Stats Notes

***arithmetic mean*** or ***mean*** (also called ***average***), computed as the data values' sum divided by the number of values. Ex: The mean of 1, 6, 7, 6 is (1 + 6 + 7 + 6) / 4 = 20 / 4 = 5. The mean can be thought of graphically as the "center" of the data on a number line.

An ***outlier*** is a data value that is either much greater than or much less than the rest of the data and not representative of the rest of the data being considered. Ex: The net worth in dollars of 5 particular people in Medina, Washington in 2015 was: 0.3 million, 0.4 million, 0.25 million, 80 billion, and 0.6 million. That 80 billion is due to Bill Gates (Microsoft co-founder) living in Medina. The resulting mean of 16 million is a poor data summary, potentially misleading a viewer into believing all 5 people are wealthy multi-millionaires.

Thus, another common data summary is the ***median***, which is the middle value in the sorted list of data values. Ex: Given unsorted data values 100, 3, 6, 9, 2, then the sorted values are 2, 3, 6, 9, 100, and the median is 6. Note that the mean is 120 / 5 = 24, being influenced by the outlier of 100.

If the number of data values N is:

* Odd: The **median** is item (N + 1) / 2 in the sorted value list. Ex: Given sorted data values 10, 20, 20, 30, 60, 60, 80, then N is 7, so (7 + 1) / 2 = 8 / 2 = 4. Item 4 in the list is 30, which is thus the median.
* Even: No single middle item exists, so the median is computed as the mean of items at N / 2 and (N / 2) + 1 (the middle two items). Ex: Given sorted data 10, 20, 20, 30, 60, 60, 80, 99, then N is 8, so N / 2 is 4 and (N / 2) + 1 is 5. Item 4 is 30 and item 5 is 60, yielding a median of (30 + 60) / 2 = 45.

***Mode*** is another measure of center, being the most frequently-occurring value. Ex: The mode of 1, 4, 4, 8, 8, 8, 9 is 8, because 8 appears more times (3 times) than any other value. Mode is much less commonly used than mean and median.

# 1.2 Measures of spread

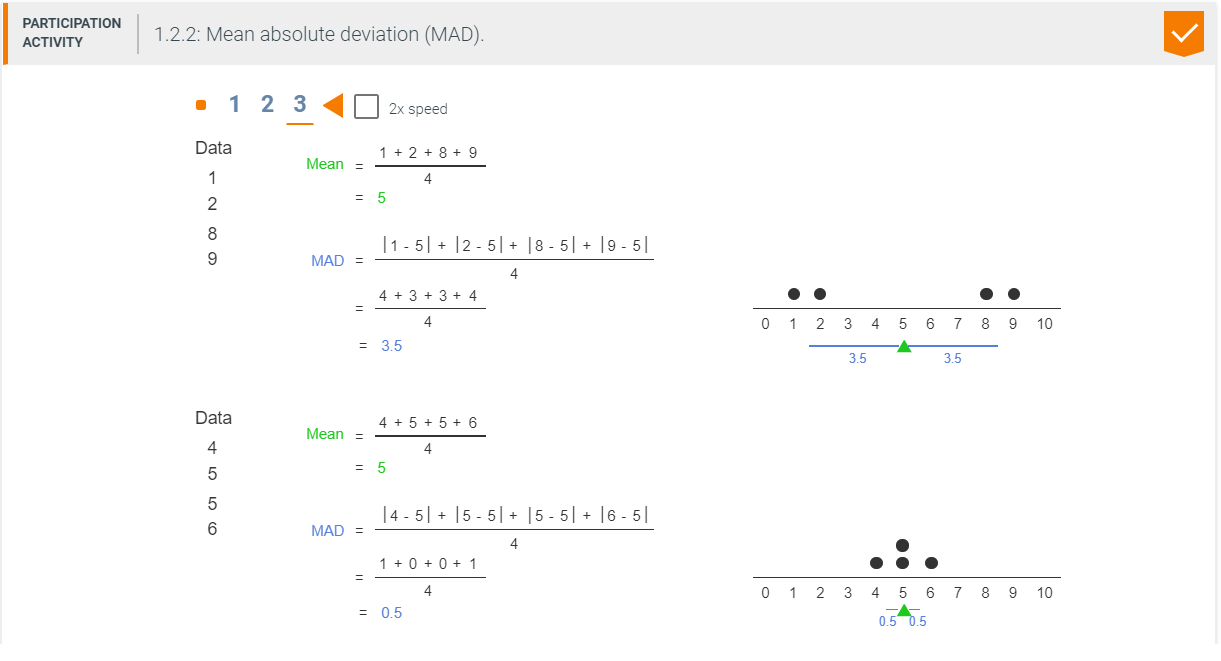
### **Minimum, maximum, range**

A measure of center alone does not indicate the extent of the spread of the data values. Ex: Data values 1, 2, 8, 9, and data values 4, 5, 5, 6 each have a mean (and median) of 5, but clearly 1, 2, 8, 9 are more spread out. Some indication of spread can help a person interpret a measure of center.

A simple measure of spread is to report the ***minimum*** and ***maximum*** values, typically along with a measure of center like mean or median. Ex: 1, 2, 8, 9 has mean = 5, min = 1, max = 9, whereas 4, 5, 5, 6 has mean = 5, min = 4, max = 6. Another simple measure of spread is ***range***, which reports the difference between the maximum and minimum values. Ex: 1, 2, 8, 9 has a range of 8, whereas 4, 5, 5, 6 has a range of 2.

### **Mean absolute deviation**

Another method of spread seeks to summarize the "average distance from the average." Because only the distance is relevant and not the direction, the absolute value (magnitude) of the difference is used to avoid negative distances. ***Mean absolute deviation*** (***MAD***) is the mean of the absolute difference between each value and the values' mean.



