**INTRODUCTION**

**Data Analysis and Reporting**

The manner in which reliability data is analyzed and reported will largely have to be tailored to the specific circumstance or organization. However, it is possible to break down the general methods of analysis/reporting into two categories: [non-parametric analysis](http://www.weibull.com/hotwire/issue44/hottopics44.htm#Non-Parametric Analysis) and [parametric analysis](http://www.weibull.com/hotwire/issue44/hottopics44.htm#Parametric Analysis). Overall, it will be necessary to tailor the analysis and reporting methods by the type of data as well as to the intended audience. Managers will generally be more interested in actual data and non-parametric analysis results, while engineers will be more concerned with parametric analysis. Of course this is a rather broad generalization and if the proper training has instilled the organization with an appreciation of the importance of reliability engineering, there should be an interest in all types of reliability reports at all levels of the organization. Nevertheless, managers are usually more interested in the "big picture" information that non-parametric analyses generally tend to provide, while not being particularly interested in the level of technical detail that parametric analyses provide. On the other hand, engineers and technicians are usually more concerned with the close-up details and technical information that parametric analyses provide. Both of these types of data analysis have a great deal of importance to any given organization, and it is merely necessary to apply the different types in the proper places.

**Non-Parametric Analysis**

Data conducive to non-parametric analysis includes information that has not or cannot be rigorously processed or analyzed. Usually, it is simply straight reporting of information, or if it has been manipulated, it is usually by simple mathematics, with no complex statistical analysis. In this respect, many types of field data lend themselves to the non-parametric type of analysis and reporting. In general, this type of information will be of most interest to managers as it usually requires no special technical know-how to interpret. Another reason it is of particular interest to managers is that most financial data falls into this category. Despite its relative simplicity, the importance of non-parametric data analysis should not be underestimated. Most of the important decisions that are made concerning the business are based on non-parametric analysis of financial data.

As mentioned in last month's issue of the *HotWire*(["Data Collection"](http://www.weibull.com/hotwire/issue43/hottopics43.htm)), ReliaSoft's Dashboard system is a powerful tool for collecting and reporting data. It especially lends itself to non-parametric data analysis and reporting, as it can be quickly processed and manipulated in accordance with the user's wishes.

# Six Types Of Analyses Every Data Scientist Should Know

The identified six(6) archetypical analyses. As presented, they range from the least to most complex, in terms of knowledge, costs, and time. In summary,

* Descriptive
* Exploratory
* Inferential
* Predictive
* Causal
* Mechanistic

**1. Descriptive** (least amount of effort):  The discipline of quantitatively describing the main features of a collection of data. In essence, it describes a set of data.

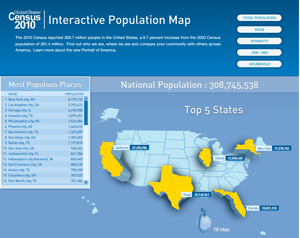
– Typically the first kind of data analysis performed on a data set

– Commonly applied to large volumes of data, such as census data

-The description and interpretation processes are different steps

– Univariate and Bivariate are two types of statistical descriptive analyses.

– Type of data set applied to: Census Data Set – a whole population

 Example: Census Data[](http://www.census.gov/2010census/popmap/)

**2. Exploratory**: An approach to analyzing data sets to find previously unknown relationships.

– Exploratory models are good for discovering new connections

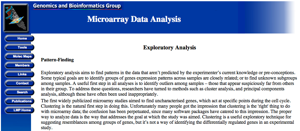
– They are also useful for defining future studies/questions

– Exploratory analyses are usually not the definitive answer to the question at hand, but only the start

– Exploratory analyses alone should not be used for generalizing and/or predicting

– Remember: correlation does not imply causation

– Type of data set applied to: Census and Convenience Sample Data Set (typically non-uniform) – a random sample with many variables measured

Example: Microarray Data Analysis [](http://discover.nci.nih.gov/microarrayAnalysis/Exploratory.Analysis.jsp)

**3. Inferential**: Aims to test theories about the nature of the world in general (or some part of it) based on samples of “subjects” taken from the world (or some part of it). That is, use a relatively small sample of data to say something about a bigger population.

– Inference is commonly the goal of statistical models

– Inference involves estimating both the quantity you care about and your uncertainty about your estimate

– Inference depends heavily on both the population and the sampling scheme

– Type of data set applied to: Observational, Cross Sectional Time Study, and Retrospective Data Set – the right, randomly sampled population

Example: Inferential Analysis [](http://www.biomedcentral.com/1471-2105/10/288)

**4. Predictive**: The various types of methods that analyze current and historical facts to make predictions about future events. In essence, to use the data on some objects to predict values for another object.

