STIA 210 : Intro to Data Science

Prof. Venkatesh Krishnamoorthy
Prof. Kyle Meyer

Week 3: Data Visualization

Edward Tufte Quotes - https://www.edwardtufte.com/tufte/

“Above all else show the data.”

“If the statistics are boring, then you've got the wrong numbers.”

- *The Visual Display of Quantitative Information*

Edward Tufte Quotes - https://www.edwardtufte.com/tufte/

“Graphical excellence is that which gives to the viewer the greatest number of ideas in the shortest time with the least ink in the smallest space.”

- *The Visual Display of Quantitative Information*

Edward Tufte Quotes - https://www.edwardtufte.com/tufte/

The commonality between science and art is in trying to see profoundly - to develop strategies of seeing and showing

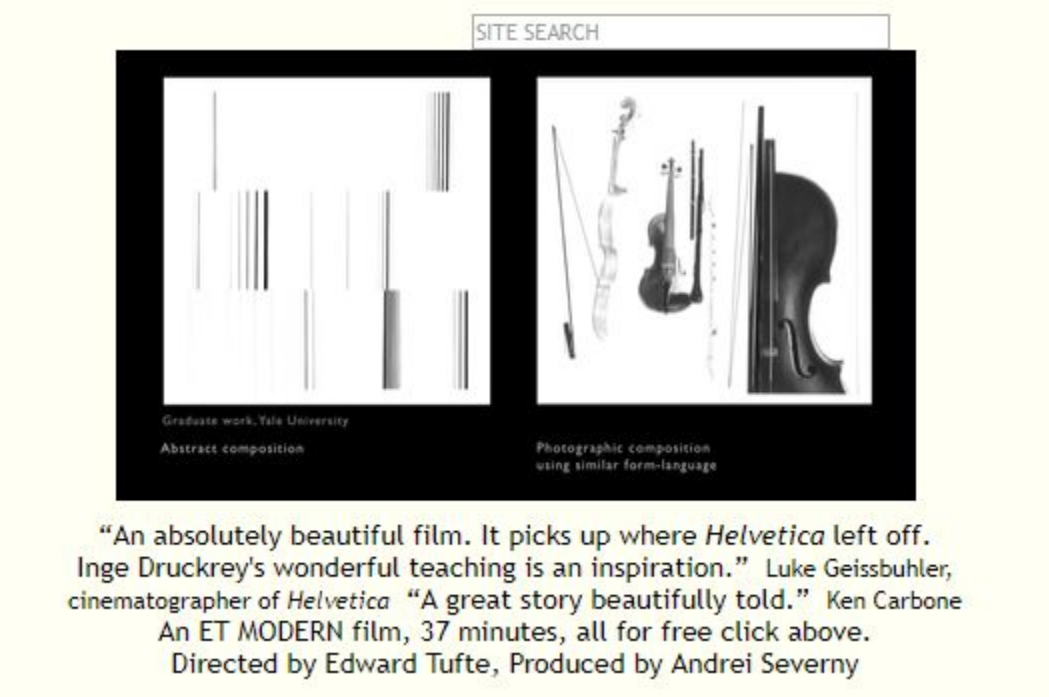
Beautiful Evidence is about the theory and practice of analytical design.

The minimum we should hope for with any display technology is that it should do no harm.

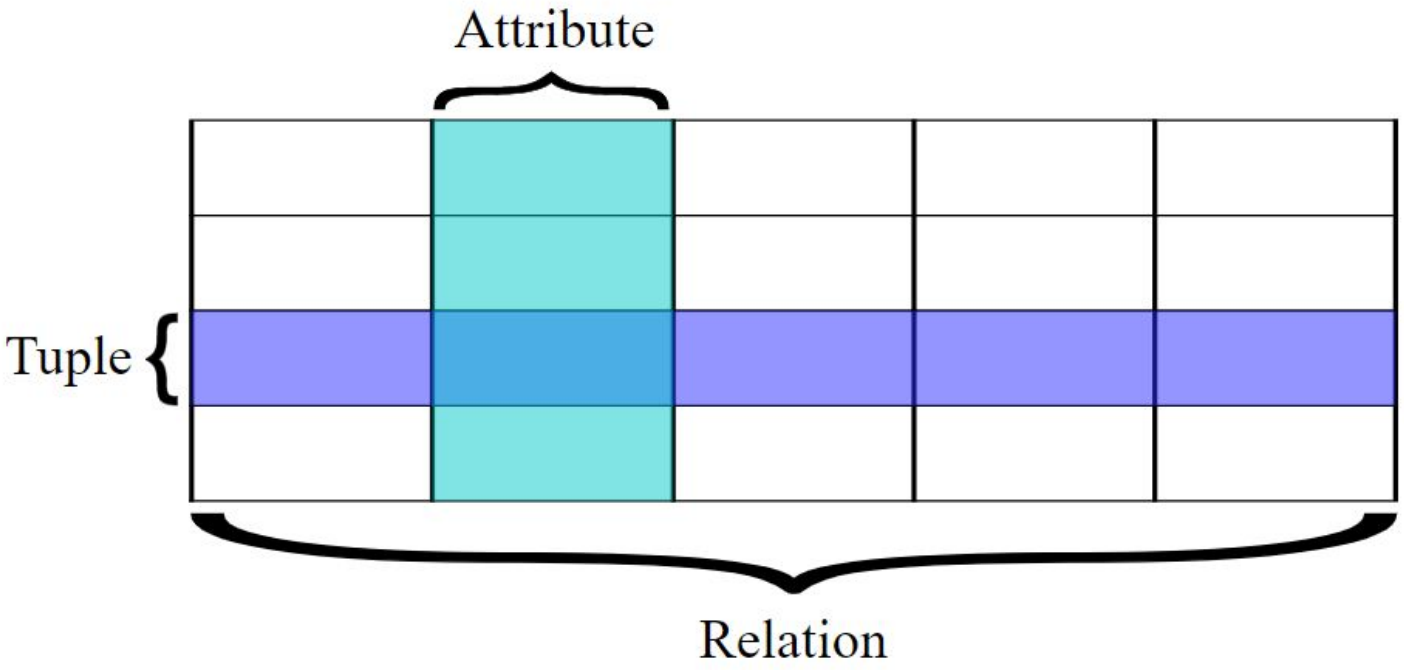
The leading edge in evidence presentation is in science; the leading edge in beauty is in high art.