A measure of spread that uses squaring rather than absolute value is called variance. ***Variance*** is



, where di is a data value and n is the number of data values. The numerator computes the square of each difference and sums those values. To compute the average of those squared differences, one would expect to divide by n; however, variance divides by n-1 instead, for reasons beyond this material's scope.

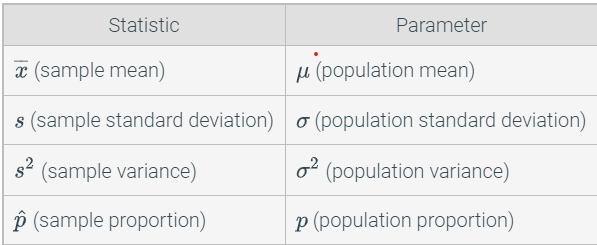
***Standard deviation*** is the square root of the variance. Standard deviation uses squaring (done to compute the variance) to eliminate negatives, but then uses the square root to arrive at a value that is similar to the mean average deviation and thus is more intuitive.

# 2.1 Parameterized population models

### **Parametric analysis**

A *parameter* describes characteristics of a population. Because population parameters are almost always unknown, sampling is performed to obtain a parameter's estimator. Given a sample, mathematical formulas are used to obtain statistics. A *statistic* is a characteristic of a sample. The idea behind statistical inference is to use a statistic to test a proposition about a population parameter. Ex: Based on the data obtained from the sample and the calculated sample mean, is the hypothesized mean the same as the population mean?

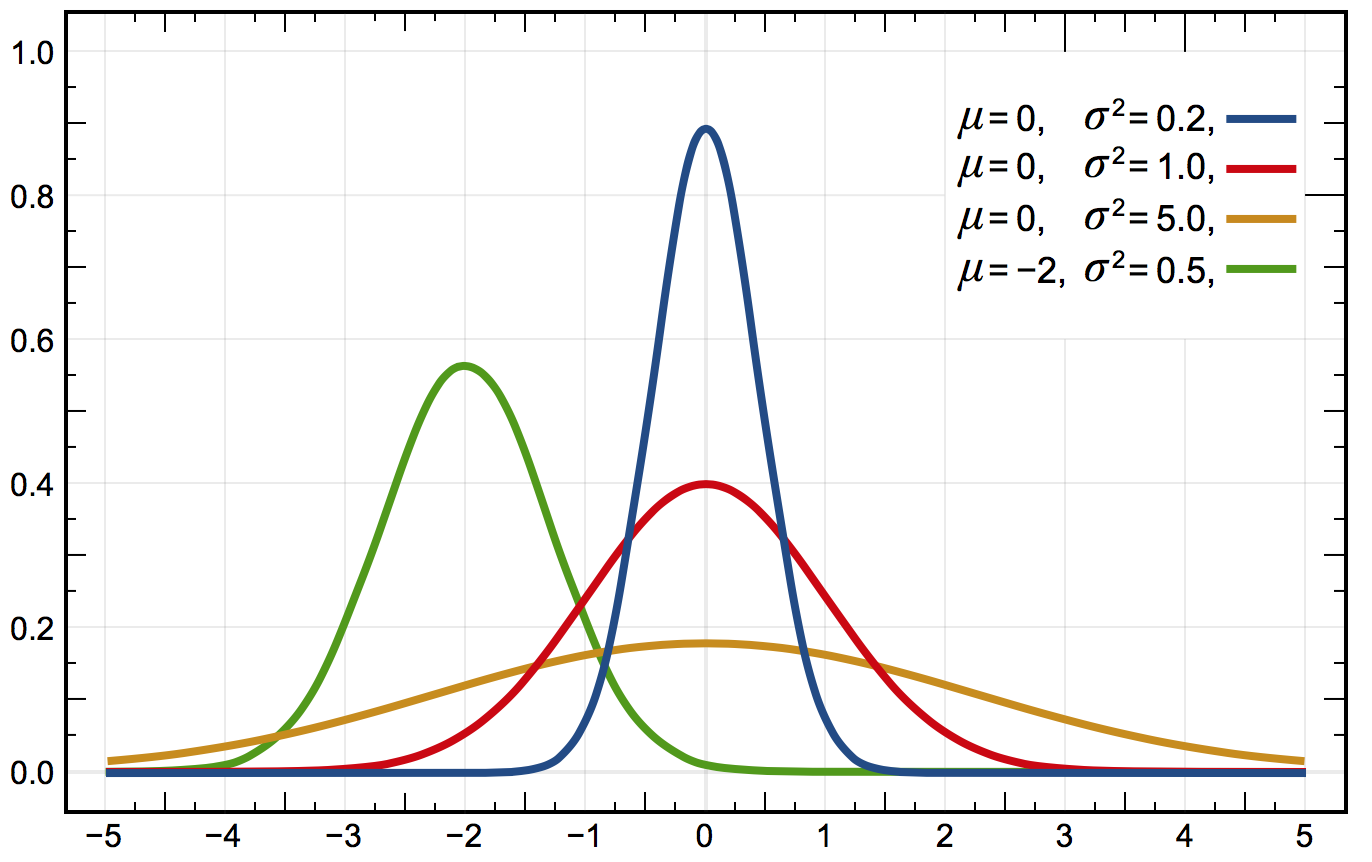
The table below lists common statistics and parameters.