– The models predicts, but it does not mean that the independent variables cause

– Accurate prediction depends heavily on measuring the right variables

– Although there are better and worse prediction models, more data and a simple model works really well

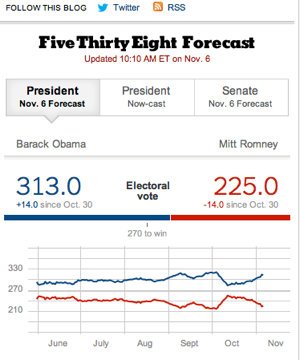
– Prediction is very hard, especially about the future references

– Type of data set applied to: Prediction Study Data Set – a training and test data set from the same population

Example: Predictive Analysis

[](http://www.heritagehealthprize.com/c/hhp)

Another Example of Predictive Analysis

[](http://fivethirtyeight.blogs.nytimes.com/)

**5. Causal**: To find out what happens to one variable when you change another.

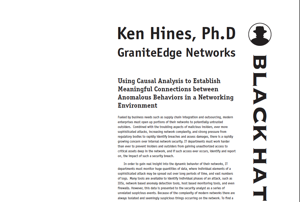
– Implementation usually requires randomized studies

– There are approaches to inferring causation in non-randomized studies

– Causal models are said to be the “gold standard” for data analysis

– Type of data set applied to: Randomized Trial Data Set – data from a randomized study

Example: Causal Analysis

[](http://www.blackhat.com/presentations/bh-usa-05/bh-us-05-hines.pdf)

**6. Mechanistic** (most amount of effort): Understand the exact changes in variables that lead to changes in other variables for individual objects.

– Incredibly hard to infer, except in simple situations

– Usually modeled by a deterministic set of equations (physical/engineering science)

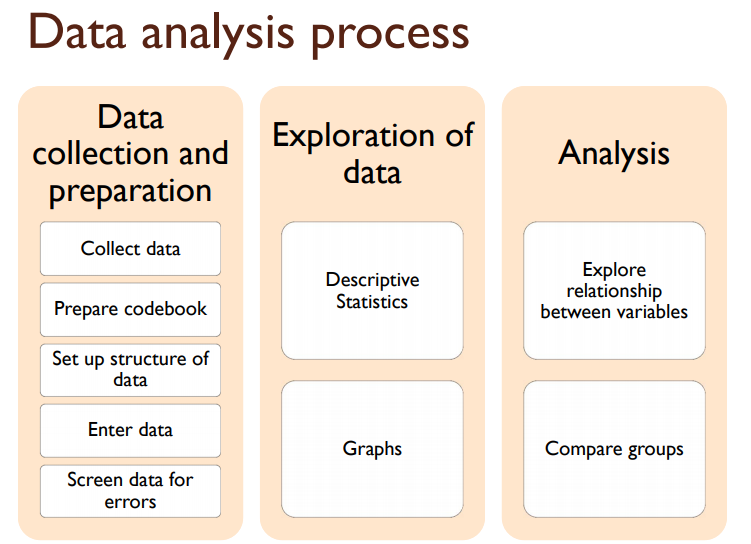
– Generally the random component of the data is measurement error

– If the equations are known but the parameters are not, they may be inferred with data analysis

– Type of data set applied to: Randomized Trial Data Set – data about all components of the system

Example: Mechanistic Analysis

[](http://www.ellisonfoundation.org/content/mechanistic-analysis-transcriptional-switch-regulating-p53-activity-and-premature-aging)

****

**0) Understanding the business:** This starts under the premise that the business sponsors need an analytical solution to a problem. In this process it is very important to keep in contact throughout the project. This part plays an essential role, hence the problem and the goal of it must be well posted and well understood.

**1) Analytic Approach:** After the problem is understood, we as data scientists must be able to find an analytical approach. We need to identify an analytical approach so as to have a top to bottom process. Assuming we are in the realm of supervised learning, we first need to identify whether the model is a classification or a regression model. Then we could choose one or more different models to rely on.

**2) Data requirements:** Step 1 determines this step. Here we identify which data do we need in order to address the analytical approach

**3) Data Collection:** Initially data scientists have to identify and collect data from the available resources. Normally we require additional investment to gather more specific data. However, it is advisable to postpone the investment decision until there is more knowledge about the data and the model. If we incorporate additional data, the predictive models can better represent events such as an earthquake or a disease.

