Data Visualization

Topics to be covered

- *1. Distribution.*
- *2. Correlation*
- *3. Ranking*
- *4. Division*
- *5. Evolution*
- *6. Maps*
- *7. Flow*
- *8. Others*

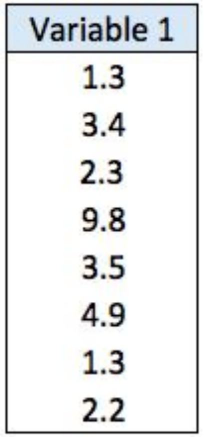
1. DISTRIBUTION

What is this? What information do we want to show?

To represent a density, you need only one vector of numbers. It can be a list, or the column of a data frame.

If you have several numerical variable, you can plot several densities and compare them, or do a boxplot or violin plot.

Distribution is basically the population (frequency) for different elements. The more some specific value (of a variable) occurs, the higher its frequency. If we plot this for all the possible values, we get the visualization of the distribution of that variable.



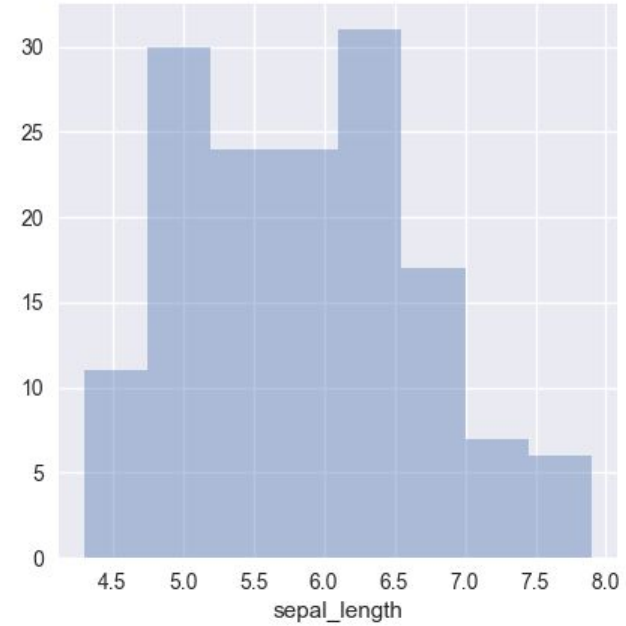
Variable 1
1.3
3.4
2.3
9.8
3.5
4.9
1.3
2.2

1. DISTRIBUTION

Histogram

A histogram is an accurate graphical representation of the distribution of numerical data. It takes as input one numerical variable only.

The variable is cut into several bins, and the number of observation per bin is represented by the height of the bar.

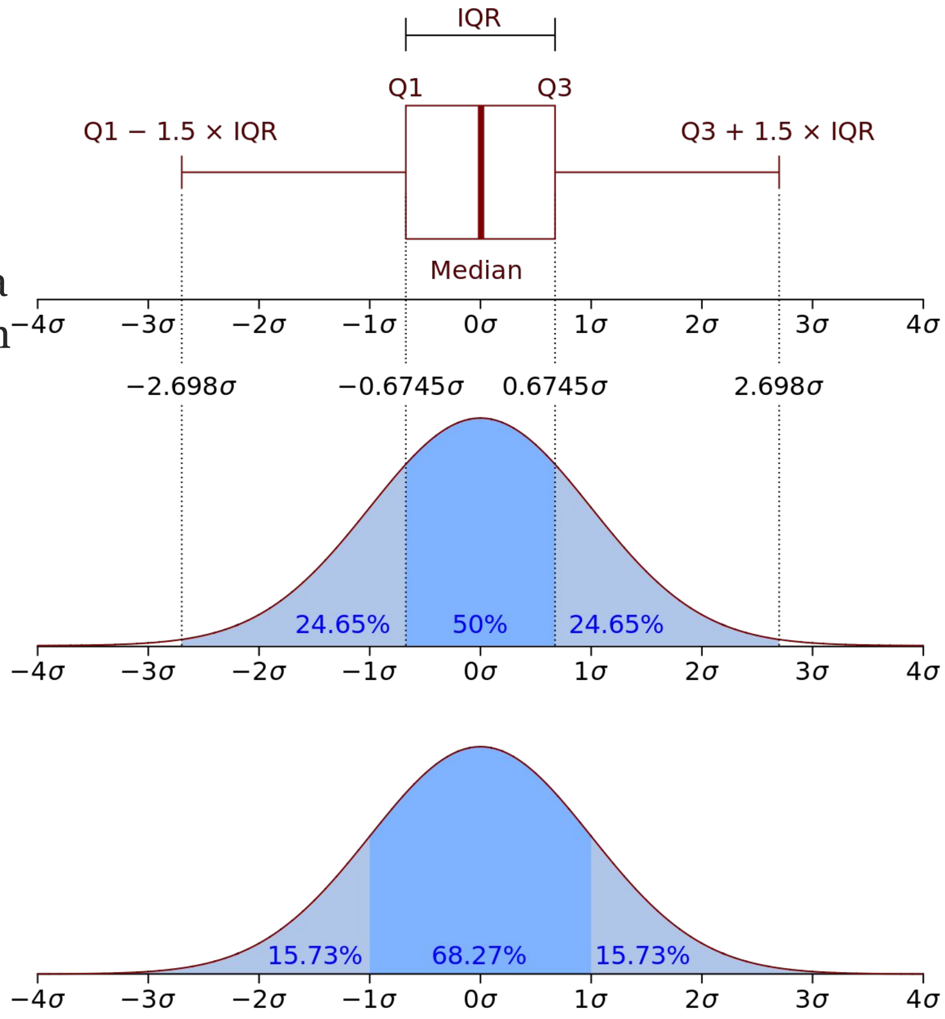


1. DISTRIBUTION

Density

Distribution can be summarized as a measure of how many things fall for different values of the independent variable(s).

In probability theory, a probability density function, or density of a continuous random variable, is a function whose value at any given sample in the sample space can be interpreted as providing a relative likelihood that the value of the random variable would equal that sample.



1. DISTRIBUTION

Box Plot (What you need to know)

Boxplots are a standardized way of displaying the distribution of data based on a five number summary (“minimum”, first quartile (Q1), median, third quartile (Q3), and “maximum”).

median (Q2/50th Percentile): the middle value of the dataset.

first quartile (Q1/25th Percentile): the middle number between the smallest number (not the “minimum”) and the median of the dataset.

third quartile (Q3/75th Percentile): the middle value between the median and the highest value (not the “maximum”) of the dataset.

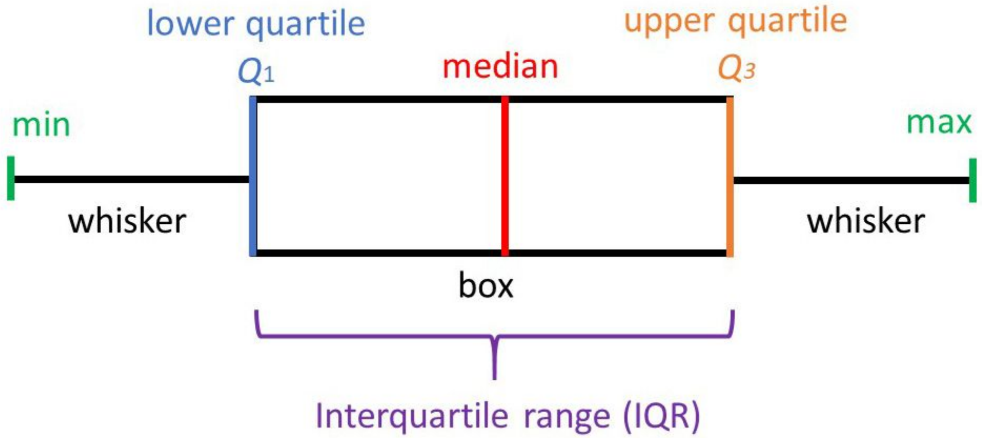
interquartile range (IQR): 25th to the 75th percentile.

whiskers (shown in blue)

outliers (shown as green circles)