***Parametric analysis*** is a classification of statistical inference tools based on assumptions about the distribution of the population and a fixed set of parameters. Sample means of random variables from non-normal distributions follow a normal distribution as the number of sample means collected increases. Thus, statisticians are able to make assumptions in many situations that aid in statistical analysis. Often statistics follow a normal distribution. Since probabilities involving normal distributions are easy to calculate, the probability of a statistic having a particular value given a hypothesized value for the parameter can easily be found. Ex: Although human height is shown to follow a normal distribution, human weight does not. If samples are repeatedly taken, does a way to show that mean weights differ between genders exist? Even without knowing the distribution of weight, parametric analysis provides the tools to answer such question. However, caution should be taken that certain assumptions are satisfied when using parametric statistical procedures, especially when sample sizes are not sufficiently large.

The following are parametric tests:

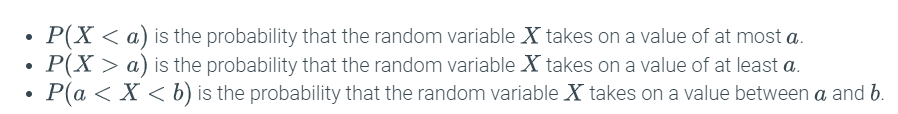
* One-way t-test - determines whether the hypothesized mean is equal to the true population mean.
* Paired t-test - used to compare two population means in the case where the two samples are dependent.
* Two-way t-test - determines whether the population means are equal.
* Analysis of variance (ANOVA) - determines whether the means of two or more populations are equal.



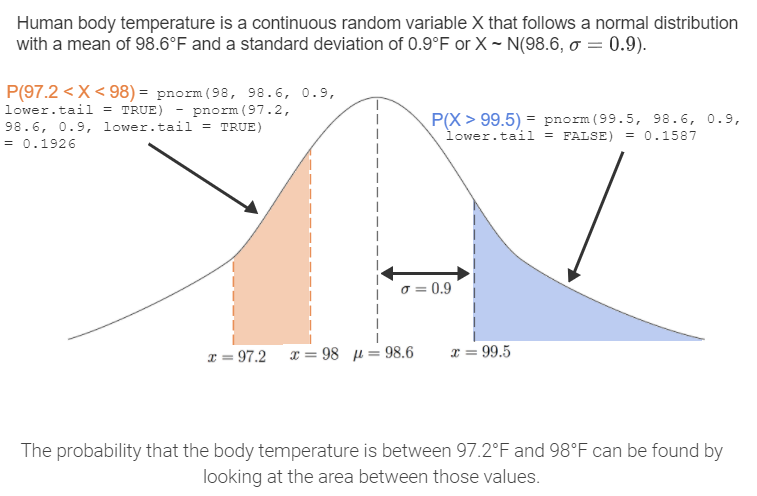
### **Finding probabilities associated with a normal distribution**

Most situations involve random variables from distributions that are continuous. A ***continuous random variable*** is a random variable that can take on any value within a range of infinitely many or uncountable values. Ex: Random variables that represent measurements with continuous values such as body mass index and radius of an O-ring joint are continuous.

The following notation will be used when working with normally distributed random variables and other random variables following continuous distributions:

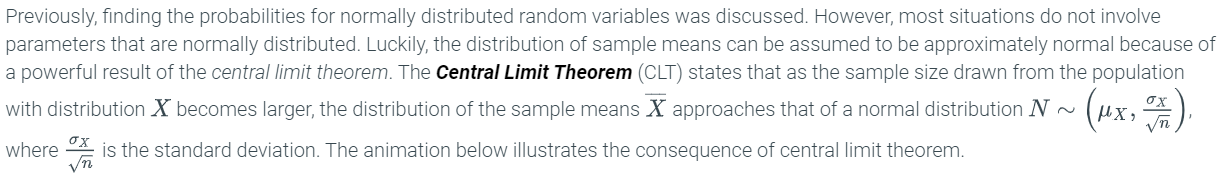


Since the normal distribution is continuous, looking at the frequency distribution to calculate probabilities would not be helpful. The area under a portion of the normal distribution gives the probability. Numerical and integral methods to find the area under the curve of a normal distribution are introduced in a calculus or advanced statistics course, but will not be covered in this material. The R-Practice below shows how to calculate different probabilities of a random variable following a normal distribution using built-in functions.

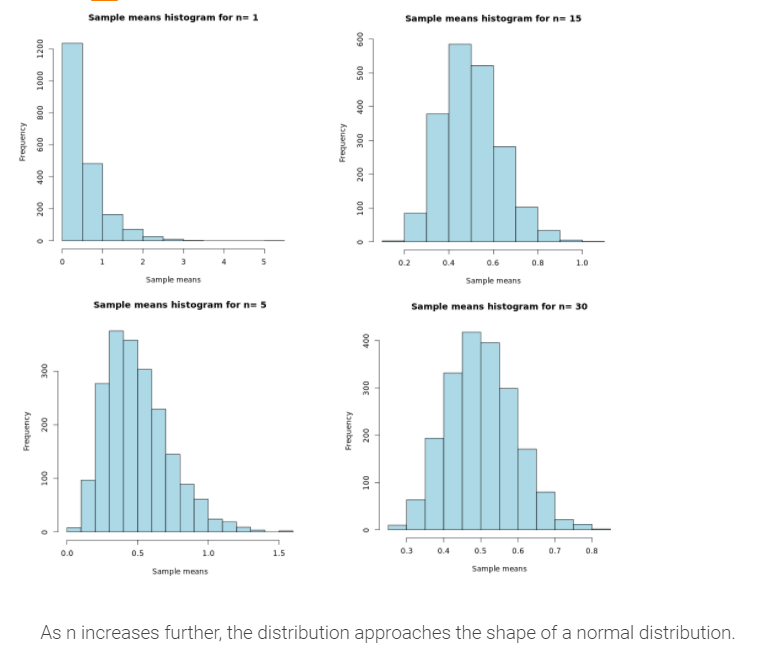


### **Central limit theorem for means**

Often, statisticians look at the distribution of a test statistic, such as the sample mean. Suppose a sample of size n is taken from a population and the sample mean is computed. Repeating this process for multiple samples, and creating a relative frequency plot of the obtained test statistic, yields a sampling distribution. The ***sampling distribution of the mean*** is the distribution of sample means when taking random samples of the same size. As seen earlier, finding the probability associated with a value of a random variable is often of interest. Ex: What is the probability that the mean heights of the starting line-ups of the last 1000 NBA games is greater than 6 foot 5 inches?



Previously, finding the probabilities for normally distributed random variables was discussed. However, most situations do not involve parameters that are normally distributed. Luckily, the distribution of sample means can be assumed to be approximately normal because of a powerful result of the *central limit theorem*. The ***Central Limit Theorem*** (CLT)



As shown in the animation above, the sample size affects the shape of the sampling distribution. For the central limit theorem to be used, assumptions and conditions in addition to sample size must be satisfied.

Central limit theorem assumptions and conditions

* Randomness assumption - samples must be randomly selected.
* Independence condition - sample values must be independent from each other.
* Sample size assumption - sample size must be large enough. A rule of thumb is that the sample size should be at least 30.
* 10% condition - sample size must be at most 10% of the population size.

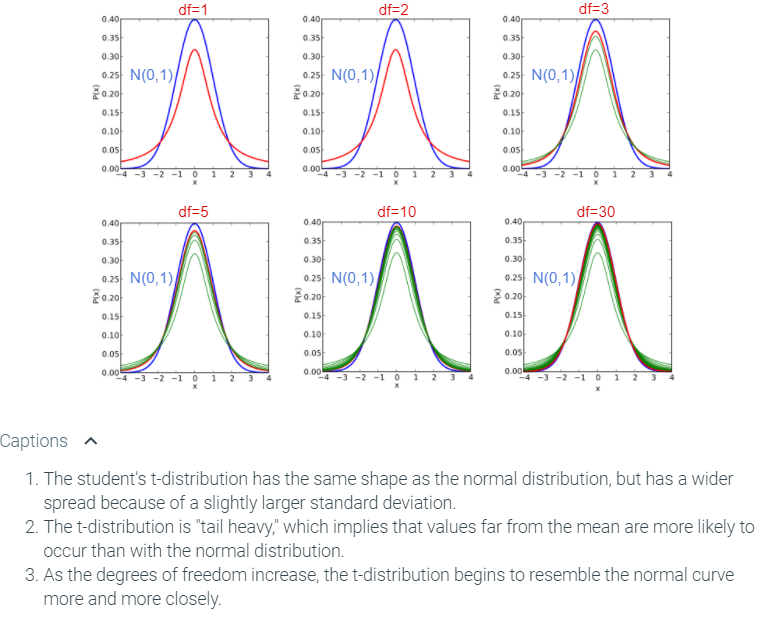
# 2.2 Student's t-test

### **Student's t-distribution**

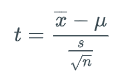
The central limit theorem provides a useful tool to calculate probabilities associated with non-normal distributions assuming the sample sizes are sufficiently large. In practice, obtaining a large enough sample may not be possible or the population standard deviation is unknown. In both cases, the sample standard deviation divided by the square of the sample size can be used in place of the population standard deviation.

The ***student's t-distribution*** or ***t-distribution*** is used in place of the normal distribution in situations where the sample size (n) is small or the population standard deviation (σ) is unknown.

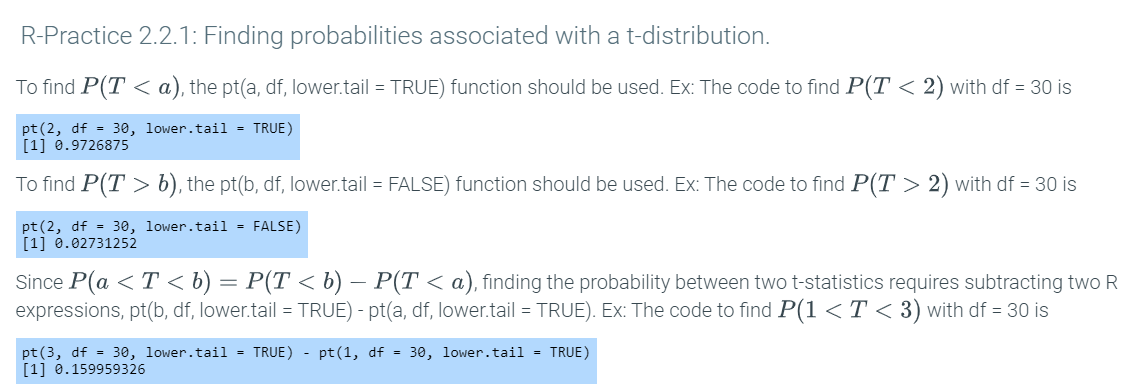
Clearly, as the sample size n increases, df increases as well. As df increases, the t-distribution approaches the normal distribution with mean of 0 and standard deviation of 1 as shown in the animation below.



The ***t-statistic*** is obtained from a sample assumed to have a t-distribution and involves the population mean and a larger variability from estimating the population standard deviation. The same process applies to the computation of probabilities involving t-distribution as shown in an earlier section with normal distribution. The formula for t-statistic is



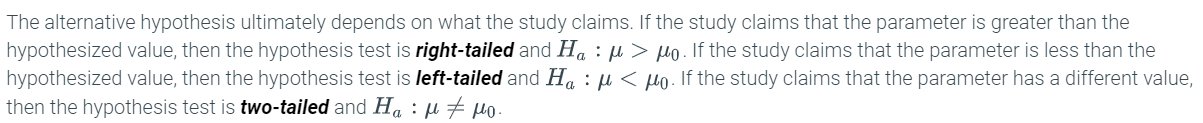
where s is the sample standard deviation, x¯ is the sample mean, μ is the population mean, and n is the sample size.



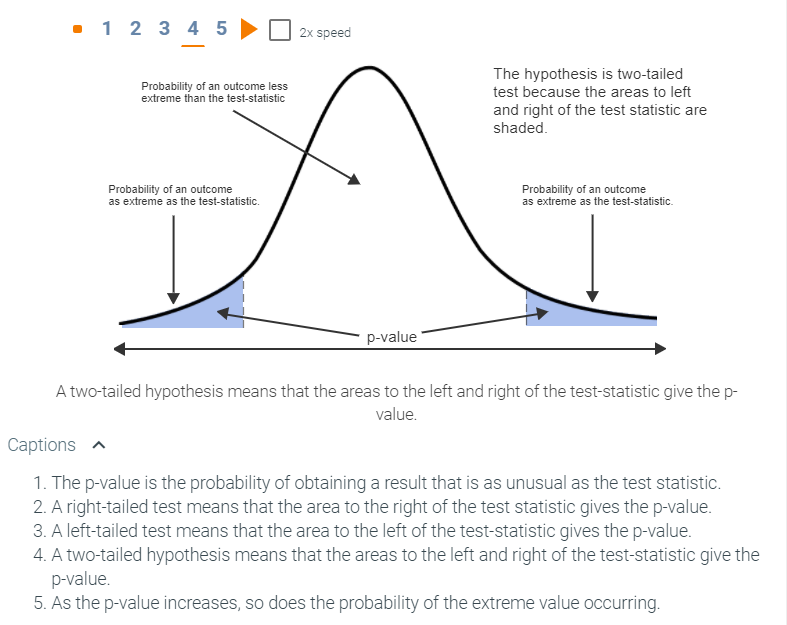
### **Student's t-test**

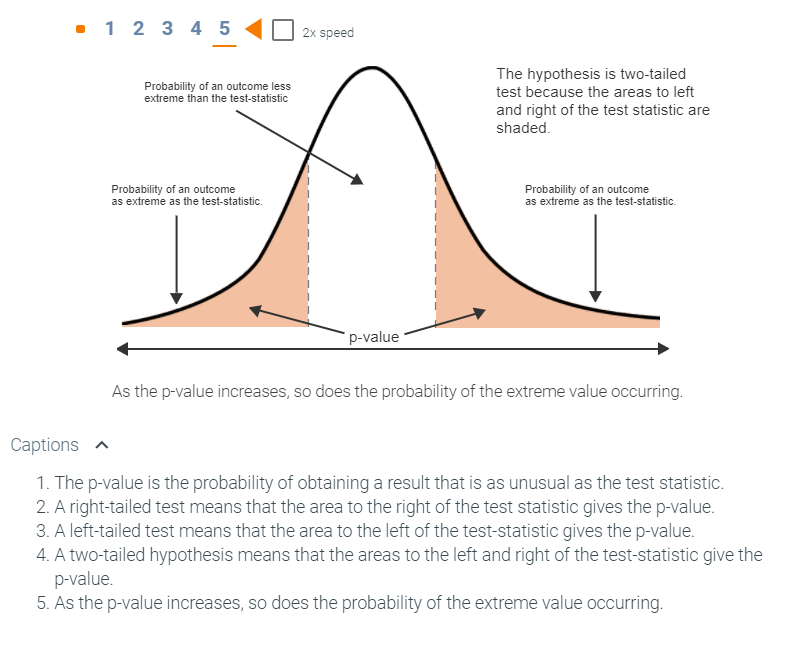
The ***student's t-test*** or ***t-test*** is a method for hypothesis testing for the mean of a sample taken from a normally distributed population when the population standard deviation is unknown. Ex: A question of interest is whether a group of households somewhere in the United States has a higher mean number of children than the rest of the households in the country.

How do statisticians answer such a question? First, statisticians frame the question in terms of hypotheses. The ***null hypothesis*** is the hypothesis of no difference. The ***alternative hypothesis*** is the claim contrary to the null hypothesis. Ex: Suppose a study claims that households in rural communities have more children than the rest of the households in the United States. The null hypothesis would be that the mean number of children from n households are the same as that of the rest of the country. The alternative hypothesis would be that the mean is actually higher.

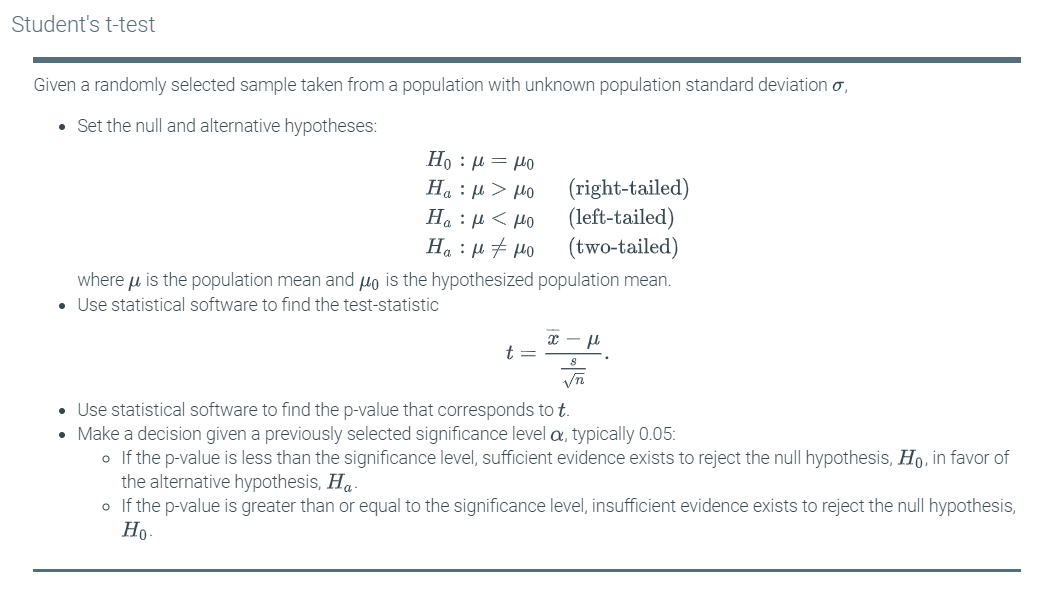


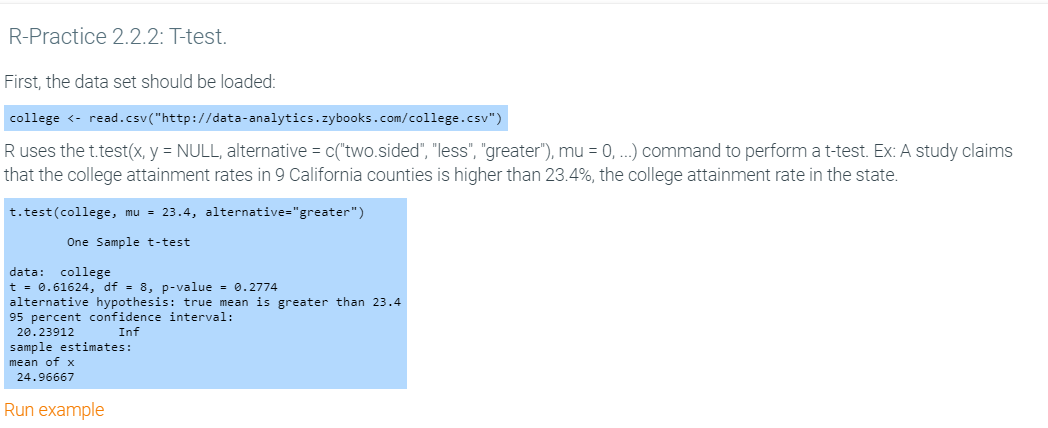
After forming the null and alternative hypotheses, sampling and data collection need to be performed. The main purpose of the t-test is to determine if the difference between the sample mean and the hypothesized population mean is statistically significant. After all, the results may just be due to chance or unique to the individuals sampled. Naturally, if the sample mean is so far from the hypothesized mean, then evidence exists that the true population mean is not what was hypothesized. Assuming the null hypothesis is true, the ***p-value*** is the probability of obtaining a result that is as unusual as the observed test statistic. The calculation of the p-value depends on the type of test (left-tailed, right-tailed, or two-tailed), as discussed in the animation below.

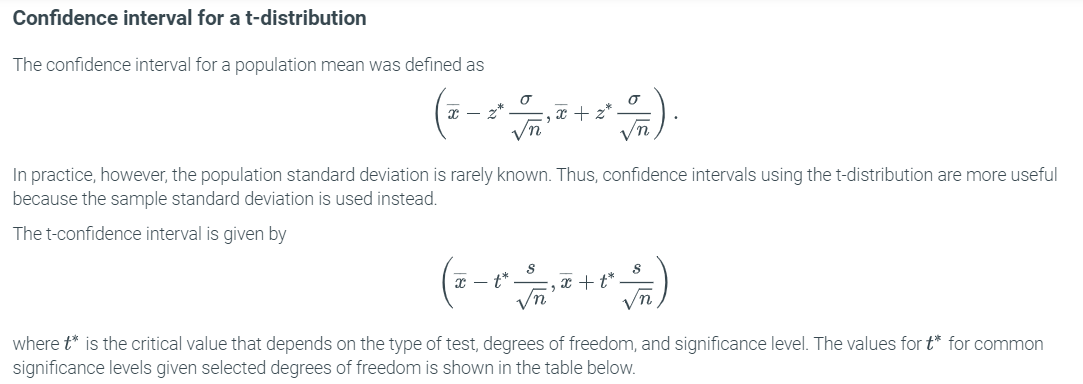


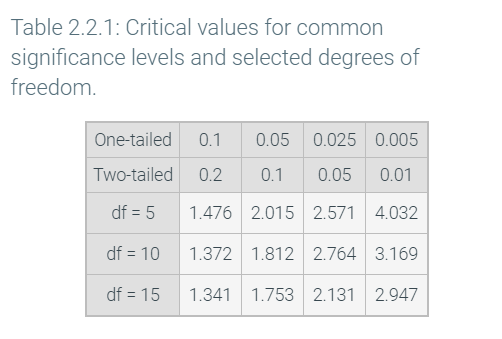


Determining how unusual is a test statistic is rather subjective. The ***significance level***, denoted by α, is the probability of rejecting the null hypothesis when the null hypothesis is actually true. Thus, if the p-value is less than the significance level, sufficient evidence exists to reject the null hypothesis. Common values for α are 0.05 and 0.01, but these values are more of a guideline than a rule and depend on the context of the study.









# 2.3 Comparing 2 samples: 2-sample t-test

### **Two-sample t-test**

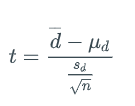
The t-test discussed analyzes the difference between the sample mean and the hypothesized value of the population mean. A similar method exists to compare the means of two different populations. The ***two-sample t-test*** is used to determine if a statistically significant difference exists between two population means. Two types of two-sample t-test exist: paired and unpaired.

In a ***paired t-test*** or ***dependent t-test***, a sample taken from one population is exposed to two different treatments. The main idea is that measurements are recorded from the same group. Ex: A group of professional cycling athletes is selected for a study on the effects of caffeine dosage on exhaustion times. The populations are the cyclists for each of two dosages. The samples are the measured exhaustion times for each dosage, which implies dependence because the measurements were taken from the same group.

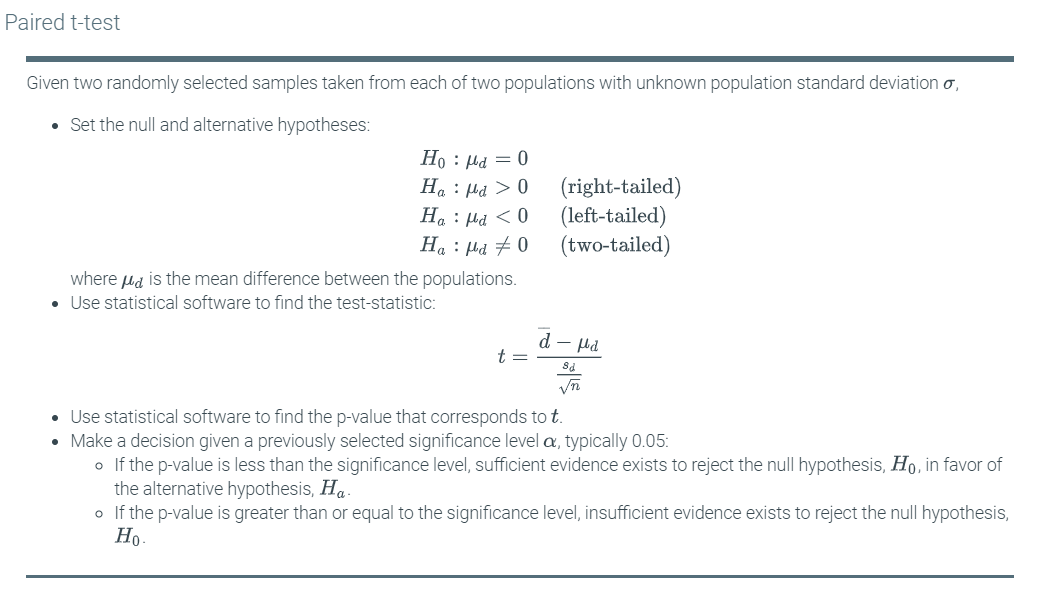
In an ***unpaired t-test*** or ***independent t-test***, a sample taken from one population is not related to a different sample taken from another population. In contrast to the paired t-test, measurements from an unpaired t-test are recorded from different groups when exposed to the same treatment. Ex: The effect of caffeine intake on exhaustion times is studied by measuring the exhaustion times of a randomly selected group of 9 professional cyclists taking caffeine pills and another group of 9 cyclists not taking caffeine pills. The two populations are all cyclists taking caffeine pills and those who are not taking the pills. The samples are the measured exhaustion times from the two groups, each with 9 cyclists, which implies independence because the times are for two different groups of cyclists.

### **Paired t-test**

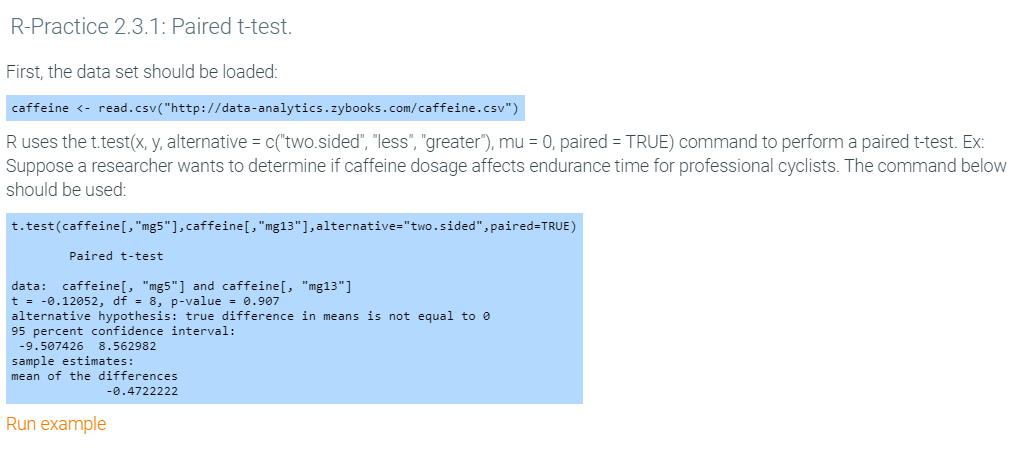
To obtain probabilities for a paired t-test, the paired t-statistic is needed. The formula involves finding the mean and standard deviation of the differences between corresponding measurements:



where sd is the sample standard deviation of the differences, d¯ is the mean difference between the samples, and n is the sample size. The most common scenario is that the hypothesized mean difference is 0. However, this need not be the case. Ex: To continue the development of a new drug, a measurable improvement in the condition of the subjects must be seen. In this situation, the null hypothesis would be that the mean difference is the minimum amount of improvement set by the manufacturer in order to continue developing the drug. The differences are assumed to come from a normal distribution. Thus, the differences can be seen as a single sample following a t-distribution, which means that a paired t-test is equivalent to a one-sample t-test.



The R-Practice below uses data from the "The Effect of Different Dosages of Caffeine on Endurance Performance Time", *International Journal of Sports Medicine* ([Source](http://www.ncbi.nlm.nih.gov/pubmed/7657415)). The data set gives the endurance times (in minutes) of 9 athletes when given a caffeine dose of 5 milligrams and 13 milligrams.