**4) Data Understanding:** This is the prequel of Exploratory Data Analysis. This starts by doing some descriptive statistics and visualization to get the first insights about the data content. Be careful because additional data may be required.

**5) Data Preparation:** This stage covers everything that involves preparing the data for the modeling stage. Its activities involve data cleaning (deal with missing values, duplicates or formatting), combining data from multiple resources and transforming the data into the useful variables. This is the most time consuming step in a Data Science project. Up to 80% of the time could be consumed if the data is messy.

**6) Modeling:** This stage focuses on developing predictive or descriptive models, following the analytical approach. When we use a predictive model, we normally split the data to leave a minority for training the model, and a majority for testing the model. Here we get intermediate insight that leads into refinement. This leads into testing multiple algorithms aiming to find the best model. Hence this stage is highly iterative.

**7) Evaluation:** Between the development of the model and its deployment, we as data scientists must evaluate how well does our model performs. We have to understand its quality and ensure it properly addresses the business problem. Here we compute various diagnostic measures and outputs such as graphs and tables. In this part we begin to understand and interpret the model's quality and efficacy. Predictive models work on a previously trained algorithm with a known dataset (in the case of supervised machine learning). The testing part uses either created data that follows the same probabilistic distribution, or part of the original data that is left exclusively for testing and tuning. Furtherly we may need to assess our model with statistical significance tests for further quality proof. Practically this proof can be used to justify the model implementation, or taking actions when the stakes are high.

**8) Deployment:** Once a model is developed and approved by business sponsors, the model is tested in the business environment or a similar test environment which is enclosed or limited for further evaluations, until it is fully assessed. It may be as simple as making a report with recommendations, or as involved as in a complex workflow scenario. This usually involves groups with additional skillsets from within the enterprise.

**9) Feedback:** After getting results from the implemented model, the organization gets feedback on the model's performance and its impact on the environment where it was deployed. This stage also has some refining to improve its accuracy and usefulness. We can automate some of the feedback gathering and the evaluation, and redeployment to respond to the dynamism from the environment.

An important aspect of this methodology is to consider that this is a highly iterative process. We as data scientists usually have to go back to previous processes to make adjustments.

**EXPLORATORY DATA ANALYSIS**

In [statistics](https://en.wikipedia.org/wiki/Statistics), **exploratory data analysis** (**EDA**) is an approach to [analyzing](https://en.wikipedia.org/wiki/Data_analysis" \o "Data analysis) [data sets](https://en.wikipedia.org/wiki/Data_set) to summarize their main characteristics, often with visual methods. A [statistical model](https://en.wikipedia.org/wiki/Statistical_model) can be used or not, but primarily EDA is for seeing what the data can tell us beyond the formal modeling or hypothesis testing task. Exploratory data analysis was promoted by [John Tukey](https://en.wikipedia.org/wiki/John_Tukey) to encourage statisticians to explore the data, and possibly formulate hypotheses that could lead to new data collection and experiments. EDA is different from [initial data analysis](https://en.wikipedia.org/w/index.php?title=Initial_data_analysis&action=edit&redlink=1) (IDA),[[1]](https://en.wikipedia.org/wiki/Exploratory_data_analysis#cite_note-1) which focuses more narrowly on checking assumptions required for model fitting and hypothesis testing, and handling missing values and making transformations of variables as needed. EDA encompasses IDA.

Tukey defined data analysis in 1961 as: "[P]rocedures for analyzing data, techniques for interpreting the results of such procedures, ways of planning the gathering of data to make its analysis easier, more precise or more accurate, and all the machinery and results of (mathematical) statistics which apply to analyzing data."[[2]](https://en.wikipedia.org/wiki/Exploratory_data_analysis#cite_note-2)

Tukey's championing of EDA encouraged the development of [statistical computing](https://en.wikipedia.org/wiki/Computational_statistics) packages, especially [*S*](https://en.wikipedia.org/wiki/S_(programming_language)) at [Bell Labs](https://en.wikipedia.org/wiki/Bell_Labs). The *S*programming language inspired the systems ['S'-PLUS](https://en.wikipedia.org/wiki/S-PLUS) and [*R*](https://en.wikipedia.org/wiki/R_(programming_language)). This family of statistical-computing environments featured vastly improved dynamic visualization capabilities, which allowed statisticians to identify [outliers](https://en.wikipedia.org/wiki/Outlier), [trends](https://en.wikipedia.org/wiki/Trend_estimation) and [patterns](https://en.wikipedia.org/wiki/Pattern_recognition) in data that merited further study.