If you like overheads, you'll love PowerPoint.

Beautiful Evidence - https://www.edwardtufte.com/tufte/



Data Organization - Rows and Columns



Relation, tuple, and attribute represented as table, row, and column respectively.

Quantitative and Categorical Data

Quantitative or **Numerical Data** are numbers,

- Can have an sorting order
- Examples are age, height, weight

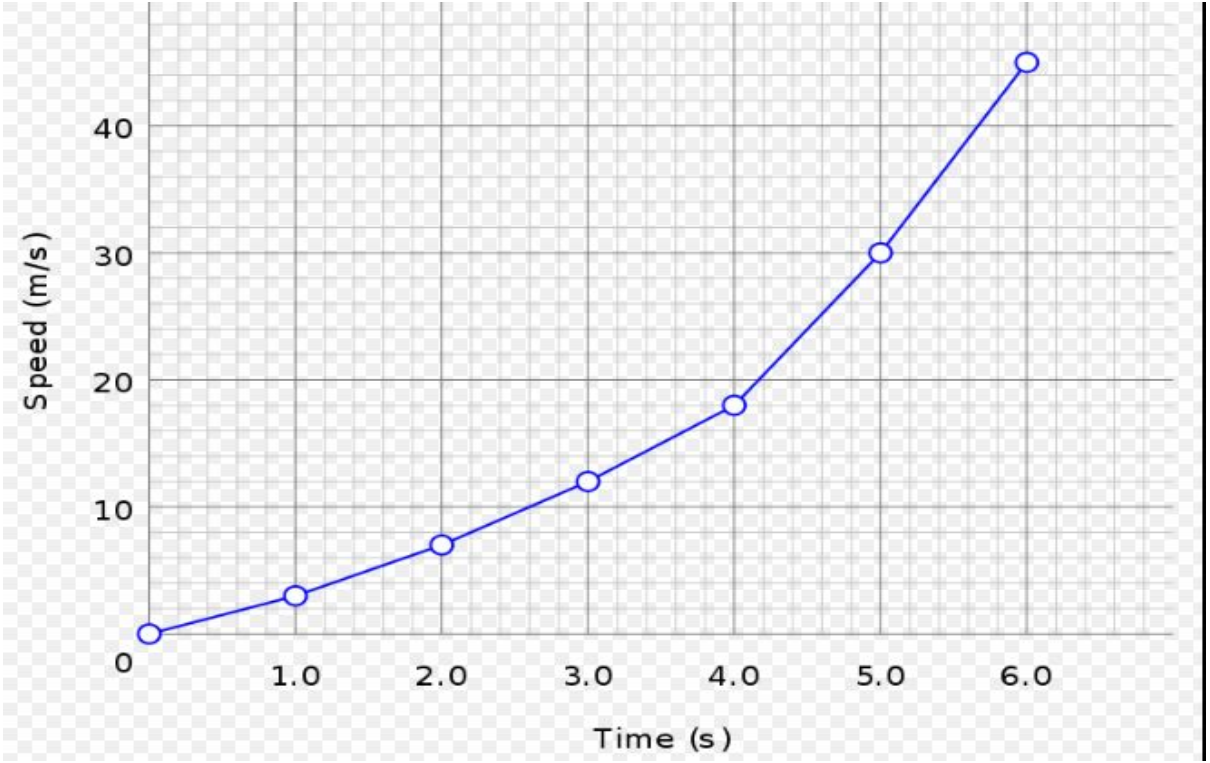
Categorical Data can take on one of a limited, and usually fixed number of possible values, assigning each individual or other unit of observation to a particular group on the basis of some qualitative property

- Blood type of a person: A, B, AB or O.
- State that a person lives in.
- Part of speech : Noun, Adjective Verb, Adverb, ...