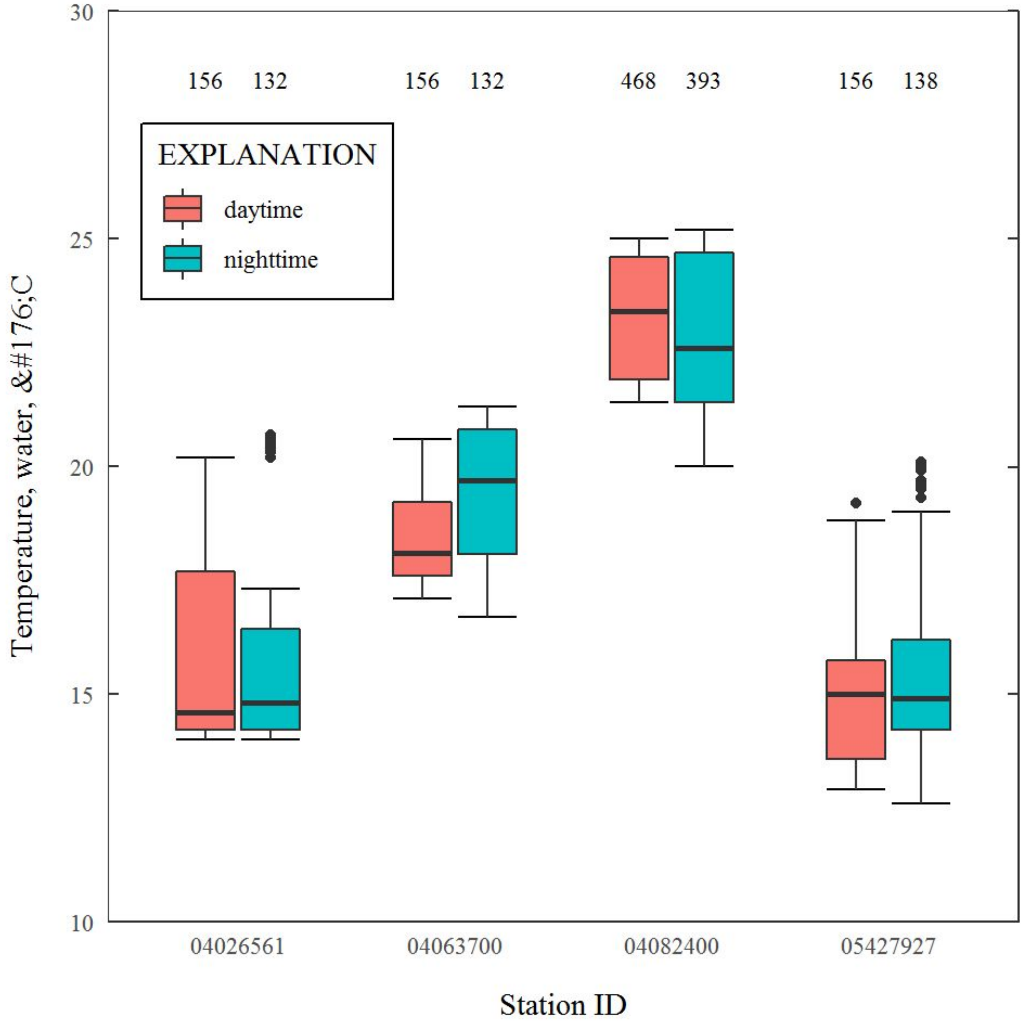
“maximum”: Q3 + 1.5*IQR

“minimum”: Q1 -1.5*IQR



1. DISTRIBUTION

Box Plot (Example)



1. DISTRIBUTION

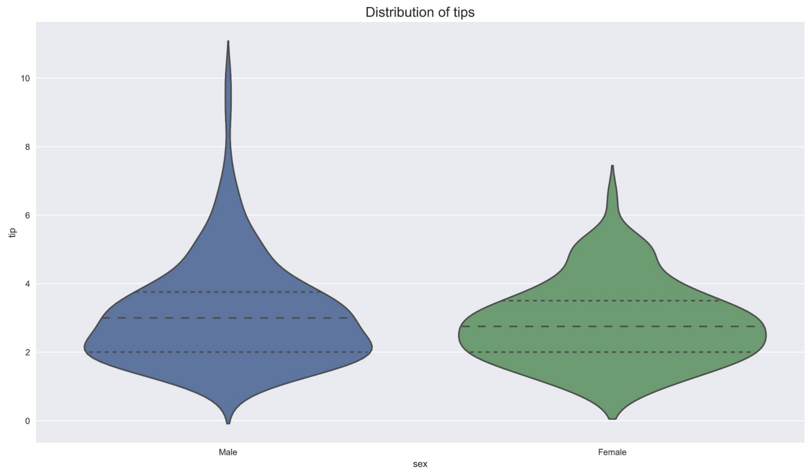
Violin Plot

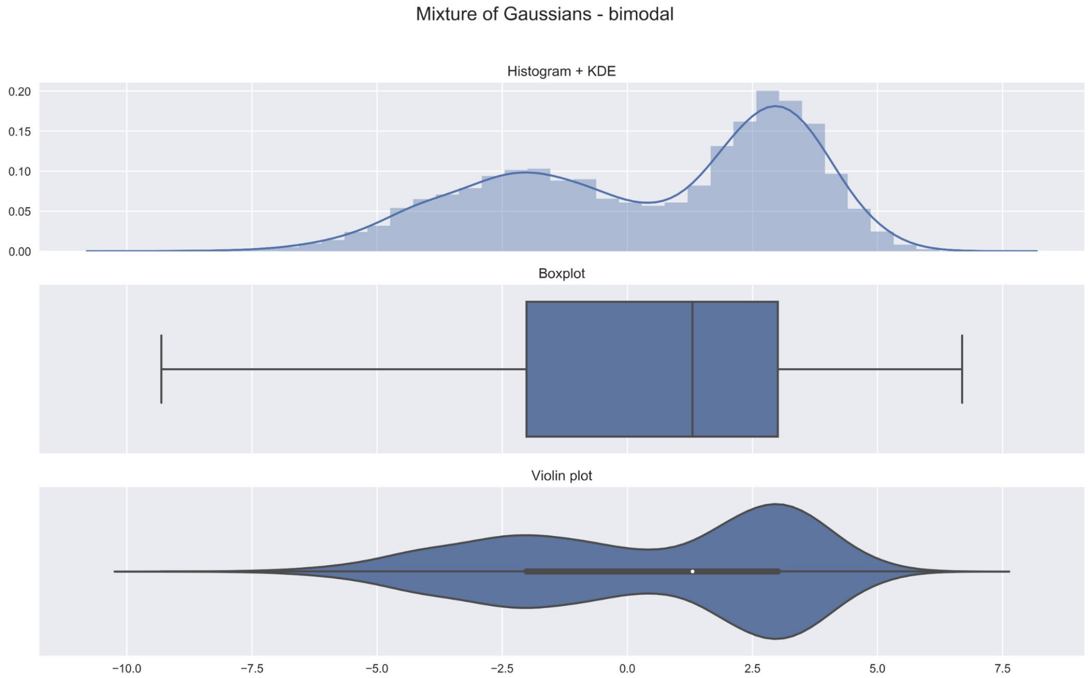
Plots Distribution of data, similar to box plots.

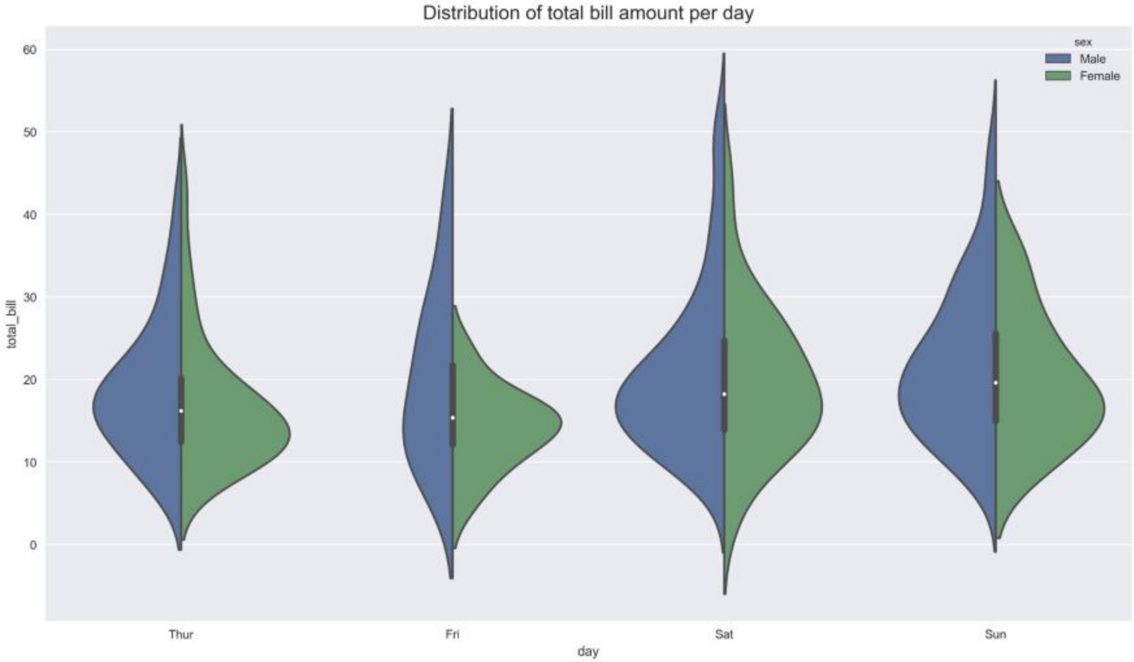
Violin Plot = Box plot + rotated kernel density plot on each side