Tukey's EDA was related to two other developments in [statistical theory](https://en.wikipedia.org/wiki/Statistical_theory): [robust statistics](https://en.wikipedia.org/wiki/Robust_statistics) and [nonparametric statistics](https://en.wikipedia.org/wiki/Nonparametric_statistics), both of which tried to reduce the sensitivity of statistical inferences to errors in formulating [statistical models](https://en.wikipedia.org/wiki/Statistical_model). Tukey promoted the use of [five number summary](https://en.wikipedia.org/wiki/Five_number_summary) of numerical data—the two [extremes](https://en.wikipedia.org/wiki/Extreme_value) ([maximum](https://en.wikipedia.org/wiki/Maximum) and [minimum](https://en.wikipedia.org/wiki/Minimum)), the [median](https://en.wikipedia.org/wiki/Median), and the [quartiles](https://en.wikipedia.org/wiki/Quartile)—because these median and quartiles, being functions of the [empirical distribution](https://en.wikipedia.org/wiki/Empirical_distribution_function) are defined for all distributions, unlike the [mean](https://en.wikipedia.org/wiki/Mean_value) and [standard deviation](https://en.wikipedia.org/wiki/Standard_deviation); moreover, the quartiles and median are more robust to [skewed](https://en.wikipedia.org/wiki/Skewness) or [heavy-tailed](https://en.wikipedia.org/wiki/Heavy-tailed_distribution) [distributions](https://en.wikipedia.org/wiki/Probability_distribution) than traditional summaries (the mean and standard deviation). The packages *S*, *S*-PLUS, and *R* included routines using [resampling statistics](https://en.wikipedia.org/wiki/Resampling_(statistics)), such as Quenouille and Tukey's [jackknife](https://en.wikipedia.org/wiki/Resampling_(statistics)" \l "Jackknife" \o "Resampling (statistics)) and [Efron](https://en.wikipedia.org/wiki/Bradley_Efron" \o "Bradley Efron)'s [bootstrap](https://en.wikipedia.org/wiki/Bootstrapping_(statistics)), which are nonparametric and robust (for many problems).

Exploratory data analysis, robust statistics, nonparametric statistics, and the development of statistical programming languages facilitated statisticians' work on scientific and engineering problems. Such problems included the fabrication of semiconductors and the understanding of communications networks, which concerned Bell Labs. These statistical developments, all championed by Tukey, were designed to complement the [analytic](https://en.wikipedia.org/wiki/Analytic_function) theory of [testing statistical hypotheses](https://en.wikipedia.org/wiki/Statistical_hypothesis_testing), particularly the [Laplacian](https://en.wikipedia.org/wiki/Pierre-Simon_Laplace) tradition's emphasis on [exponential families](https://en.wikipedia.org/wiki/Exponential_family).[[3]](https://en.wikipedia.org/wiki/Exploratory_data_analysis#cite_note-3)

**Exploratory Data Analysis**

“The greatest value of a picture is when it forces us to notice what we never expected to see.” -John W. Tukey

Biases, systematic errors and unexpected variability are common in data from the life sciences. Failure to discover these problems often leads to flawed analyses and false discoveries. As an example, consider that experiments sometimes fail and not all data processing pipelines, such as the t.test function in R, are designed to detect these. Yet, these pipelines still give you an answer. Furthermore, it may be hard or impossible to notice an error was made just from the reported results.

Graphing data is a powerful approach to detecting these problems. We refer to this as *exploratory data analysis*(EDA). Many important methodological contributions to existing techniques in data analysis were initiated by discoveries made via EDA. In addition, EDA can lead to interesting biological discoveries which would otherwise be missed through simply subjecting the data to a battery of hypothesis tests. Through this book, we make use of exploratory plots to motivate the analyses we choose. Here we present a general introduction to EDA using height data.

We have already introduced some EDA approaches for *univariate* data, namely the histograms and qq-plot. Here we describe the qq-plot in more detail and some EDA and summary statistics for paired data. We also give a demonstration of commonly used figures that we recommend against.

**TECHNIQUES:**

# Box plot

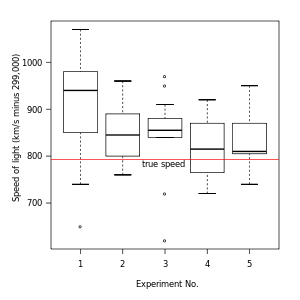
[](https://en.wikipedia.org/wiki/File:Michelsonmorley-boxplot.svg)

Figure 1. Box plot of data from the [Michelson–Morley experiment](https://en.wikipedia.org/wiki/Michelson%E2%80%93Morley_experiment)

a simple way of representing statistical data on a plot in which a rectangle is drawn to represent the second and third quartiles, usually with a vertical line inside to indicate the median value. The lower and upper quartiles are shown as horizontal lines either side of the rectangle.

In [descriptive statistics](https://en.wikipedia.org/wiki/Descriptive_statistics), a **box plot** or **boxplot** is a method for graphically depicting groups of numerical data through their [quartiles](https://en.wikipedia.org/wiki/Quartile). 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** and **box-and-whisker diagram**. [Outliers](https://en.wikipedia.org/wiki/Outlier) may be plotted as individual points. Box plots are [non-parametric](https://en.wikipedia.org/wiki/Non-parametric): they display variation in samples of a [statistical population](https://en.wikipedia.org/wiki/Statistical_population) without making any assumptions of the underlying [statistical distribution](https://en.wikipedia.org/wiki/Probability_distribution). The spacings between the different parts of the box indicate the degree of [dispersion](https://en.wikipedia.org/wiki/Statistical_dispersion) (spread) and [skewness](https://en.wikipedia.org/wiki/Skewness) in the data, and show [outliers](https://en.wikipedia.org/wiki/Outlier). In addition to the points themselves, they allow one to visually estimate various [L-estimators](https://en.wikipedia.org/wiki/L-estimator), notably the [interquartile range](https://en.wikipedia.org/wiki/Interquartile_range), [midhinge](https://en.wikipedia.org/wiki/Midhinge" \o "Midhinge), [range](https://en.wikipedia.org/wiki/Range_(statistics)), [mid-range](https://en.wikipedia.org/wiki/Mid-range), and [trimean](https://en.wikipedia.org/wiki/Trimean" \o "Trimean). Box plots can be drawn either horizontally or vertically. Box plots received their name from the box in the middle.

The box plot is a quick way of examining one or more sets of data graphically. Box plots may seem more primitive than a [histogram](https://en.wikipedia.org/wiki/Histogram) or [kernel density estimate](https://en.wikipedia.org/wiki/Kernel_density_estimation) but they do have some advantages. They take up less space and are therefore particularly useful for comparing distributions between several groups or sets of data (see Figure 1 for an example). Choice of [number and width of bins](https://en.wikipedia.org/wiki/Histogram#Number_of_bins_and_width) techniques can heavily influence the appearance of a histogram, and choice of bandwidth can heavily influence the appearance of a kernel density estimate.

As looking at a statistical distribution is more commonplace than looking at a box plot, comparing the box plot against the probability density function (theoretical histogram) for a normal N(0,1σ2) distribution may be a useful tool for understanding the box plot

#### **Box Plot**

import matplotlib.pyplot as plt

import pandas as pd

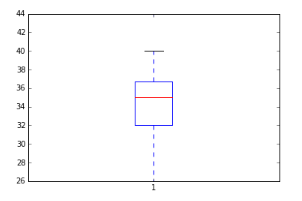
fig=plt.figure()

ax = fig.add\_subplot(1,1,1)

#Variable

ax.boxplot(df['Age'])

plt.show()

[](https://www.analyticsvidhya.com/wp-content/uploads/2015/05/Box_Plot_Violin_Plot1.png)

# Histogram

A **histogram** is an accurate graphical representation of the [distribution](https://en.wikipedia.org/wiki/Frequency_distribution) of numerical data. It is an estimate of the [probability distribution](https://en.wikipedia.org/wiki/Probability_distribution)of a [continuous variable](https://en.wikipedia.org/wiki/Continuous_variable) (quantitative variable) and was first introduced by [Karl Pearson](https://en.wikipedia.org/wiki/Karl_Pearson).[[1]](https://en.wikipedia.org/wiki/Histogram#cite_note-pearson-1) It is a kind of bar graph. To construct a histogram, the first step is to "[bin](https://en.wikipedia.org/wiki/Data_binning)" the range of values—that is, divide the entire range of values into a series of intervals—and then count how many values fall into each interval. The bins are usually specified as consecutive, non-overlapping [intervals](https://en.wikipedia.org/wiki/Interval_(mathematics)) of a variable. The bins (intervals) must be adjacent, and are often (but are not required to be) of equal size.[[2]](https://en.wikipedia.org/wiki/Histogram#cite_note-2)

If the bins are of equal size, a rectangle is erected over the bin with height proportional to the [frequency](https://en.wikipedia.org/wiki/Frequency_(statistics)) — the number of cases in each bin. A histogram may also be [normalized](https://en.wikipedia.org/wiki/Normalization_(statistics)) to display "relative" frequencies. It then shows the proportion of cases that fall into each of several [categories](https://en.wikipedia.org/wiki/Categorization), with the sum of the heights equaling 1.

However, bins need not be of equal width; in that case, the erected rectangle is defined to have its *area* proportional to the frequency of cases in the bin.[[3]](https://en.wikipedia.org/wiki/Histogram#cite_note-3) The vertical axis is then not the frequency but *frequency density* — the number of cases per unit of the variable on the horizontal axis. Examples of variable bin width are displayed on Census bureau data below.

As the adjacent bins leave no gaps, the rectangles of a histogram touch each other to indicate that the original variable is continuous.[[4]](https://en.wikipedia.org/wiki/Histogram#cite_note-4)

Histograms give a rough sense of the density of the underlying distribution of the data, and often for [density estimation](https://en.wikipedia.org/wiki/Density_estimation): estimating the [probability density function](https://en.wikipedia.org/wiki/Probability_density_function) of the underlying variable. The total area of a histogram used for probability density is always normalized to 1. If the length of the intervals on the *x*-axis are all 1, then a histogram is identical to a [relative frequency](https://en.wikipedia.org/wiki/Relative_frequency) plot.

A histogram can be thought of as a simplistic [kernel density estimation](https://en.wikipedia.org/wiki/Kernel_density_estimation), which uses a [kernel](https://en.wikipedia.org/wiki/Kernel_(statistics)) to smooth frequencies over the bins. This yields a [smoother](https://en.wikipedia.org/wiki/Smooth_function) probability density function, which will in general more accurately reflect distribution of the underlying variable. The density estimate could be plotted as an alternative to the histogram, and is usually drawn as a curve rather than a set of boxes.

Another alternative is the average shifted histogram,[[5]](https://en.wikipedia.org/wiki/Histogram" \l "cite_note-5) which is fast to compute and gives a smooth curve estimate of the density without using kernels.

The histogram is one of the [seven basic tools of quality control](https://en.wikipedia.org/wiki/Seven_Basic_Tools_of_Quality).[[6]](https://en.wikipedia.org/wiki/Histogram#cite_note-6)

Histograms are sometimes confused with bar charts. A histogram is used for [continuous data](https://en.wikipedia.org/wiki/Continuous_data), where the bins represent ranges of data, while a [bar chart](https://en.wikipedia.org/wiki/Bar_chart) is a plot of categorical variables. Some authors recommend that bar charts have gaps between the rectangles to clarify the distinction.

#### **Histogram :**

fig=plt.figure() #Plots in matplotlib reside within a figure object, use plt.figure to create new figure

#Create one or more subplots using add\_subplot, because you can't create blank figure

ax = fig.add\_subplot(1,1,1)

#Variable

ax.hist(df['Age'],bins = 7) # Here you can play with number of bins

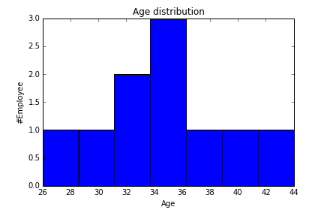
Labels and Tit

plt.title('Age distribution')

plt.xlabel('Age')

plt.ylabel('#Employee')

plt.show()



# Scatter plot

|  |  |
| --- | --- |
|  | |
|  | |
|  |  |
|  |  |

A **scatter plot** (also called a **scatter graph**, **scatter chart**, **scattergram**, or **scatter diagram**)[[3]](https://en.wikipedia.org/wiki/Scatter_plot#cite_note-3) is a type of [plot](https://en.wikipedia.org/wiki/Plot_(graphics)) or [mathematical diagram](https://en.wikipedia.org/wiki/Mathematical_diagram) using [Cartesian coordinates](https://en.wikipedia.org/wiki/Cartesian_coordinate_system) to display values for typically two [variables](https://en.wikipedia.org/wiki/Variable_(mathematics)) for a set of data. If the points are color-coded, one additional variable can be displayed. The data is displayed as a collection of points, each having the value of one variable determining the position on the horizontal axis and the value of the other variable determining the position on the [vertical axis](https://en.wikipedia.org/wiki/Vertical_axis).[[4]](https://en.wikipedia.org/wiki/Scatter_plot#cite_note-4)

A scatter plot can be used either when one continuous variable that is under the control of the experimenter and the other depends on it or when both continuous variables are independent. If a [parameter](https://en.wikipedia.org/wiki/Parameter) exists that is systematically incremented and/or decremented by the other, it is called the *control parameter* or [independent variable](https://en.wikipedia.org/wiki/Independent_variable) and is customarily plotted along the horizontal axis. The measured or [dependent variable](https://en.wikipedia.org/wiki/Dependent_variable) is customarily plotted along the vertical axis. If no dependent variable exists, either type of variable can be plotted on either axis and a scatter plot will illustrate only the degree of [correlation](https://en.wikipedia.org/wiki/Correlation) (not [causation](https://en.wikipedia.org/wiki/Causality)) between two variables.

A scatter plot can suggest various kinds of correlations between variables with a certain [confidence interval](https://en.wikipedia.org/wiki/Confidence_interval). For example, weight and height, weight would be on y axis and height would be on the x axis. Correlations may be positive (rising), negative (falling), or null (uncorrelated). If the pattern of dots slopes from lower left to upper right, it indicates a positive [correlation](https://en.wikipedia.org/wiki/Correlation) between the variables being studied. If the pattern of dots slopes from upper left to lower right, it indicates a negative correlation. A line of [best fit](https://en.wikipedia.org/wiki/Curve_fitting) (alternatively called 'trendline') can be drawn in order to study the relationship between the variables. An equation for the correlation between the variables can be determined by established best-fit procedures. For a linear correlation, the best-fit procedure is known as [linear regression](https://en.wikipedia.org/wiki/Linear_regression) and is guaranteed to generate a correct solution in a finite time. No universal best-fit procedure is guaranteed to generate a correct solution for arbitrary relationships. A scatter plot is also very useful when we wish to see how two comparable data sets agre to show nonlinear relationships between variables. The ability to do this can be enhanced by adding a smooth line such as [LOESS](https://en.wikipedia.org/wiki/Local_regression).[[5]](https://en.wikipedia.org/wiki/Scatter_plot#cite_note-5) Furthermore, if the data are represented by a mixture model of simple relationships, these relationships will be visually evident as superimposed patterns.

The scatter diagram is one of the [seven basic tools](https://en.wikipedia.org/wiki/Seven_Basic_Tools_of_Quality) of [quality control](https://en.wikipedia.org/wiki/Quality_control).[[6]](https://en.wikipedia.org/wiki/Scatter_plot#cite_note-6)

Scatter charts can be built in the form of [bubble](https://en.wikipedia.org/wiki/Bubble_chart), marker, or/and [line charts](https://en.wikipedia.org/wiki/Line_chart).[[7]](https://en.wikipedia.org/wiki/Scatter_plot#cite_note-7)

**VIOLIN PLOT**

A **violin plot** is a method of plotting numeric data. It is similar to [box plot](https://en.wikipedia.org/wiki/Box_plot) with a rotated [kernel density plot](https://en.wikipedia.org/wiki/Kernel_density_estimation) on each side.[[1]](https://en.wikipedia.org/wiki/Violin_plot#cite_note-1)

The violin plot is similar to [box plots](https://en.wikipedia.org/wiki/Box_plot), except that they also show the [probability density](https://en.wikipedia.org/wiki/Probability_density_function) of the data at different values (in the simplest case this could be a [histogram](https://en.wikipedia.org/wiki/Histogram)). Typically violin plots will include a marker for the median of the data and a box indicating the interquartile range, as in standard box plots. Overlaid on this box plot is a [kernel density estimation](https://en.wikipedia.org/wiki/Kernel_density_estimation). Like box plots, violin plots are used to represent comparison of a variable distribution (or sample distribution) across different "categories". For example temperature distribution compared between day and night or distribution of car prices compared across different car makers.