Graphing Data

- Line Plot
- Histogram
- Scatter Plot
- Bar Plot
- Box Plot

Line Plot - series of data points connected by straight line segments

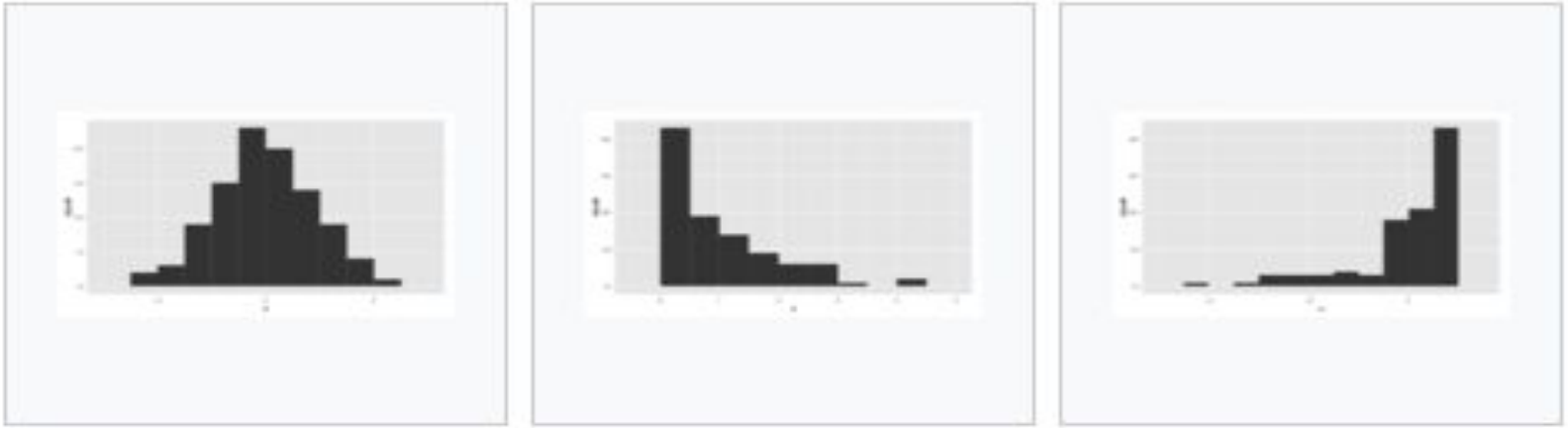


What does a Line Plot Show?

- Trends
- Cyclical Patterns
- Variable Comparison

Histogram - distribution of numeric variables

- A **histogram** is an accurate representation of the distribution of numerical data. It is an estimate of the probability distribution of a continuous variable and was first introduced by Karl Pearson. It differs from a bar graph, in the sense that a bar graph relates two variables, but a histogram relates only one.



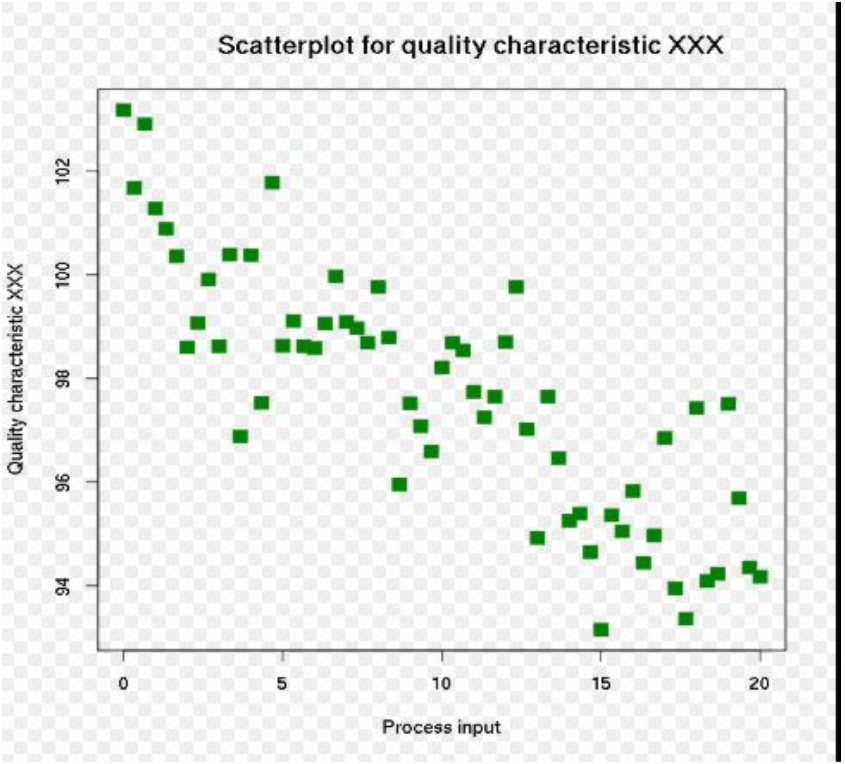
What a Histogram shows?

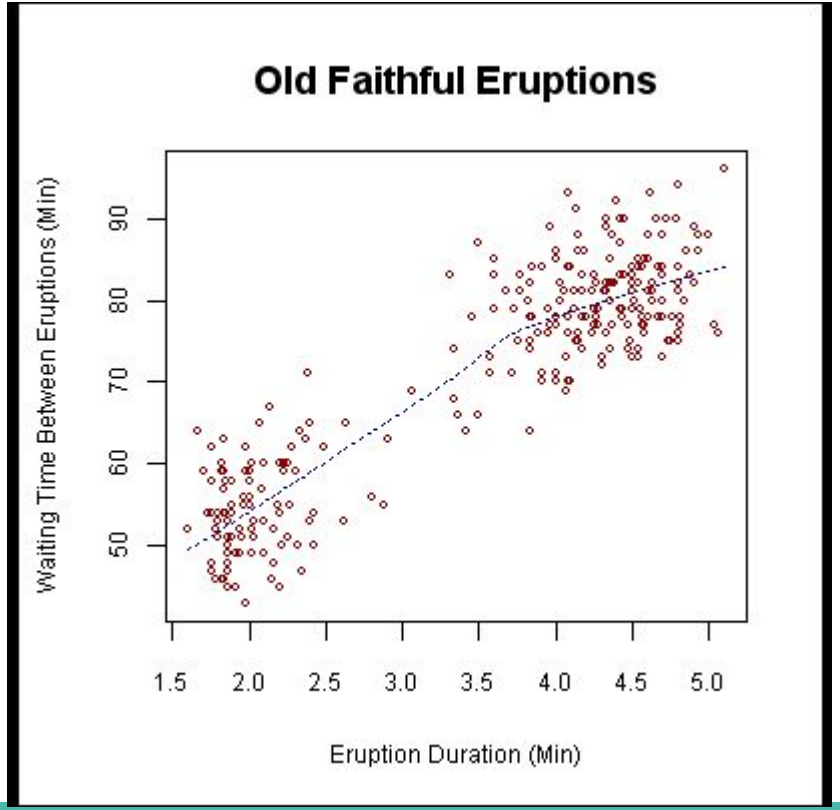
Skewness

Central Tendency

Outliers

Scatter Plot - relationship between two variables



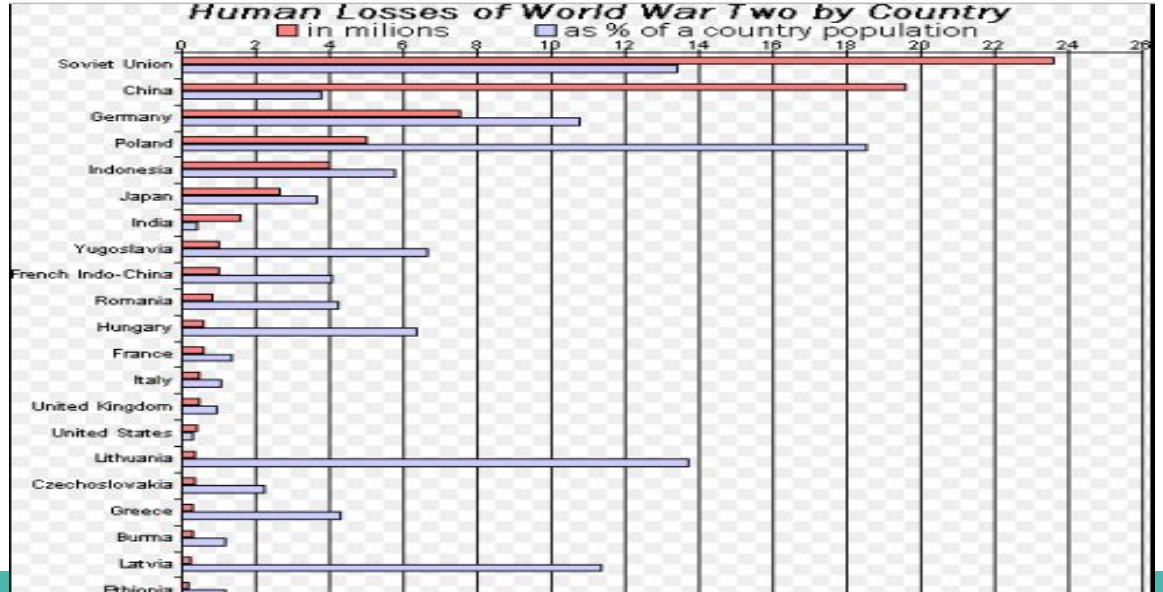


What a Scatter Plot Shows?

- Positive Correlation
- Negative Correlation
- No Correlation
- Non-Linear Correlation

Bar Plot - distribution of categorical variable

- A **bar chart** or **bar graph** is a chart or graph that presents categorical data with rectangular bars with heights or lengths proportional to the values that they represent. The bars can be plotted vertically or horizontally. A vertical bar chart is sometimes called a **line graph**



What a bar plot shows?

Bar graphs can also be used for more complex comparisons of data with grouped bar charts and stacked bar charts.

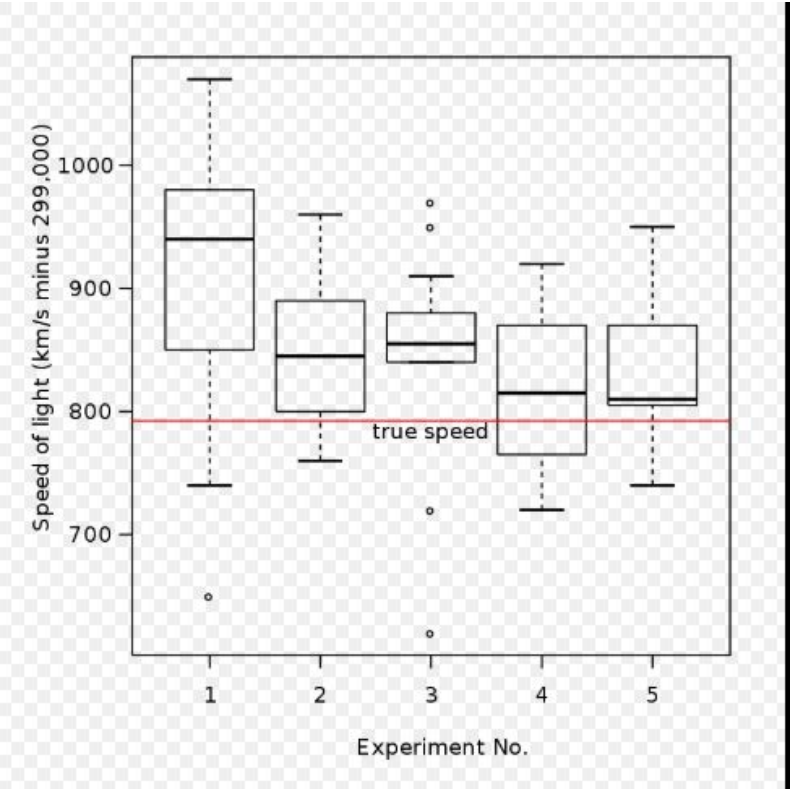
Grouped bar chart, for each categorical group there are two or more bars. These bars are color-coded to represent a particular grouping. For example, a business owner with two stores might make a grouped bar chart with different colored bars to represent each store: the horizontal axis would show the months of the year and the vertical axis would show the revenue.

Alternatively, a **stacked bar chart** could be used. The stacked bar chart stacks bars that represent different groups on top of each other. The height of the resulting bar shows the combined result of the groups. However, stacked bar charts are not suited to data sets where some groups have negative values. In such cases, grouped bar chart are preferable.

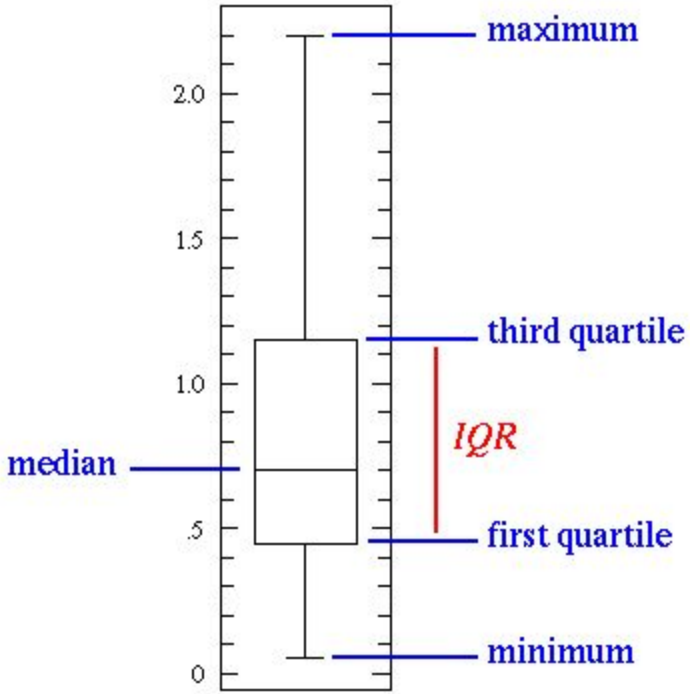
Box Plot - Compares distributions of variables

A boxplot is a method for graphically depicting groups of numerical data through their quartiles.

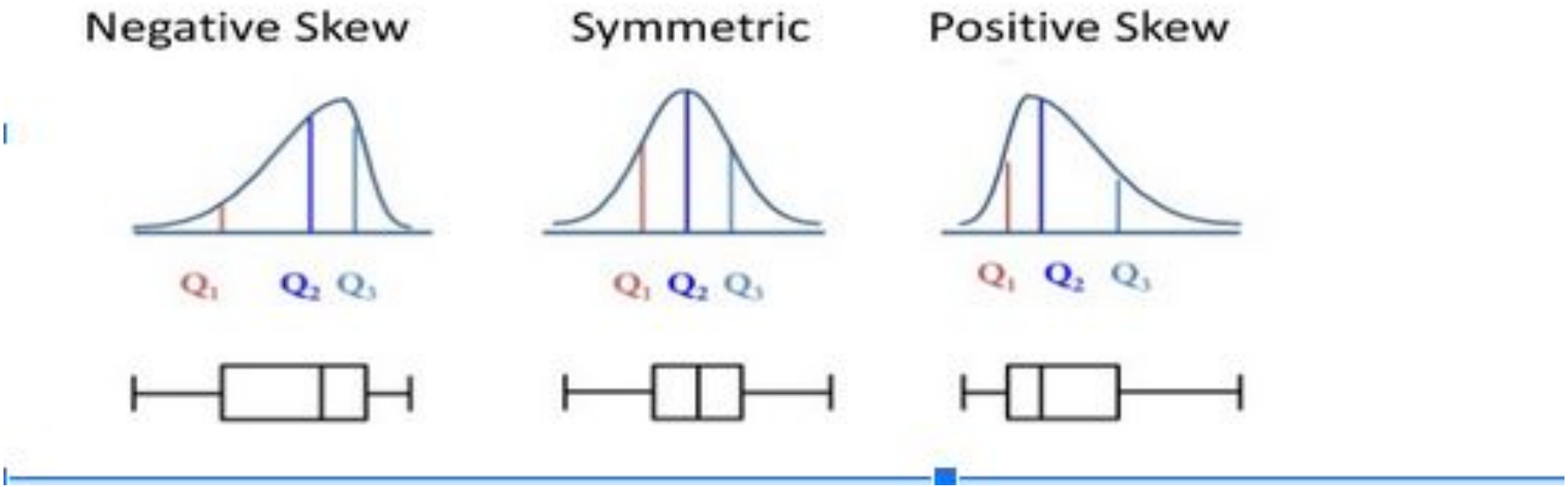
Box plots may also have lines extending vertically from the boxes (*whiskers*) indicating variability outside the upper and lower quartiles, hence the terms **box-and-whisker plot**



Components of a Box Plot

- 

What a Box Plot Shows?



Summary

- Look at data intuitively
- Useful for data exploration
- Aids communication of results

Matplotlib

- Matplotlib and pyplot libraries basics
- Use matplotlib and generate simple plots with python

Matplotlib

Matplotlib is a powerful library that can be used to generate quick visualizations and quality graphics

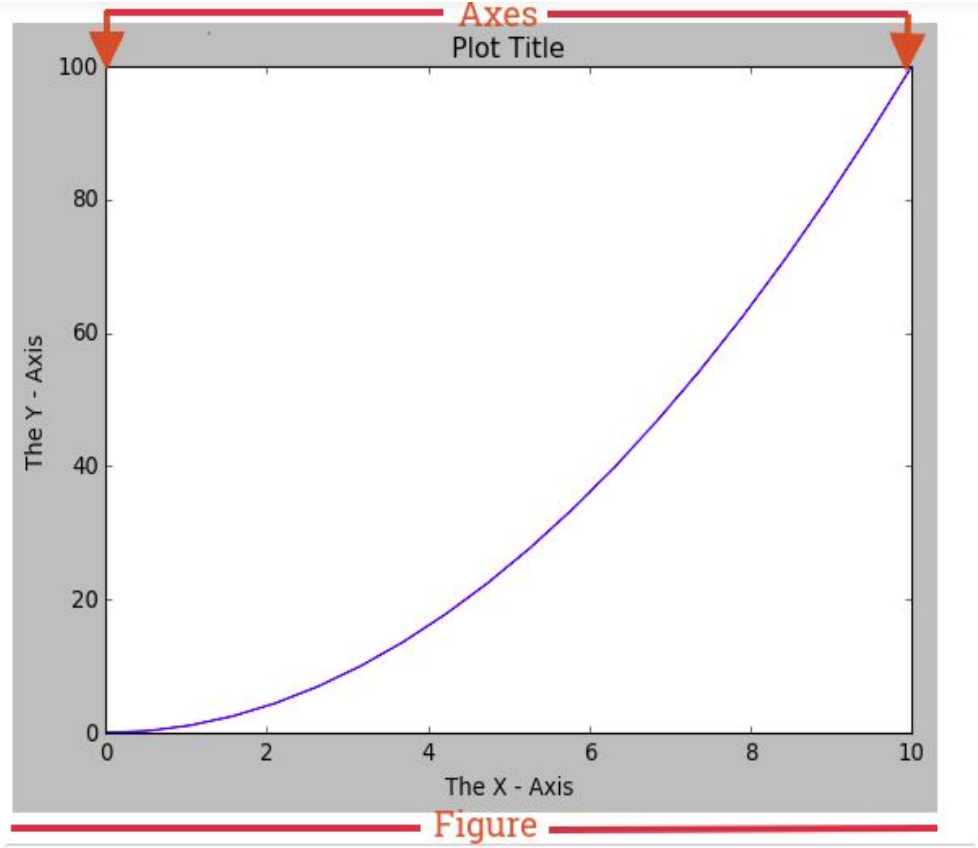
• Let us introduce basic functionality (via pyplot), such as bar and line plots

Examples of code of the types of plots that can be generated are available at **https://matplotlib.org/gallery/index.html**

fig, ax=plt.subplots

- fig, ax=plt.subplots
- plt is PYPLOT - pyplot is matplotlib's graphing framework.
- fig is the FIGURE
- The weird part is, though, your actual graph is not the figure. The figure is the part around your graph. Your chart sits on top of the figure. So what's your visualization?
- Your graph is what's called a subplot or axis. Or, technically, an AxesSubplot.
- ax is the AXIS or SUBPLOT

Components of a Plot - Figure & Azes



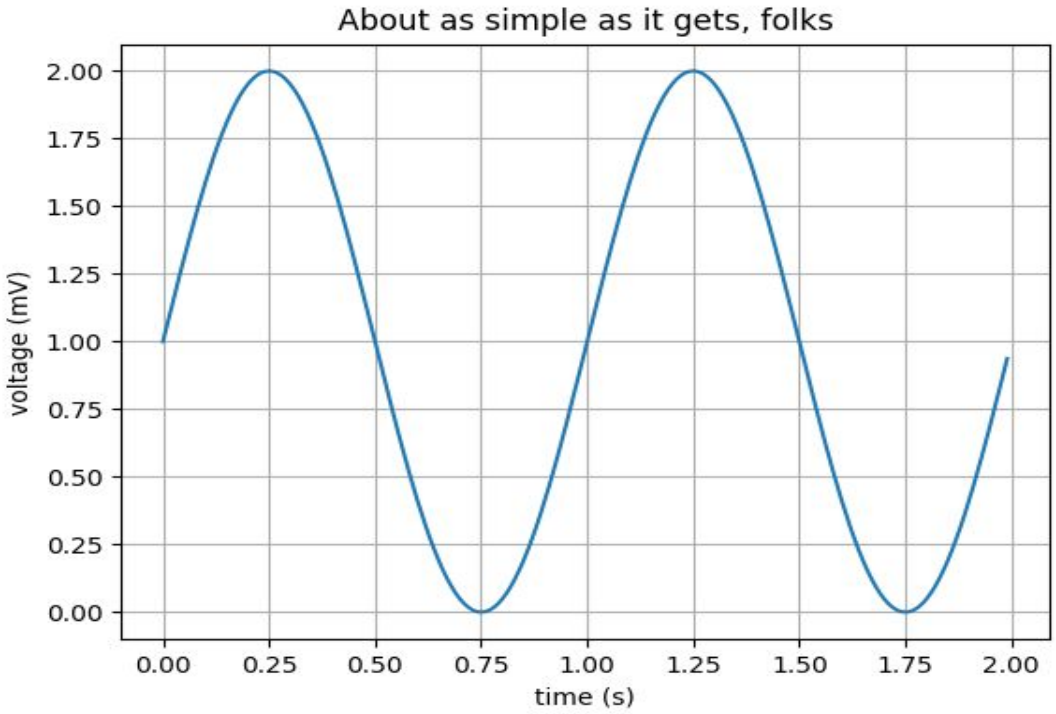
Generate simple statistics - code

Draw a Simple Line Plot - code

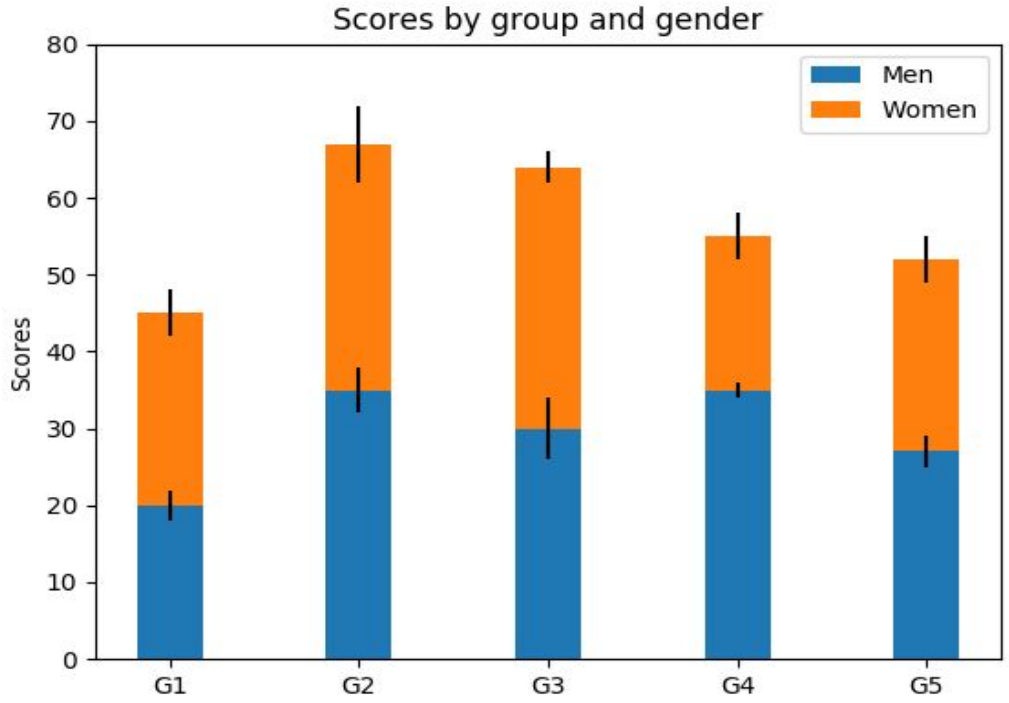
Create a simple plot.

```
import matplotlib
import matplotlib.pyplot as plt
import numpy as np

# Data for plotting
t = np.arange(0.0, 2.0, 0.01)
s = 1 + np.sin(2 * np.pi * t)

fig, ax = plt.subplots()
ax.plot(t, s)

ax.set(xlabel='time (s)', ylabel='voltage (mV)',
       title='About as simple as it gets, folks' )
ax.grid()

fig.savefig("test.png")
plt.show()
```

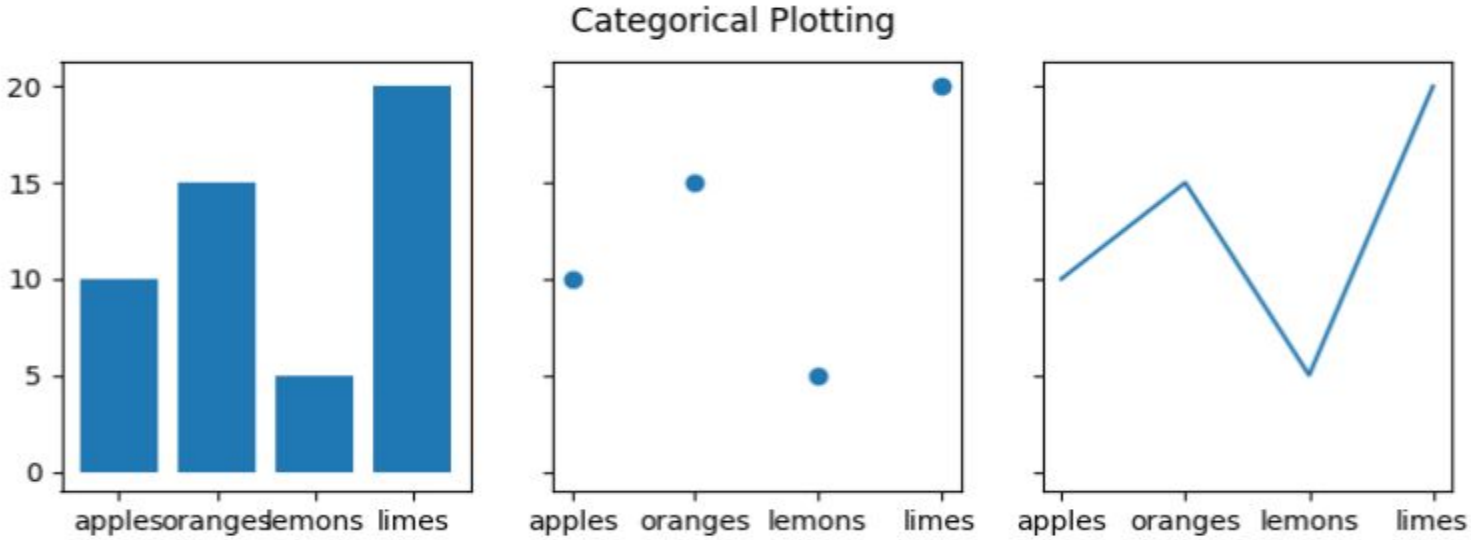
Draw a Simple Line Plot



Draw a Stacked Bar Graph - code



Plotting categorical variables



NUMPY - Basics

NumPy’s main object is the homogeneous multidimensional array. It is a table of elements (usually numbers), all of the same type, indexed by a tuple of non-negative integers. In NumPy dimensions are called *axes*.

NumPy’s array class is called `ndarray`. It is also known by the alias `array`. The more important attributes of an `ndarray` object are:

NUMPY - Array Attributes

Ndarray.ndim - the number of axes (dimensions) of the array.

Ndarray.shape - the dimensions of the array. This is a tuple of integers indicating the size of the array in each dimension. For a matrix with *n* rows and *m* columns, `shape` will be `(n,m)`. The length of the `shape` tuple is therefore the number of axes, `ndim`.

Ndarray.size - the total number of elements of the array. This is equal to the product of the elements of `shape`.

NUMPY - Array Attributes

Ndarray.dtype - an object describing the type of the elements in the array. One can create or specify dtype's using standard Python types. Additionally NumPy provides types of its own. numpy.int32, numpy.int16, and numpy.float64 are some examples.

Ndarray.itemsize - the size in bytes of each element of the array. For example, an array of elements of type `float64` has `itemsize` 8 (=64/8), while one of type `complex32` has `itemsize` 4 (=32/8). It is equivalent to `ndarray.dtype.itemsize`.

Ndarray.data - the buffer containing the actual elements of the array. Normally, we won't need to use this attribute because we will access the elements in an array using indexing facilities.

Data Visualization - Analyzing the IRIS Dataset

- Iris Data Set
- Analysis and Visualization

Iris flower data set (Fisher's *Iris* data set)

- Multivariate data set - by the British statistician-biologist Ronald Fisher in 1936 - classic, The Use of Multiple Measurements in Taxonomic Problems, and can also be found on the UCI Machine Learning Repository.
- Data to quantify the morphologic variation of *Iris* flowers of three related species
- The data set consists of 50 samples from each of three species of *Iris* (*Iris setosa*, *Iris virginica* and *Iris versicolor*).
- Four features were measured from each sample: the length and the width of the sepals and petals, in centimeters.
- Based on the combination of these four features, Fisher developed a linear discriminant model to distinguish the species from each other.

Morphologic variation of *3* of related species

Iris setosa

Iris versicolor

Iris virginica

About this Dataset

One flower species is linearly separable from the other two, but the other two are not linearly separable from each other.

The columns in this dataset are:

- Id
- SepalLengthCm
- SepalWidthCm
- PetalLengthCm
- PetalWidthCm
- Species