More informative than a plain box plot

Compare the distribution of a given variable across some categories





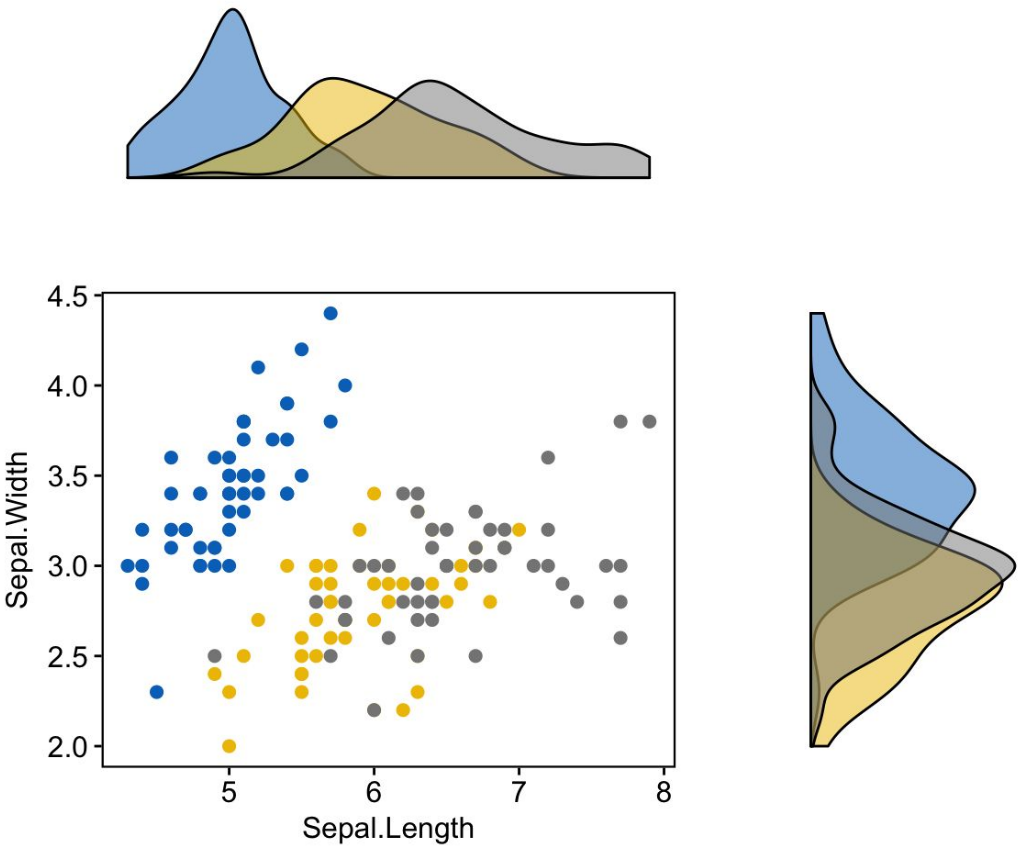


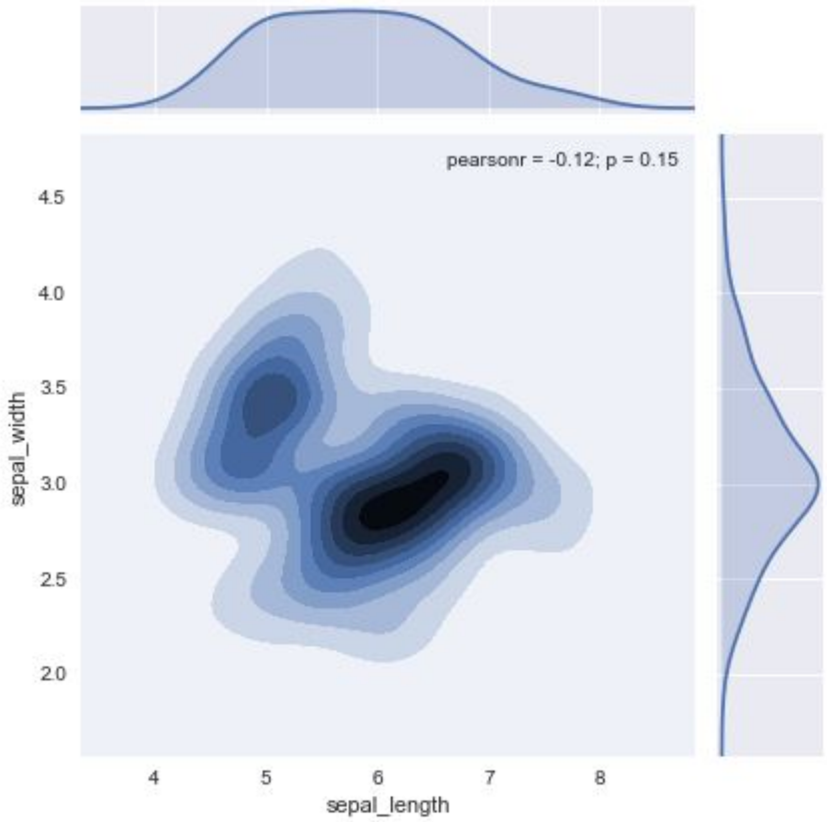
For better understanding: https://towardsdatascience.com/violin-plots-explained-fb1d115e023d

1. DISTRIBUTION

Marginal plot

When there are more than one variables, you can plot the joint effect of the distribution and the separate ones all in a single curve through a marginal plot.





2. Correlation

What is this?

There are always multiple variables in our distribution. We need to come up with good visualization to easily understand the relationships between them and also make in understandable for the general masses.

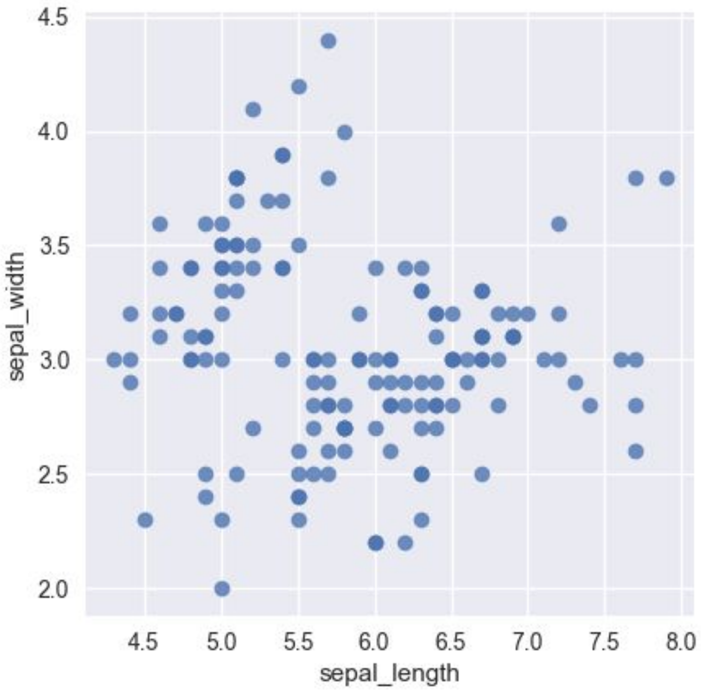
The ones shown here help us do that.

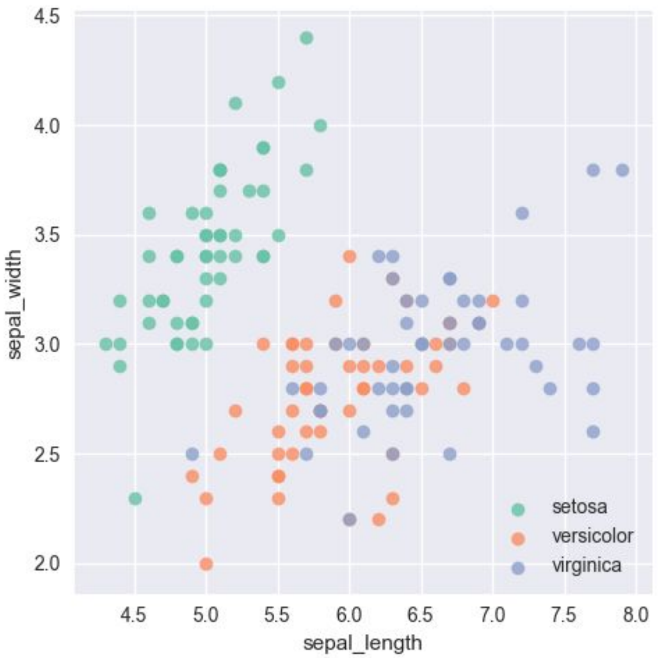
2. Correlation

Scatter Plot

Plotting the *coordinates* of different samples keeping feature vectors as axis.

We can do ALL sorts of things to make the vizualization more informative.





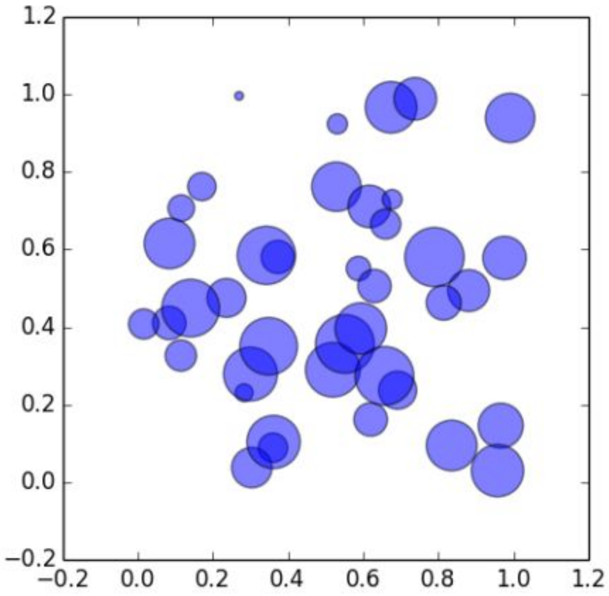
2. Correlation

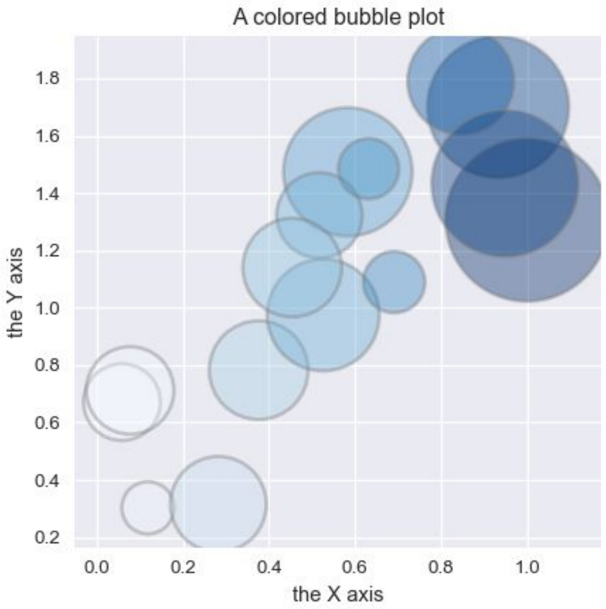
Bubble Plot

Scatter plot with a third dimension.

Where every co-ordinate on 2D feature space also has another value (another feature) that is shown by putting a bubble in that specific coordinate. The bigger the value is, the bigger the bubble.

A bubble plot is a scatterplot with a third dimension: the size of the markers. It is even possible to add a *fourth* dimension using colors.

