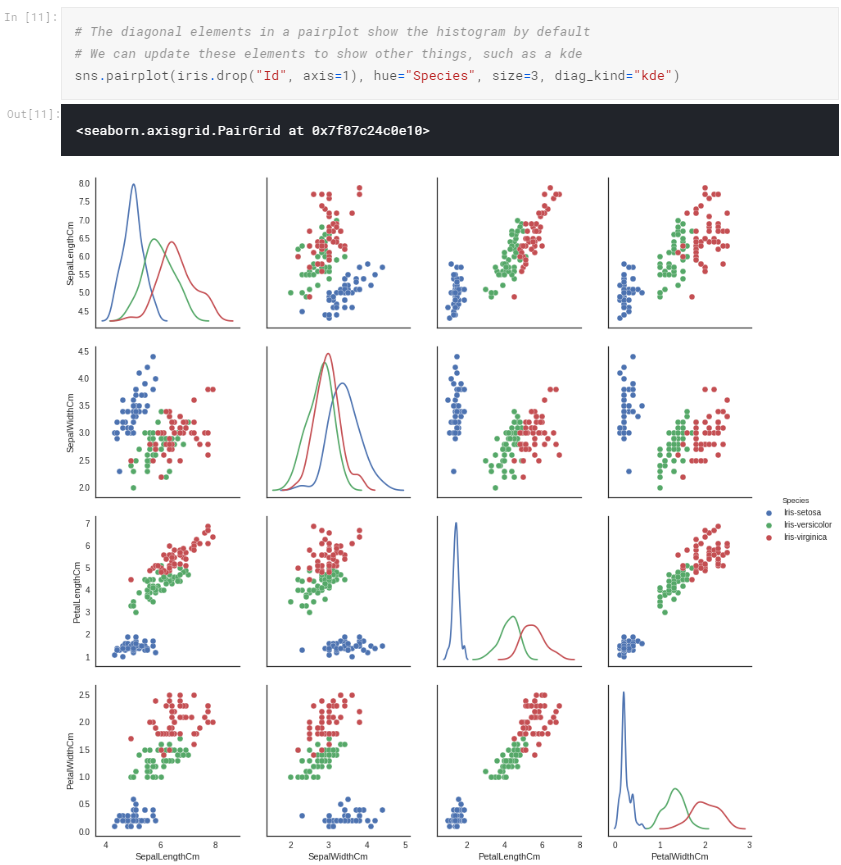
### Interpretation

Diagonal values and upper triangle are ignored (melted the upper triangle through **np.tril** and **df.where**).  
Naturally, we find:

* a **high positive correlation** between PetalWidth and PetalLength (**0.96**)
* a **high positive correlation** between PetalLength and SepalLength (**0.87**)
* a **high positive correlation** between PetalWidth and SepalLength (**0.81**)

As such, we observe correlations between these main attributes: **PetalWidth**, **PetalLength** and **SepalLength**.

1. Matrix Map Display Type





1. Additional Features

Even #D Plots are seen:

