

# Basic Machine Learning: Data Visualization

Goal

# Goal

To understand how important data visualization is in term of building machine learning model

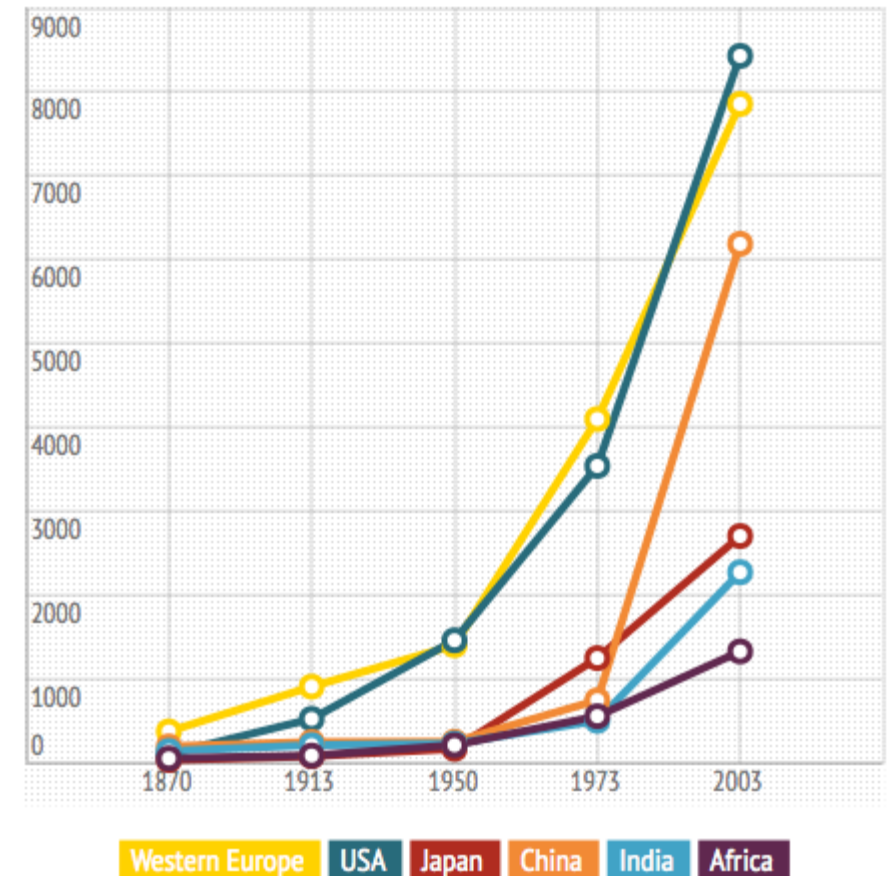
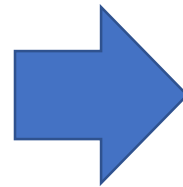
- Help answer questions
- Generate hypothesis for further analysis
- Prepare data for machine learning modeling process

# Why Data Visualization?

## Past Country's GDP Data

	A	B	C	D	E	F
1	Past GDP	1870	1913	1950	1973	2003
2	Western Europe	367	902	1396	4096	7857
3	USA	98	517	1455	3536	8430
4	Japan	25	71	160	1242	2699
5	China	189	241	244	739	6187
6	India	134	204	222	494	2267
7	Africa	45	79	203	549	1322

Data Table



Data Visualization

# Outline

# Outline

- Matplotlib
- Visual Categories
- Single, double and multiple variables

Content

# About Matplotlib

Matplotlib is a comprehensive library for creating static, animated, and interactive visualizations in Python. Matplotlib makes easy things easy and hard things possible.

“A picture is worth a thousand words”





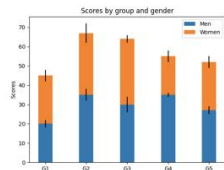
# About Matplotlib

## Gallery

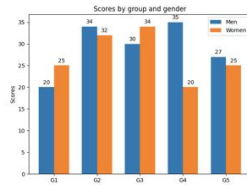
This gallery contains examples of the many things you can do with Matplotlib. Click on any image to see the full image and source code.

For longer tutorials, see our [tutorials page](#). You can also find [external resources](#) and a [FAQ](#) in our [user guide](#).

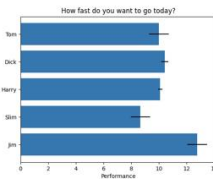
## Lines, bars and markers



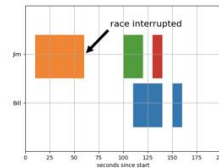
Stacked bar chart



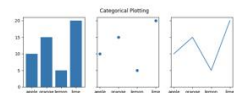
Grouped bar chart  
with labels



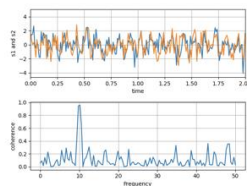
Horizontal bar chart



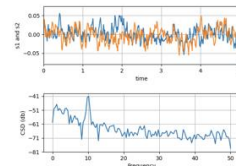
Broken Barh



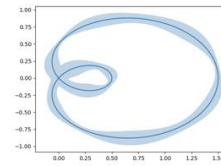
Plotting categorical  
variables



Plotting the  
coherence of two  
signals



CSD Demo



Curve with error  
band

[Matplotlib Galery](#)

# About Matplotlib

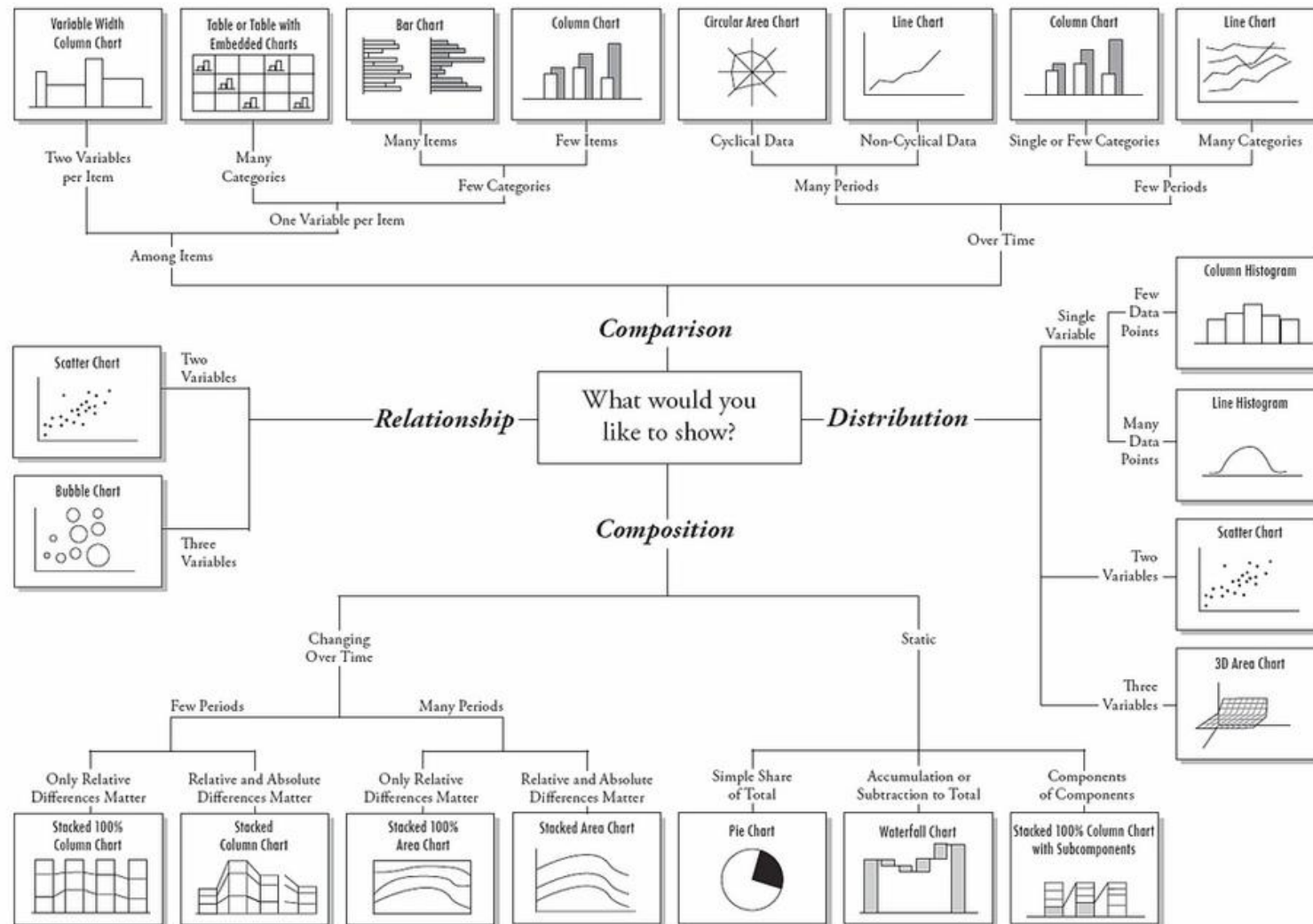
We can divide the chart messages in four categories namely:

1. Distribution
2. Comparison
3. Relationship
4. Composition



# About Matplotlib

## Chart Suggestions—A Thought-Starter



# About Matplotlib

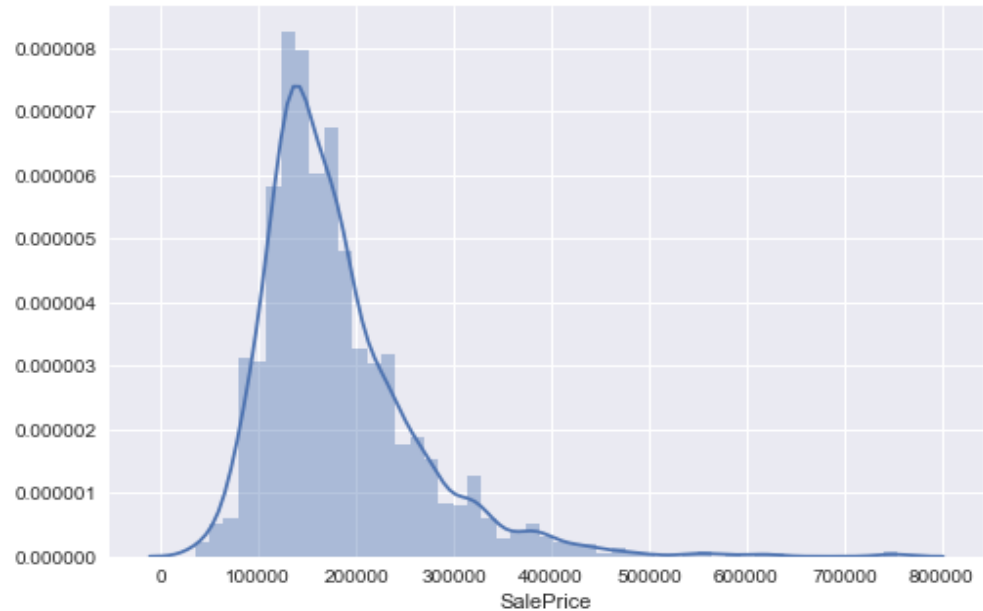
We can also divide the chart messages in three categories namely

1. Single Variable Visualization
2. Two Variable Visualization
3. Multi Variable Visualization



# Single Variable – Displot

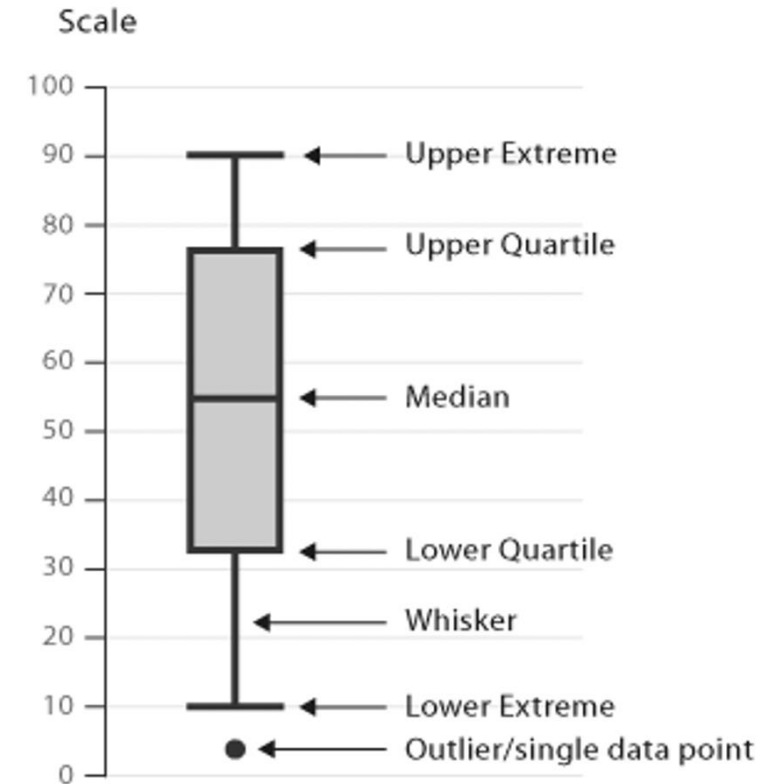
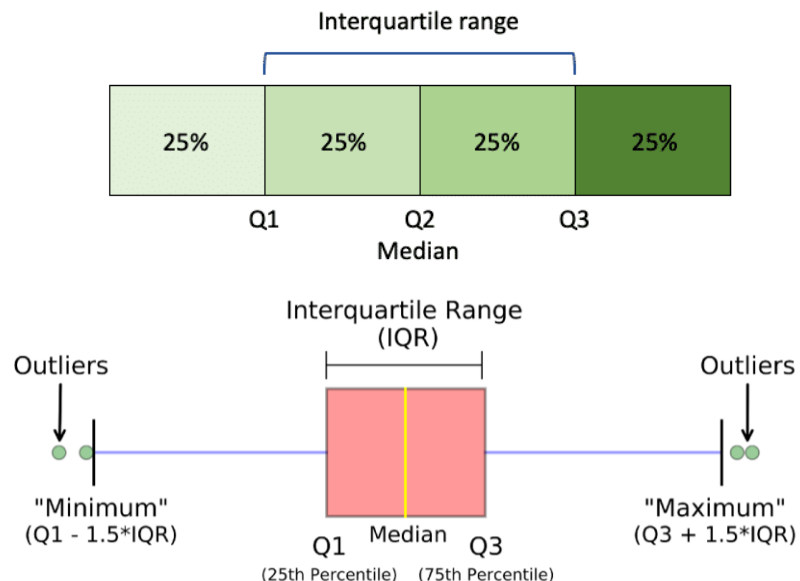
- Summarize the distribution of a single variable dataset
- Show centre, spread, skewness, outliers, multiple modes