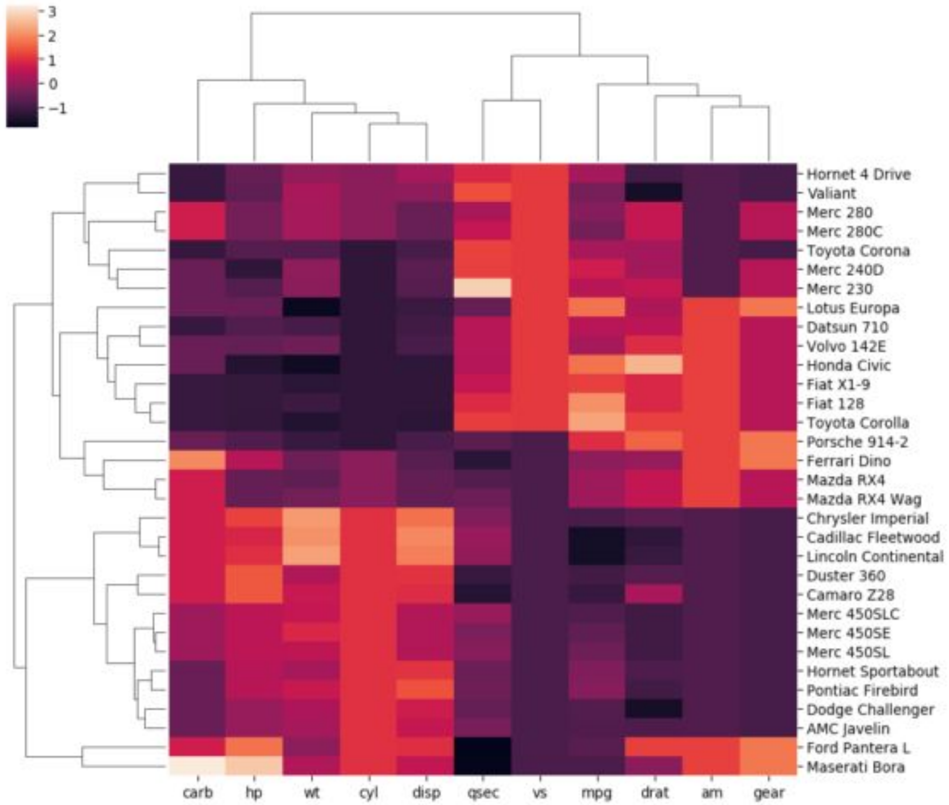


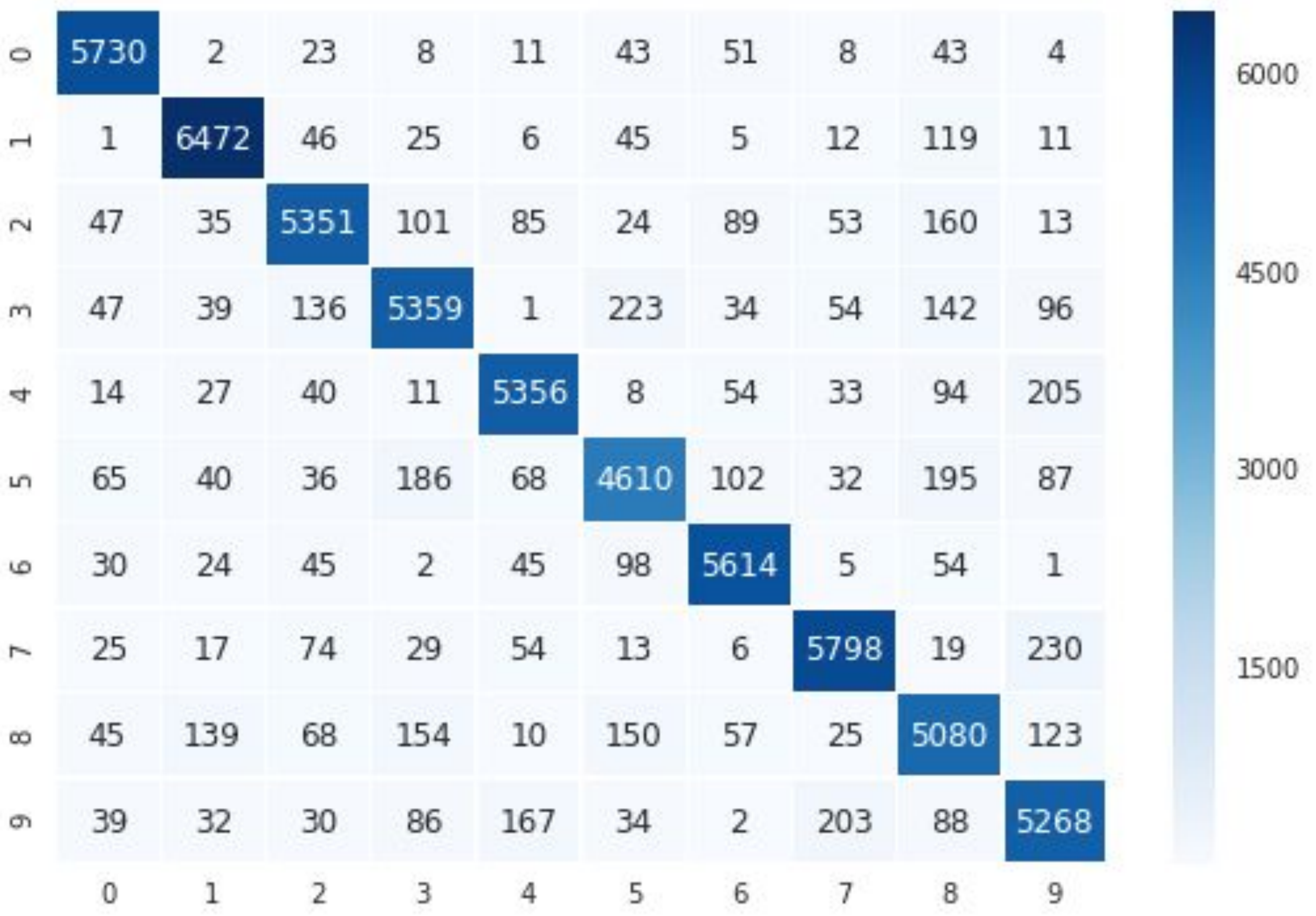


2. Correlation

Heatmap

Heatmap is an essential tool for visualizing the performance of classifiers.
We can use this to visualize the confusion of models after we have trained them. You can always use this for other tasks, too. Also, you can modify heatmaps for better visualization.



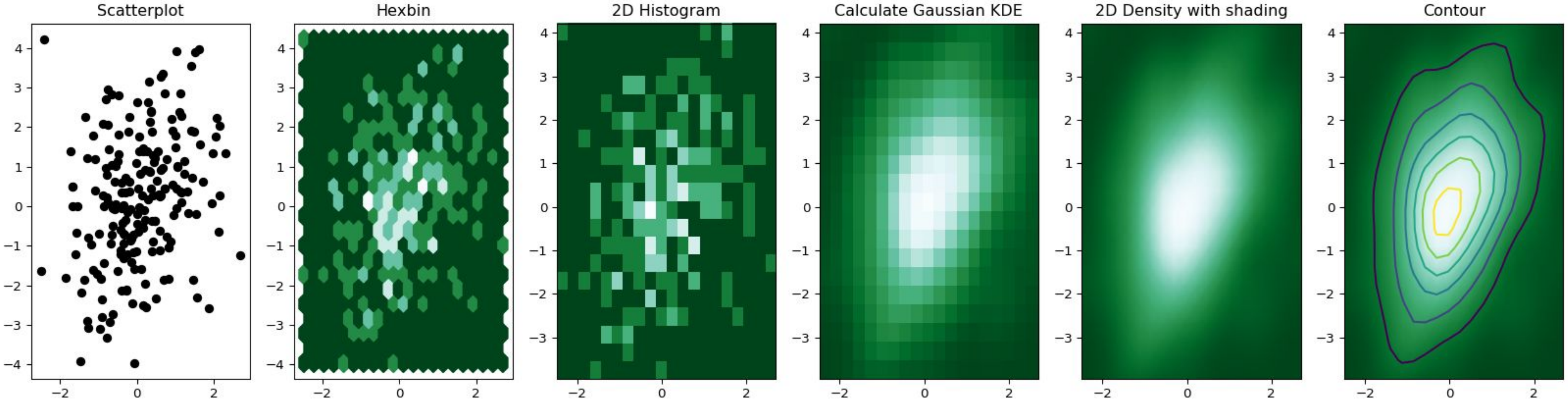


	0	1	2	3	4	5	6	7	8	9
0	5730	2	23	8	11	43	51	8	43	4
1	1	6472	46	25	6	45	5	12	119	11
2	47	35	5351	101	85	24	89	53	160	13
3	47	39	136	5359	1	223	34	54	142	96
4	14	27	40	11	5356	8	54	33	94	205
5	65	40	36	186	68	4610	102	32	195	87
6	30	24	45	2	45	98	5614	5	54	1
7	25	17	74	29	54	13	6	5798	19	230
8	45	139	68	154	10	150	57	25	5080	123
9	39	32	30	86	167	34	2	203	88	5268

2. Correlation

2D Histogram/ 2D Density Plot

A 2D density plot or 2D histogram is an extension of the well known histogram. It shows the distribution of values in a data set across the range of two quantitative variables.

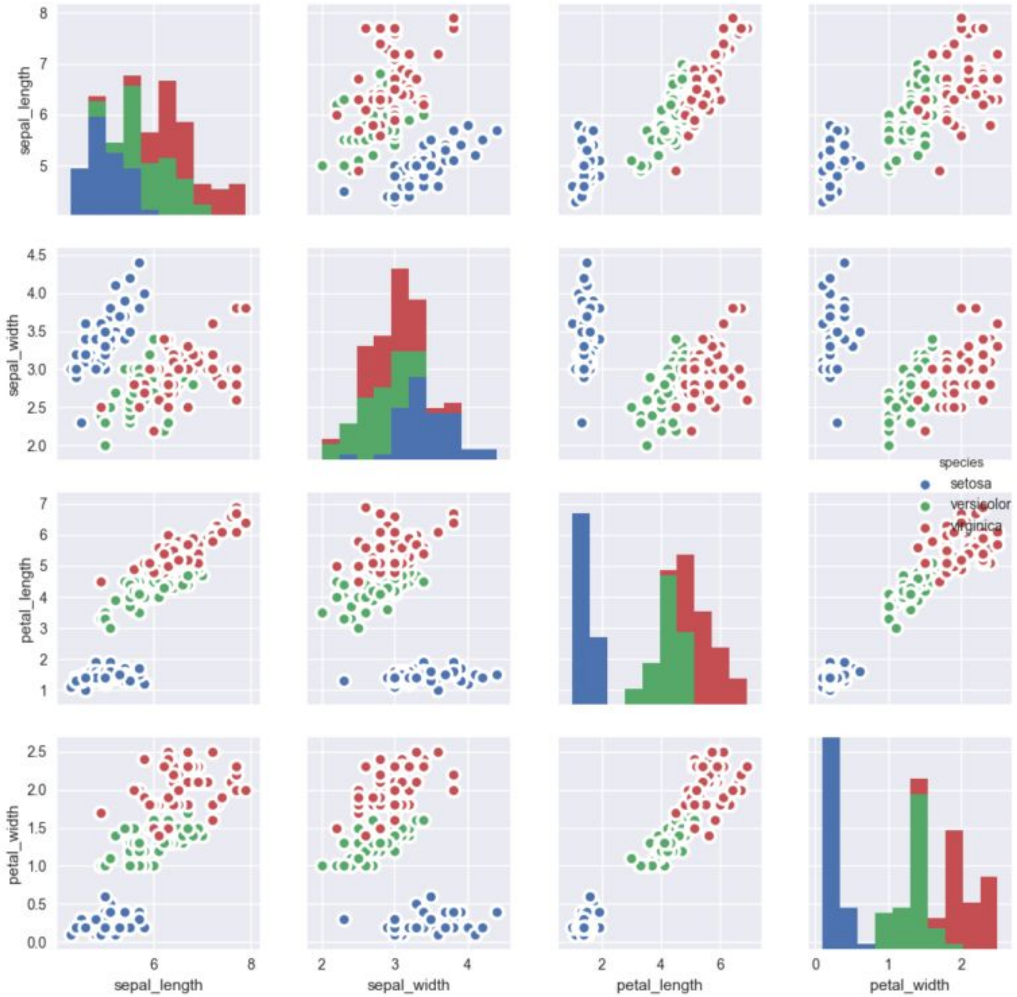


2. Correlation

Corrgram

In the paper paper, the authors described it as a set of techniques subsumed under the name “corrgram”, based on
two main schemes: (a) rendering the value of a correlation to depict its sign and magnitude. We consider some of the properties of several iconic representations, in relation to the kind of task to be performed. (b) re-ordering the variables in a correlation matrix so that “similar” variables are positioned adjacently, facilitating perception.

There can be many versions of corrgrams. Here, one with modified scatter plots is shown for Iris flower classification dataset.



3. RANKING

What is this?

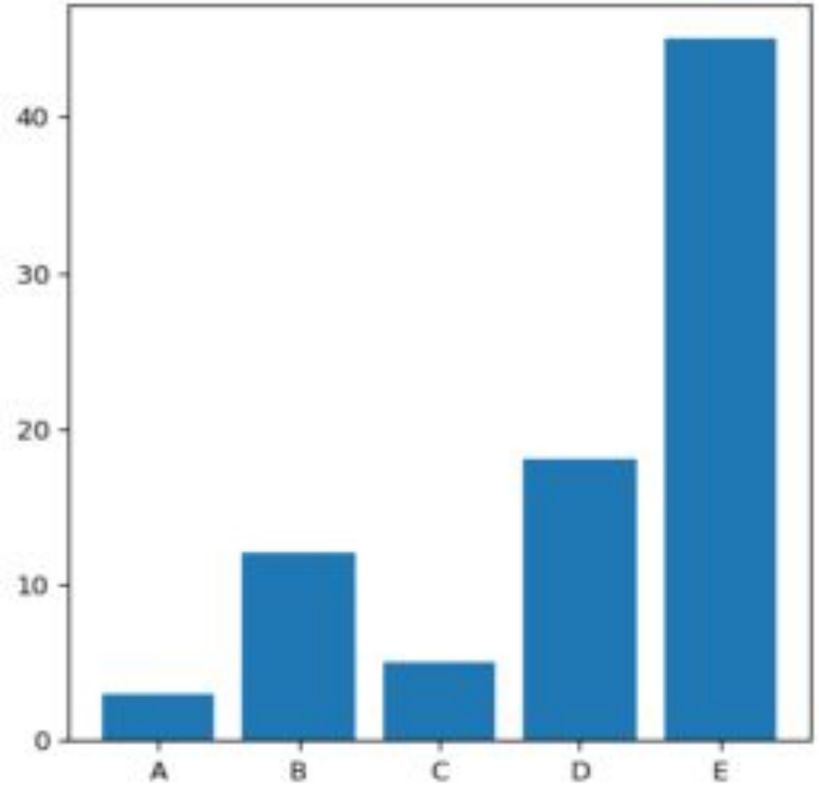
Techniques to give a sense of ranking among various data points.

3. RANKING

Bar Plot

A barplot (or barchart) is one of the most common type of plot. It shows the relationship between a numerical variable and a categorical variable. For example, you can display the height of several individuals using bar chart.

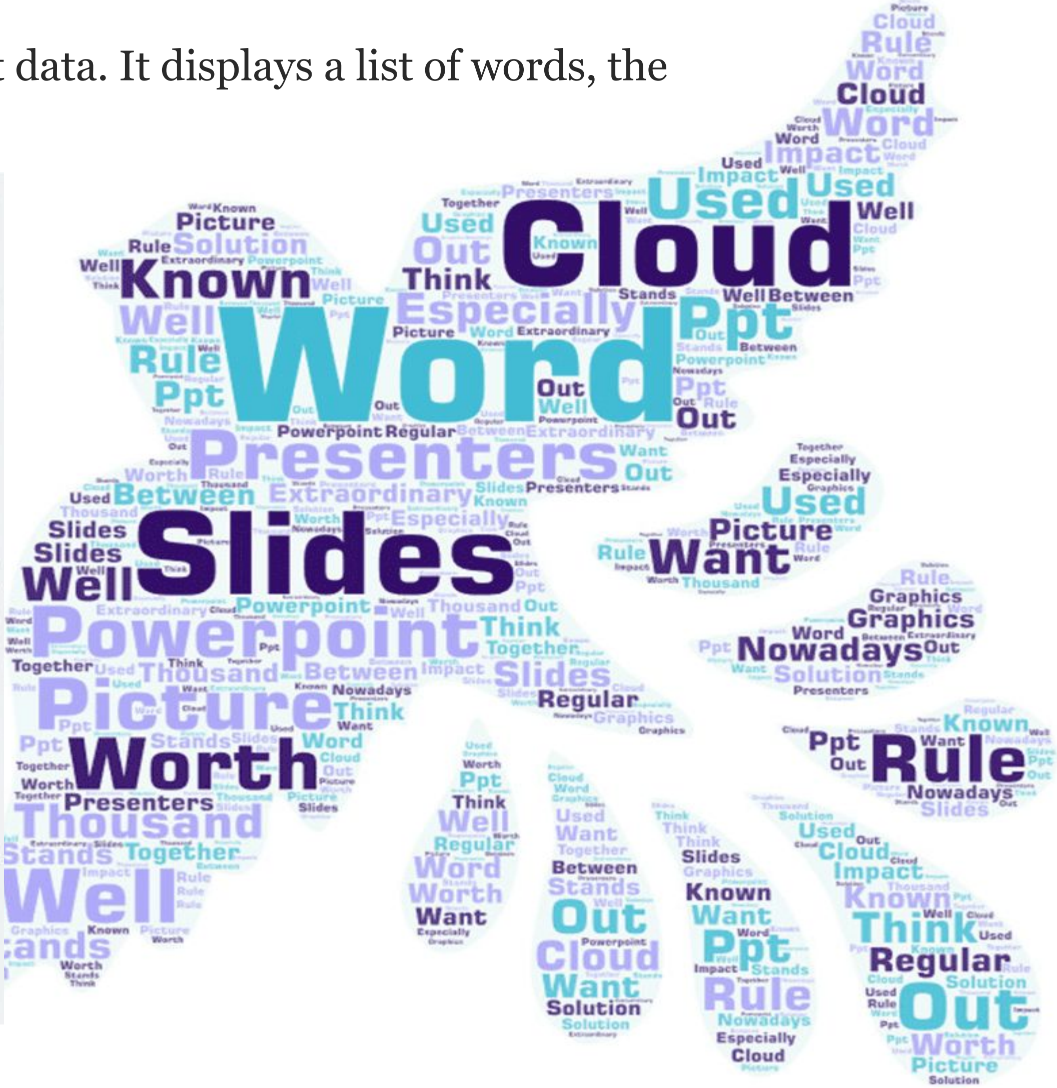
Barcharts are often confounded with histograms, which is highly different. (It has only a numerical variable as input and shows its distribution).



3. RANKING

Wordcloud

A Wordcloud (or Tag cloud) is a visual representation of text data. It displays a list of words, the importance of each beeing shown with font size or color.

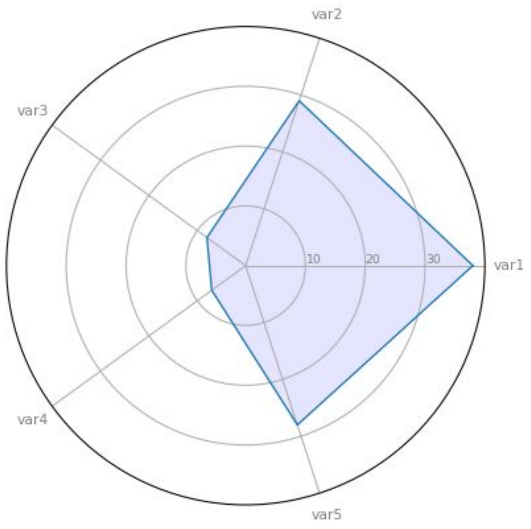


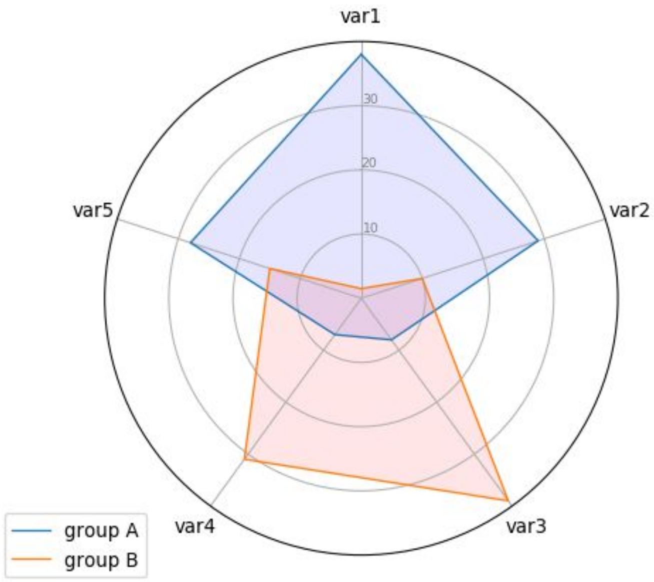
3. RANKING

Spider Plot

A Radar chart or Spider plot or Polar chart or Web chart allows to study the feature of one or several individuals for several numerical variables.

It is possible to represent several individuals on the same graph but be careful, the chart can quickly become unreadable. Instead, try to use faceting: display as many chart as the number of individual, it makes easy to compare the shape of each.



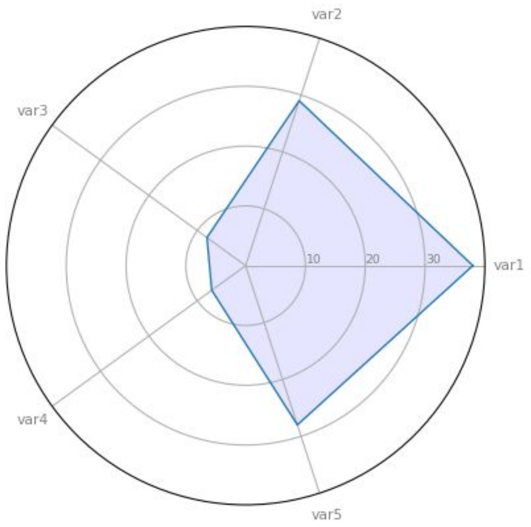


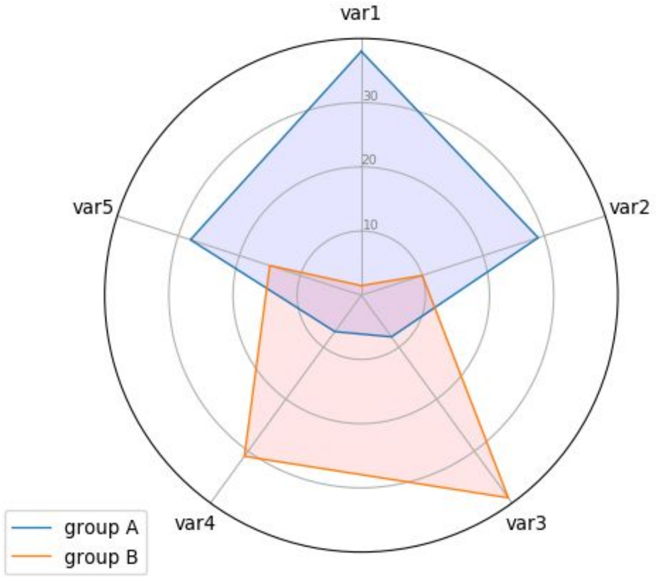
3. RANKING

Spider Plot

A Radar chart or Spider plot or Polar chart or Web chart allows to study the feature of one or several individuals for several numerical variables.

It is possible to represent several individuals on the same graph but be careful, the chart can quickly become unreadable. Instead, try to use faceting: display as many chart as the number of individual, it makes easy to compare the shape of each.



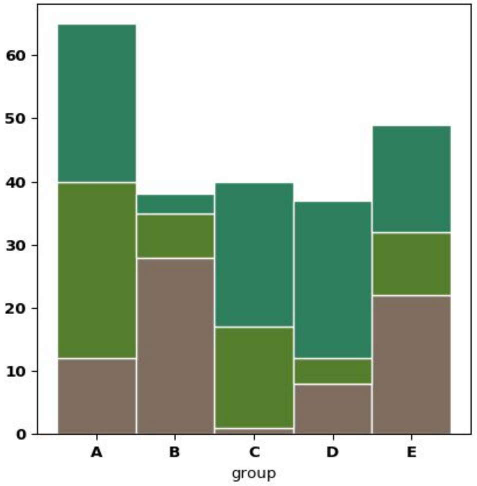


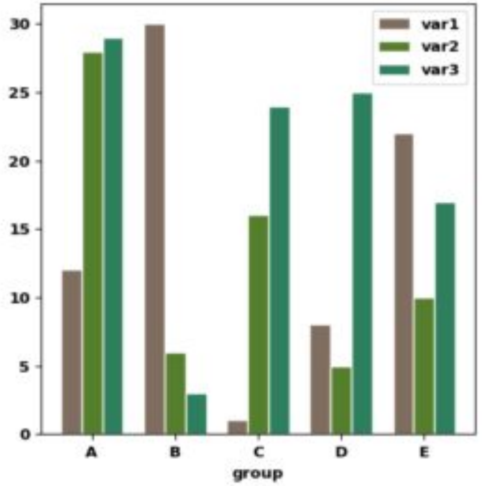
4. Division

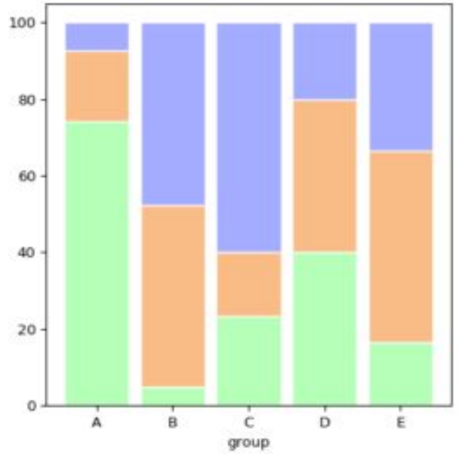
Stacked Bar Plot

There are three similar things. Grouped barcharts, stacked barcharts and percent stacked barcharts. This 3 types of barplot variation have the same objective. It displays a numerical value for several entities, organised into groups and subgroups.

A grouped barplot display the subgroups one beside each other, whereas the stacked ones display them on top of each other. The percent variation normalise the data to make in sort the value of each group is 100. It allows the compare the importance of each subgroups in each group more effectively



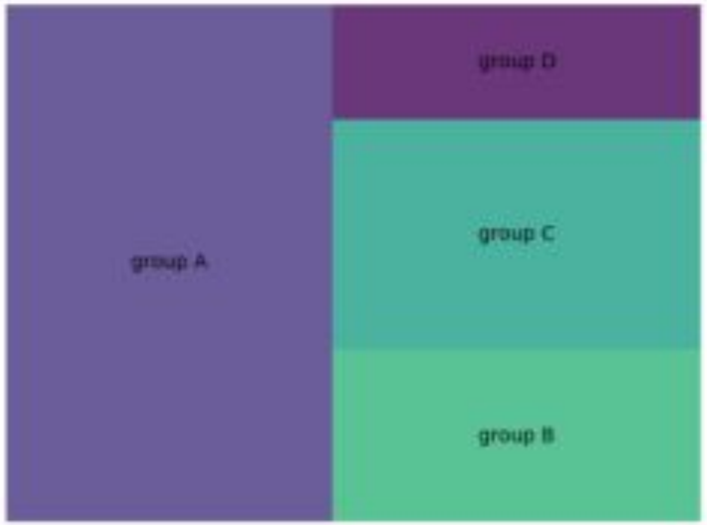




4. Division

Treemaps

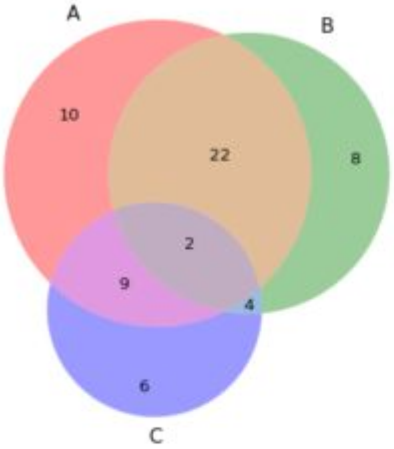
Treemaps display hierarchical data as a set of nested rectangles. Each group is represented by a rectangle, which area is proportional to its value. Using color schemes, it is possible to represent several dimensions: groups, subgroups...
Treemaps have the advantage to make efficient use of space, what makes them useful to represent a big amount of data.

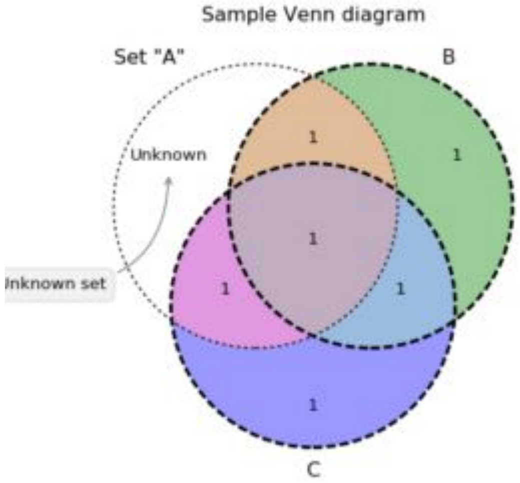


4. Division

Venn Diagram

A Venn diagram (also called primary diagram, set diagram or logic diagram) is a diagram that shows all possible logical relations between a finite collection of different sets.



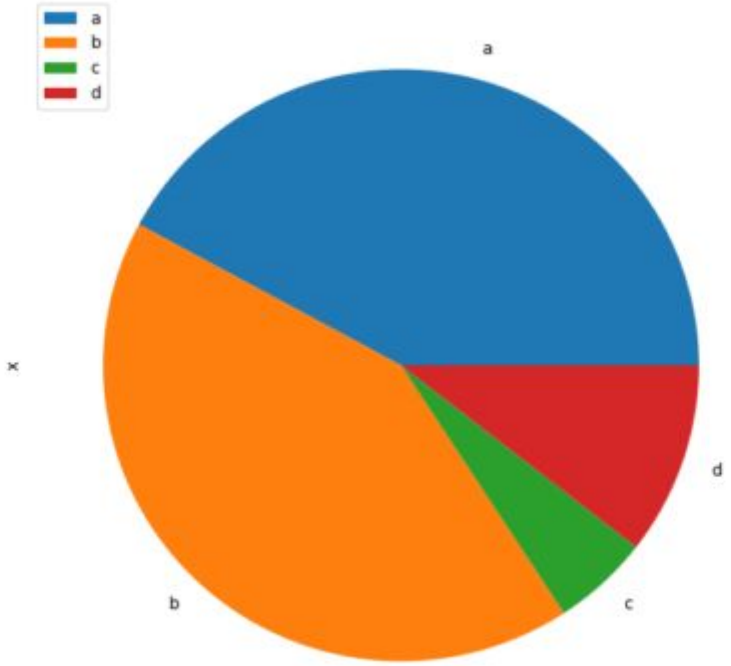


4. Division

Pie Chart

No need to explain ig.

Please do not use pie charts for anything. :3
Use bar charts instead. Or the next one.

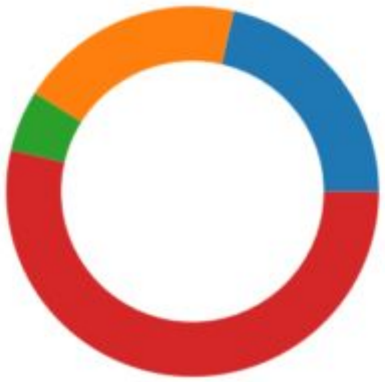


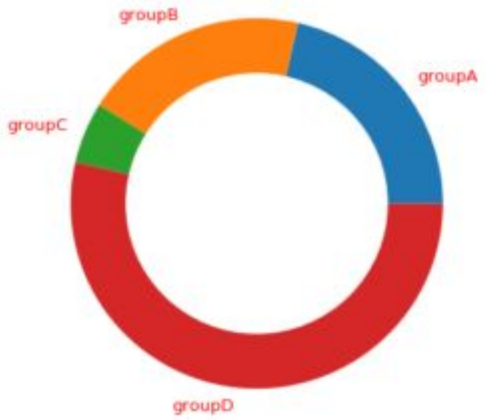
4. Division

Donut chart

A donut chart is essentially a Pie Chart with an area of the center cut out.

Looks nice





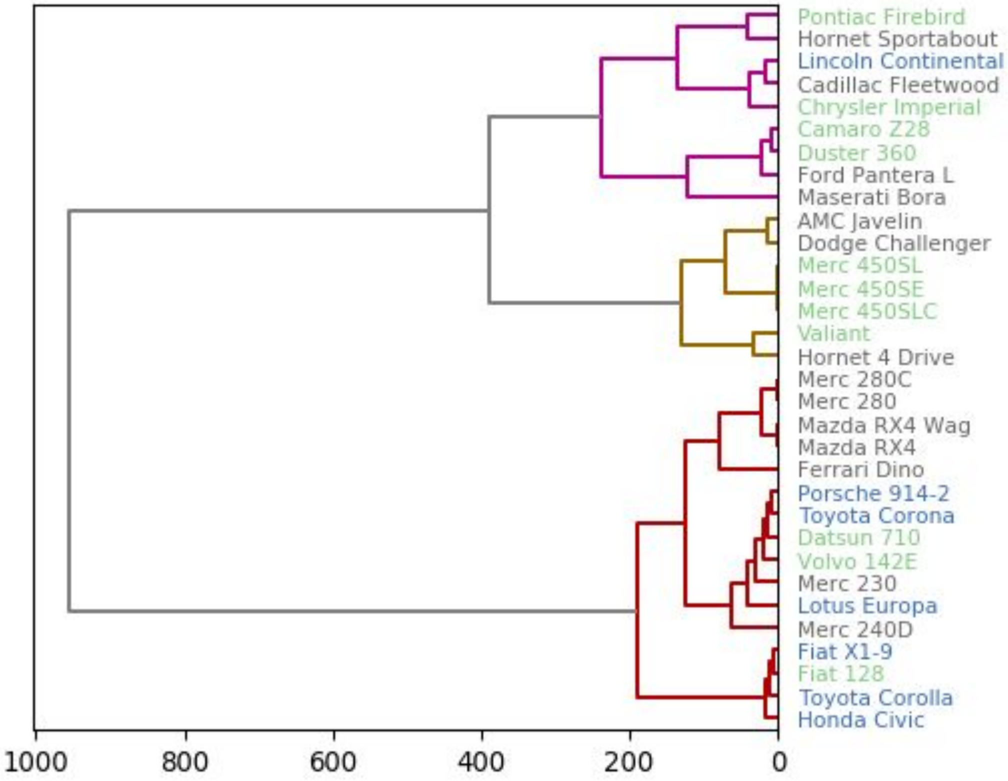
4. Division

Tree chart

A dendrogram or tree diagram allows to illustrate the hierarchical organisation of several entities. For example, we often use it to make family trees. It is constituted of a root node, which give birth to several nodes that ends by giving leaf nodes (the bottom of the tree).
Dendrogram can be made with 2 types of dataset.

i/ a numeric matrix where several variables describe the features of individuals. We can then calculate the distance between individuals and clustering them.

ii/ A hierarchical dataset where the relationship between entities is provided directly. Note that for clusterization, it is a good practice to provide the corresponding heat map that illustrates the structure.



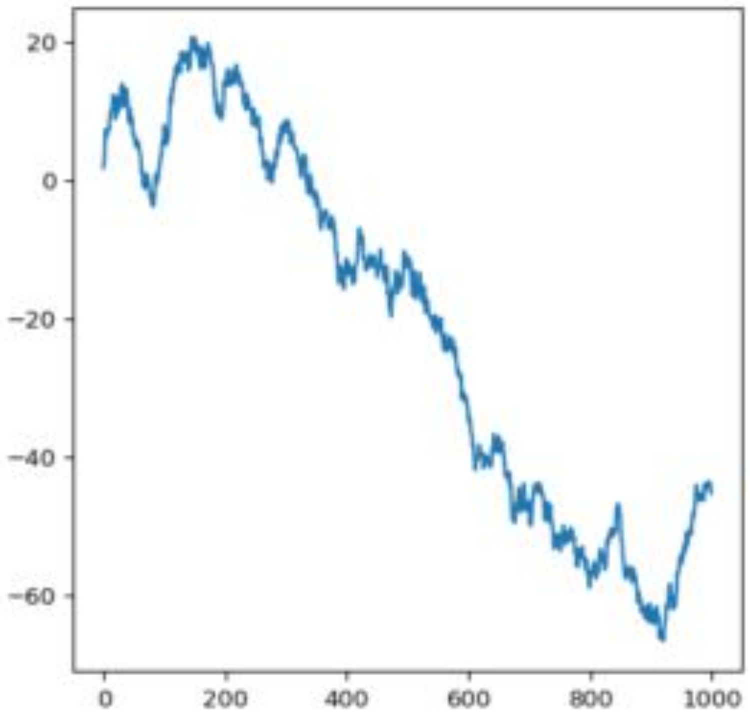
5. Evolution

Line Chart

A line chart or line graph is a type of chart which displays information as a series of data points called ‘markers’ connected by straight line segments. It is a basic type of chart common in many fields.

 It is similar to a scatter plot except that the measurement points are ordered (typically by their x-axis value) and joined with straight line segments.

A line chart is often used to visualize a trend in data over intervals of time – a time series – thus the line is often drawn chronologically. In these cases they are known as run charts.



7. Flow

Chord Diagram

Chord diagrams allow to visualize flows between several entities. Each entity is represented by a fragment on the outside of the circle. Then, arcs are drawn between each entities. The size of the arc is proportional to the importance of the flow

8. General Instruction

What should we do now?

Keep an open mind, look for new interesting ways of plotting
Try to understand the context of each plot. Why is this used to plot this specific kind of data?
GOOGLE
GOOGLE Even more. Everythings written somewhere.
You just need to pick things up efficiently. That's all.

A good place to start from: https://python-graph-gallery.com/