A violin plot is more informative than a plain box plot. In fact while a box plot only shows summary statistics such as mean/median and interquartile ranges, the violin plot shows the full distribution of the data. The difference is particularly useful when the data distribution is multimodal (more than one peak). In this case a violin plot clearly shows the presence of different peaks, their position and relative amplitude. This information could not be represented with a simple box plot which only reports summary statistics. The inner part of a violin plot usually shows the mean (or median) and the interquartile range. In other cases, when the number of samples is not too high, the inner part can show all sample points (with a dot or a line for each sample).

Although more informative than box plots, a disadvantage of violin plots is that they are less popular. Because of their unpopularity, their meaning can be harder to grasp for many readers not familiar with the violin plot representation. In this case, a more accessible alternative can be plotting a series of stacked histograms or [Kernel density distributions](https://en.wikipedia.org/wiki/Kernel_density_estimation).

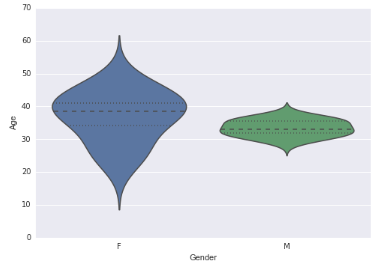
Violin plots are available as extensions to a number of software packages, including the [R](https://en.wikipedia.org/wiki/R_(programming_language)) libraries vioplot, wvioplot, caroline, UsingR, lattice and ggplot2, the [Stata](https://en.wikipedia.org/wiki/Stata) add-on command vioplot,[[2]](https://en.wikipedia.org/wiki/Violin_plot#cite_note-2) and the [Python](https://en.wikipedia.org/wiki/Python_(programming_language)) libraries [matplotlib](https://en.wikipedia.org/wiki/Matplotlib" \o "Matplotlib)[[3]](https://en.wikipedia.org/wiki/Violin_plot#cite_note-3) and Seaborn.[[4]](https://en.wikipedia.org/wiki/Violin_plot#cite_note-4)

#### **Violin Plot**

import seaborn as sns

sns.violinplot(df['Age'], df['Gender']) #Variable Plot

sns.despine()



**BUBBLE PLOT**

A **bubble chart** is a type of [chart](https://en.wikipedia.org/wiki/Chart) that displays three dimensions of data. Each entity with its triplet (*v*1, *v*2, *v*3) of associated data is plotted as a disk that expresses two of the *vi* values through the disk's *xy* location and the third through its size. Bubble charts can facilitate the understanding of social, economical, medical, and other scientific relationships.

Bubble charts can be considered a variation of the [scatter plot](https://en.wikipedia.org/wiki/Scatter_plot), in which the data points are replaced with bubbles. As the documentation for [Microsoft Office](https://en.wikipedia.org/wiki/Microsoft_Office) explains, "You can use a bubble chart instead of a scatter chart if your data has three data series that each contain a set of values. The sizes of the bubbles are determined by the values in the third data series.".[[1]](https://en.wikipedia.org/wiki/Bubble_chart" \l "cite_note-1)

## **Choosing bubble sizes correctly[[edit](https://en.wikipedia.org/w/index.php?title=Bubble_chart&action=edit&section=1" \o "Edit section: Choosing bubble sizes correctly)]**

The [human visual system](https://en.wikipedia.org/wiki/Human_visual_system) naturally experiences a disk's size in terms of its [area](https://en.wikipedia.org/wiki/Area). And the area of a disk—unlike its [diameter](https://en.wikipedia.org/wiki/Diameter) or [circumference](https://en.wikipedia.org/wiki/Circumference)—is not proportional to its [radius](https://en.wikipedia.org/wiki/Radius), but to the square of the radius. So if one chooses to scale the disks' radii to the third data values directly, then the apparent size differences among the disks will be non-linear and misleading. To get a properly weighted scale, one must scale each disk's radius to the [square root](https://en.wikipedia.org/wiki/Square_root) of the corresponding data value *v*3 This scaling issue can lead to extreme misinterpretations, especially where the range of the data has a large spread. And because many people are unfamiliar with—or do not stop to consider—the issue and its impact on perception, those who are aware of it often have to hesitate in interpreting a bubble chart because they cannot assume that the scaling correction was indeed made. So it is important that bubble charts not only be scaled in this way, but also be clearly labeled to document that it is area, rather than radius or diameter, that conveys the data.[[2]](https://en.wikipedia.org/wiki/Bubble_chart#cite_note-2)

#### **Bubble Plot**

fig = plt.figure()

ax = fig.add\_subplot(1,1,1)

ax.scatter(df['Age'],df['Sales'], s=df['Income']) # Added third variable income as size of the bubble

plt.show()