# Single Variable – Box Plot

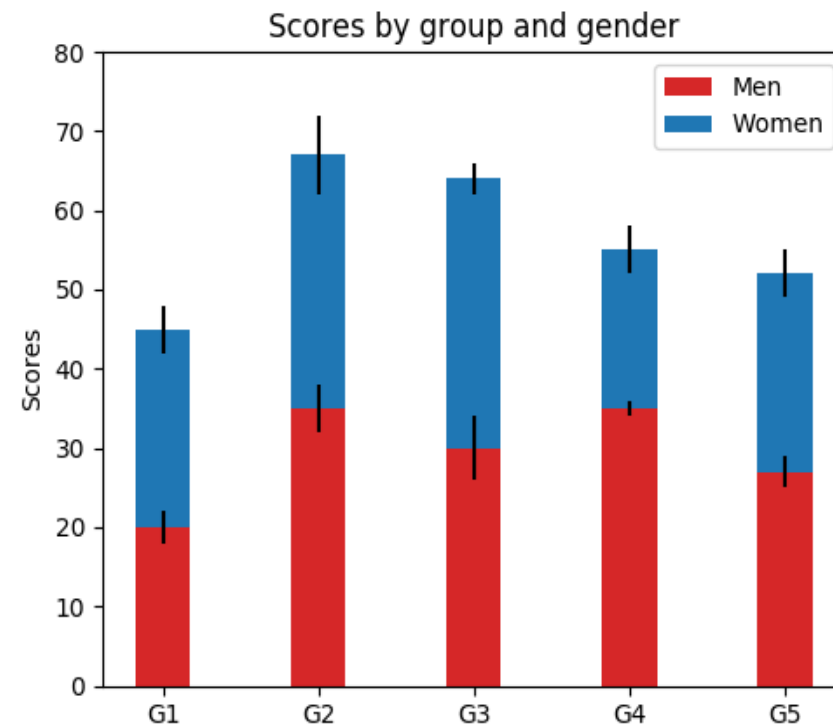
- Depicting groups of numerical data through their quartiles
- It shows a lot of information about a variable in one plot

1. Median
2. Inter Quartile Range (IQR)
3. Outliers
4. Range



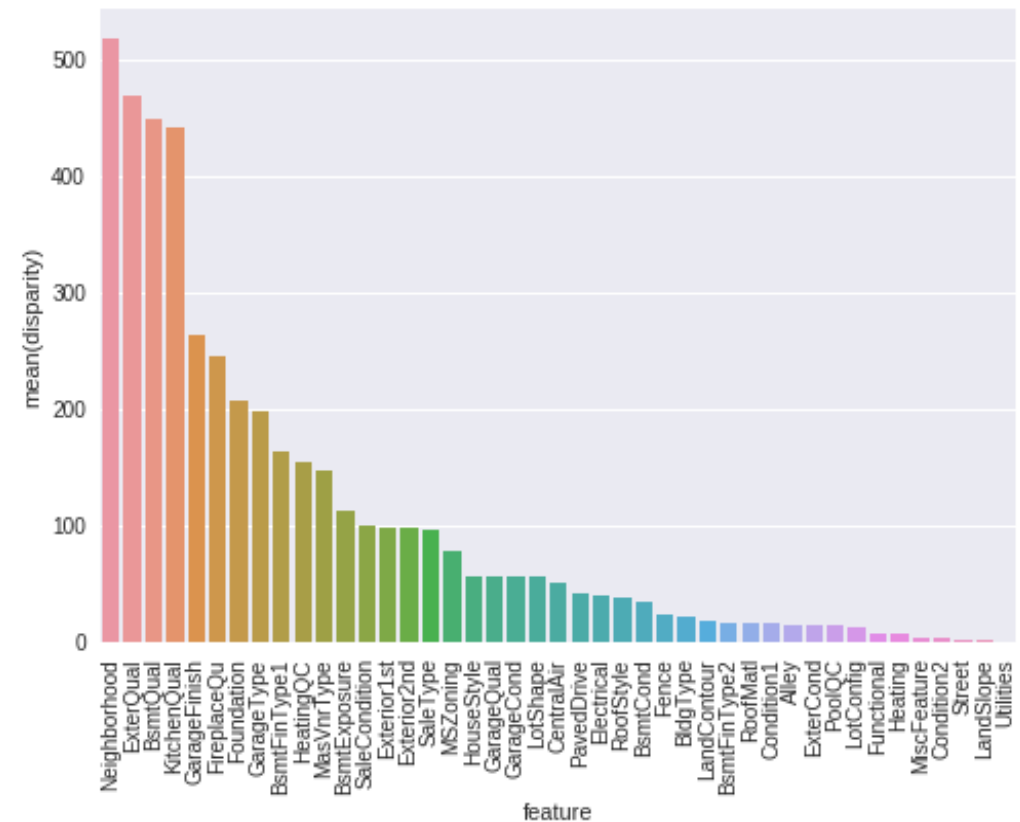
# Single Variable – Stacked Bar

- Show comparisons between categories of data, but with ability to break down and compare parts of a whole



# Single Variable – Bar Plot

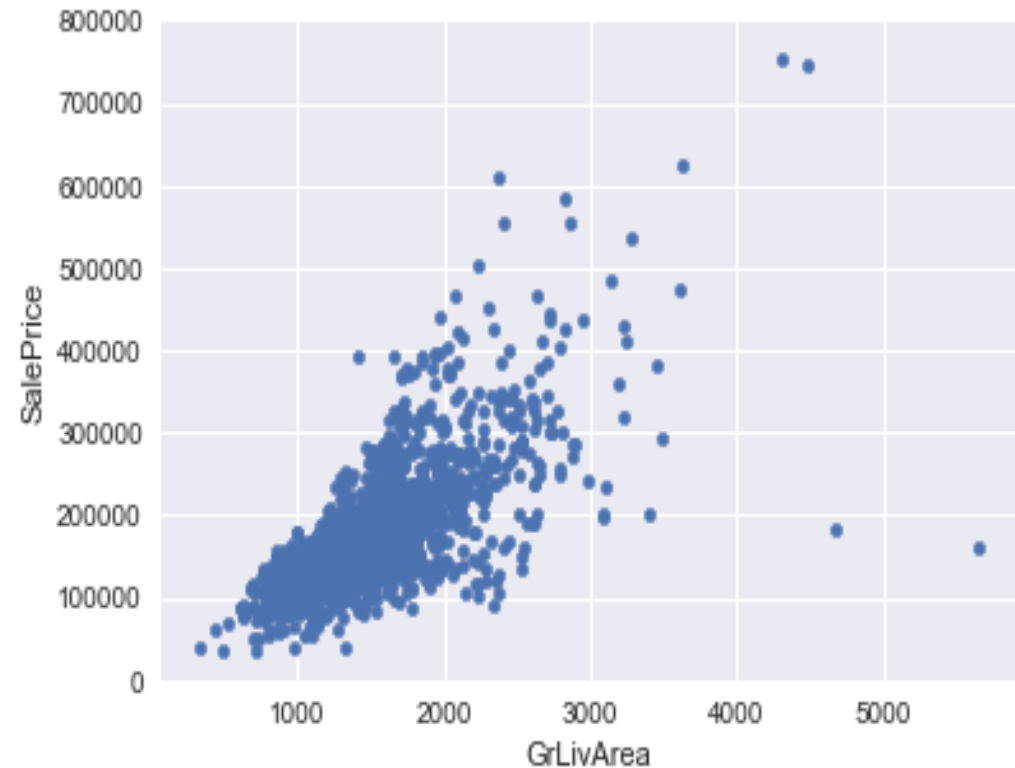
- A bar chart or bar graph is a chart or graph that presents grouped data with rectangular bars with lengths proportional to the values that they represent



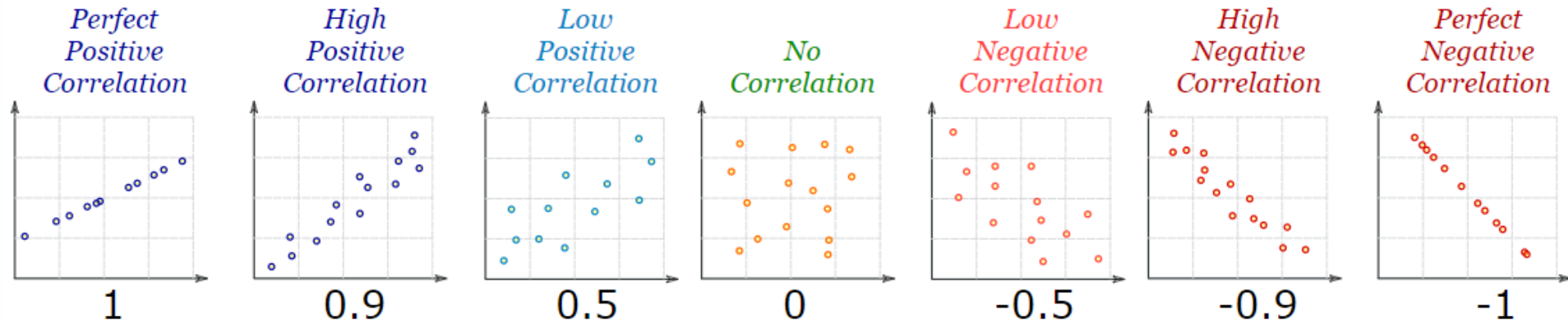


# Two Variable – Scatter Plot

- For two numeric variables
- Use cartesian coordinates to display values for typically two variables for a set of data
- This sample plot reveals a linear relationship between the two variables

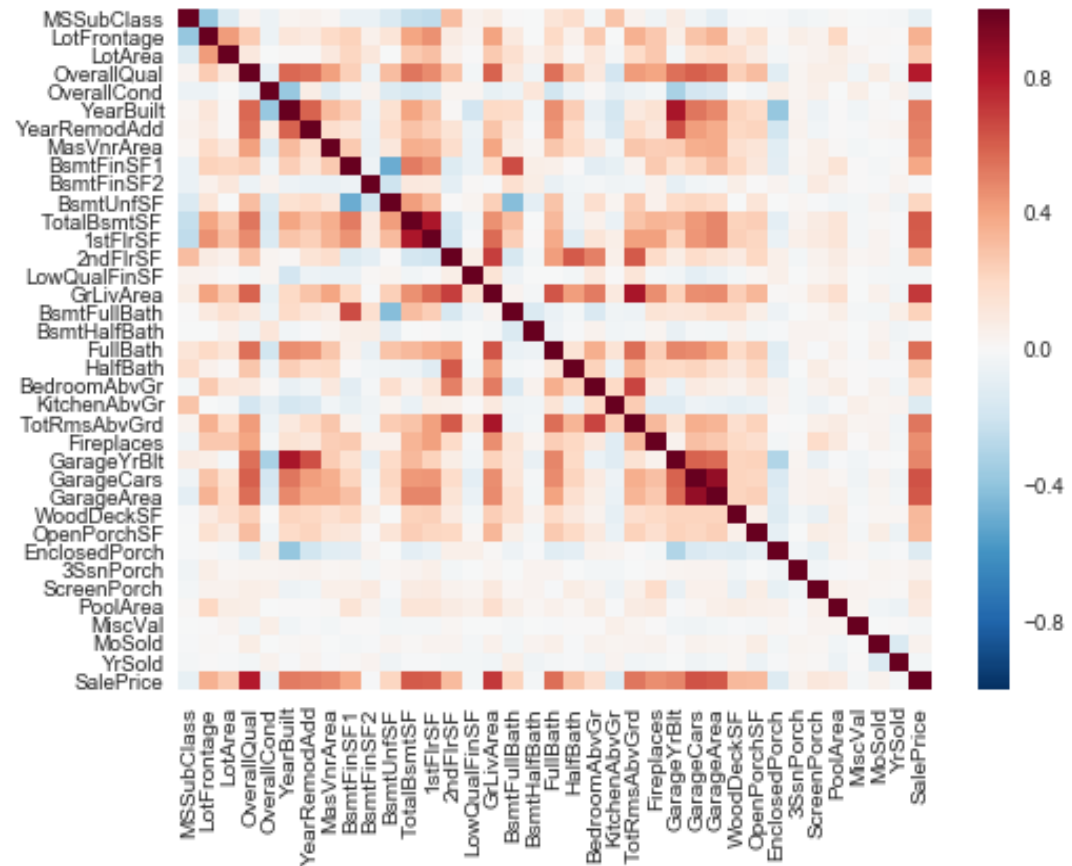


# Two Variable – Scatter Plot



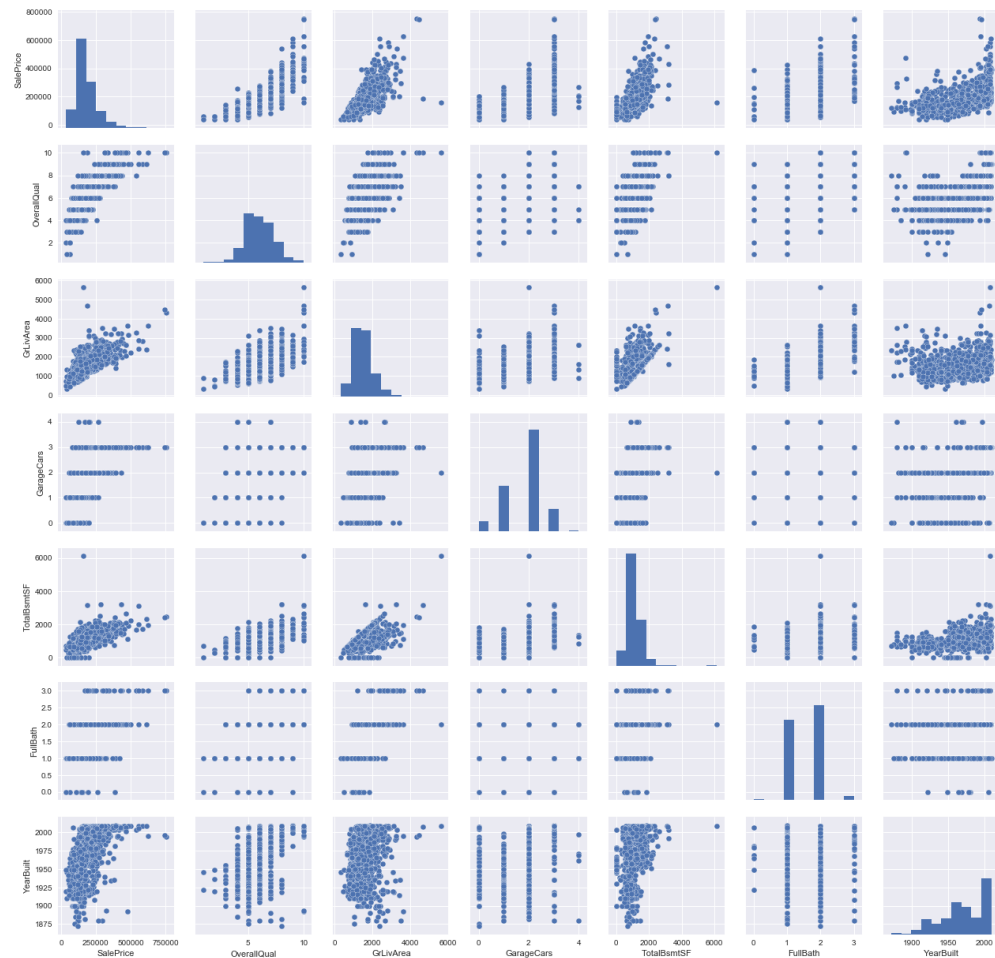
# Multi Variable – Heat Map

- Graphical representation of data where the individual values contained in a matrix are represented as colours
- Helpful to represent the correlation of data



# Multi Variable – Pair Plot

- Another useful seaborn plot is the pair plot which shows the bivariate relation
- Show two across all feature combinations




# Assignment 1

- Lakukan Data Visualization menggunakan datasets *titanic.csv* untuk menganalisa informasi *Single Variable*, *Two Variable* dan *Multi Variable*!
- Akan dipilih dua mentee untuk mempresentasikan hasilnya pada pertemuan berikutnya!



# Hands On Activity

← → ↻ archive.ics.uci.edu/ml/datasets/iris




**UCI**  
Machine Learning Repository  
Center for Machine Learning and Intelligent Systems

## Iris Data Set

Download: [Data Folder](#), [Data Set Description](#)

**Abstract:** Famous database; from Fisher, 1936



<b>Data Set Characteristics:</b>	Multivariate	<b>Number of Instances:</b>	150	<b>Area:</b>	Life
<b>Attribute Characteristics:</b>	Real	<b>Number of Attributes:</b>	4	<b>Date Donated</b>	1988-07-01
<b>Associated Tasks:</b>	Classification	<b>Missing Values?</b>	No	<b>Number of Web Hits:</b>	3759453

**iris setosa**



petal

sepal

**iris versicolor**



petal

sepal

**iris virginica**



petal

sepal



# Hands On Activity

## THE USE OF MULTIPLE MEASUREMENTS IN TAXONOMIC PROBLEMS

By R. A. FISHER, Sc.D., F.R.S.

### I. DISCRIMINANT FUNCTIONS

WHEN two or more populations have been measured in several characters,  $x_1, \dots, x_s$ , special interest attaches to certain linear functions of the measurements by which the populations are best discriminated. At the author's suggestion use has already been made of this fact in craniometry (a) by Mr E. S. Martin, who has applied the principle to the sex differences in measurements of the mandible, and (b) by Miss Mildred Barnard, who showed how to obtain from a series of dated series the particular compound of cranial measurements showing most distinctly a progressive or secular trend. In the present paper the application of the same principle will be illustrated on a taxonomic problem; some questions connected with the precision of the processes employed will also be discussed.

### II. ARITHMETICAL PROCEDURE

Table I shows measurements of the flowers of fifty plants each of the two species *Iris setosa* and *I. versicolor*, found growing together in the same colony and measured by Dr E. Anderson, to whom I am indebted for the use of the data. Four flower measurements are given. We shall first consider the question: What linear function of the four measurements

$$X = \lambda_1 x_1 + \lambda_2 x_2 + \lambda_3 x_3 + \lambda_4 x_4$$

will maximize the ratio of the difference between the specific means to the standard deviations within species? The observed means and their differences are shown in Table II. We may represent the differences by  $d_p$ , where  $p = 1, 2, 3$  or 4 for the four measurements.

The sums of squares and products of deviations from the specific means are shown in Table III. Since fifty plants of each species were used these sums contain 98 degrees of freedom. We may represent these sums of squares or products by  $S_{pq}$ , where  $p$  and  $q$  take independently the values 1, 2, 3 and 4.

Then for any linear function,  $X$ , of the measurements, as defined above, the difference between the means of  $X$  in the two species is

$$D = \lambda_1 d_1 + \lambda_2 d_2 + \lambda_3 d_3 + \lambda_4 d_4,$$

while the variance of  $X$  within species is proportional to

$$S = \sum_{p=1}^4 \sum_{q=1}^4 \lambda_p \lambda_q S_{pq}.$$

The particular linear function which best discriminates the two species will be one for

## 180 MULTIPLE MEASUREMENTS IN TAXONOMIC PROBLEMS

Table I

<i>Iris setosa</i>				<i>Iris versicolor</i>				<i>Iris virginica</i>			
Sepal length	Sepal width	Petal length	Petal width	Sepal length	Sepal width	Petal length	Petal width	Sepal length	Sepal width	Petal length	Petal width
5.1	3.5	1.4	0.2	7.0	3.2	4.7	1.4	6.3	3.3	6.0	2.5
4.9	3.0	1.4	0.2	6.4	3.2	4.5	1.5	5.8	2.7	5.1	1.9
4.7	3.2	1.3	0.2	6.9	3.1	4.9	1.5	7.1	3.0	5.9	2.1
4.6	3.1	1.5	0.2	5.5	2.3	4.0	1.3	6.3	2.9	5.6	1.8
5.0	3.6	1.4	0.2	6.5	2.8	4.6	1.5	6.5	3.0	5.8	2.2
5.4	3.9	1.7	0.4	5.7	2.8	4.5	1.3	7.6	3.0	6.6	2.1
4.6	3.4	1.4	0.3	6.3	3.3	4.7	1.6	4.9	2.5	4.5	1.7
5.0	3.4	1.5	0.2	4.9	2.4	3.3	1.0	7.3	2.9	6.3	1.8
4.4	2.9	1.4	0.2	6.6	2.9	4.6	1.3	6.7	2.5	5.8	1.8
4.9	3.1	1.5	0.1	5.2	2.7	3.9	1.4	7.2	3.6	6.1	2.5
5.4	3.7	1.5	0.2	5.0	2.0	3.5	1.0	6.5	3.2	5.1	2.0
4.8	3.4	1.6	0.2	5.9	3.0	4.2	1.5	6.4	2.7	5.3	1.9
4.8	3.0	1.4	0.1	6.0	2.2	4.0	1.0	6.8	3.0	5.5	2.1
4.3	3.0	1.1	0.1	6.1	2.9	4.7	1.4	5.7	2.5	5.0	2.0
5.8	4.0	1.2	0.2	5.6	2.9	3.6	1.3	5.8	2.8	5.1	2.4
5.7	4.4	1.5	0.4	6.7	3.1	4.4	1.4	6.4	3.2	5.3	2.3
5.4	3.9	1.3	0.4	5.6	3.0	4.5	1.5	6.5	3.0	5.5	1.8
5.1	3.5	1.4	0.3	5.8	2.7	4.1	1.0	7.7	3.8	6.7	2.2
5.7	3.8	1.7	0.3	6.2	2.2	4.5	1.5	7.7	2.6	6.9	2.3
5.1	3.8	1.5	0.3	5.6	2.5	3.9	1.1	6.0	2.2	5.0	1.5
5.4	3.4	1.7	0.2	5.9	3.2	4.8	1.8	6.9	3.2	5.7	2.3
5.1	3.7	1.5	0.4	6.1	2.8	4.0	1.3	5.6	2.8	4.9	2.0
4.6	3.6	1.0	0.2	6.3	2.5	4.9	1.5	7.7	2.8	6.7	2.0
5.1	3.3	1.7	0.5	6.1	2.8	4.7	1.2	6.3	2.7	4.9	1.8
4.8	3.4	1.9	0.2	6.4	2.9	4.3	1.3	6.7	3.3	5.7	2.1
5.0	3.0	1.6	0.2	6.6	3.0	4.4	1.4	7.2	3.2	6.0	1.8
5.0	3.4	1.6	0.4	6.8	2.8	4.8	1.4	6.2	2.8	4.8	1.8
5.2	3.5	1.5	0.2	6.7	3.0	5.0	1.7	6.1	3.0	4.9	1.8
5.2	3.4	1.4	0.2	6.0	2.9	4.5	1.5	6.4	2.8	5.6	2.1
4.7	3.2	1.6	0.2	5.7	2.6	3.5	1.0	7.2	3.0	5.8	1.6
4.8	3.1	1.6	0.2	5.5	2.4	3.8	1.1	7.4	2.8	6.1	1.9
5.4	3.4	1.5	0.4	5.5	2.4	3.7	1.0	7.9	3.8	6.4	2.0
5.2	4.1	1.5	0.1	5.8	2.7	3.9	1.2	6.4	2.8	5.6	2.2
5.5	4.2	1.4	0.2	6.0	2.7	5.1	1.6	6.3	2.8	5.1	1.5
4.9	3.1	1.5	0.2	5.4	3.0	4.5	1.5	6.1	2.6	5.6	1.4
5.0	3.2	1.2	0.2	6.0	3.4	4.5	1.6	7.7	3.0	6.1	2.3
5.5	3.5	1.3	0.2	6.7	3.1	4.7	1.5	6.3	3.4	5.6	2.4
4.9	3.6	1.4	0.1	6.3	2.3	4.4	1.3	6.4	3.1	5.5	1.8

Thanks!