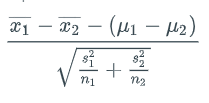


### 

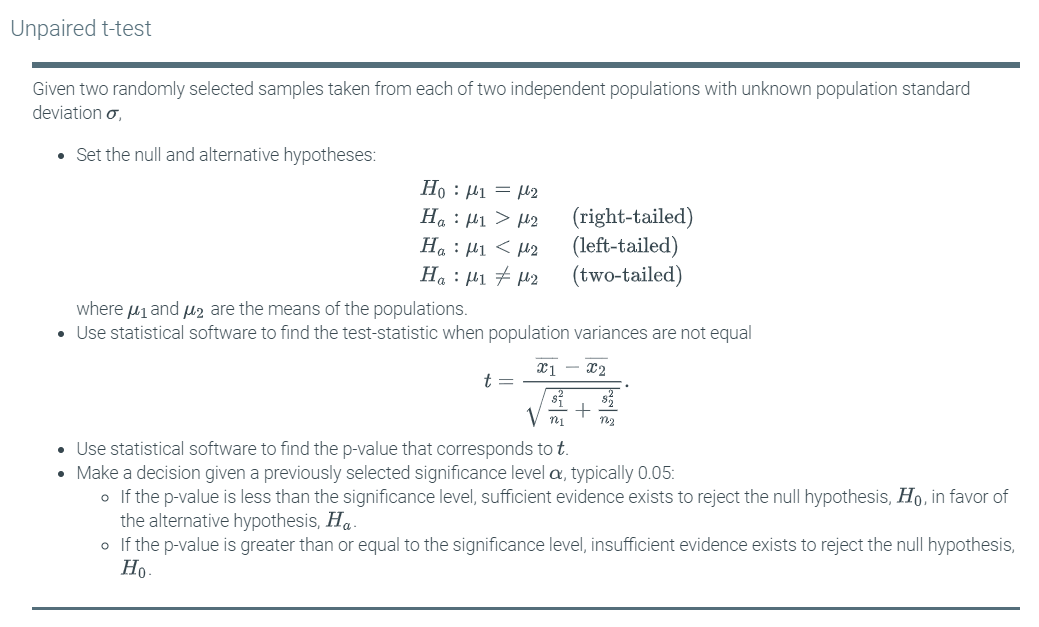
### 

### **Unpaired t-test**

The t-test statistic for unpaired data is different from that of paired data. The formula involves subtracting the means of the two samples:



where x1¯, s1, and n1 are the mean, standard deviation, and sample size of the sample drawn from the first population respectively; and x2¯, s2, and n2 are the mean, standard deviation, and sample size of the sample drawn from the second population. Since measurements are not subtracted, the sample sizes are not necessarily the same. The degrees of freedom are df=n1+n2−2. Although μ1−μ2 can be any number based on the hypothesized means for the two populations, most of the time, the accepted difference between the means of the populations is 0. Finally, the formula for the t-statistic above assumes that the variances are unequal. In most practical instances, the homogeneity of variances should be verified using Fisher's F-test before performing the unpaired t-test. However, this is beyond the scope of the material.



The data set in the R-Practice below contains the number of errors made while completing a memory-related task by a group taking a fictional memory enhancement drug and by another group not taking the drug.



# 2.4 Comparing 3+ samples: ANOVA

### **Comparing three or more populations**

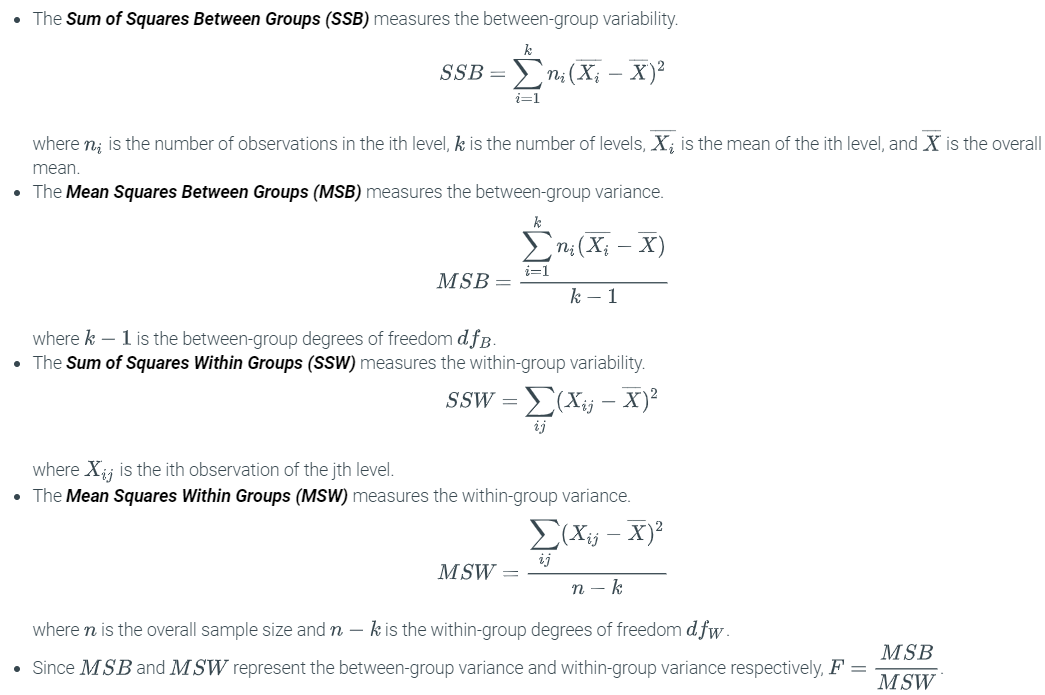
The unpaired t-test determines whether a statistically significant difference exists between the means of two populations. However, many situations require that means be compared among three or more populations. Ex: A researcher wants to classify iris species based on sepal length by using a method called k-means clustering. As a first step, the researcher checks whether the mean sepal length differs among three species of iris: setosa, virginica, and versicolor. A possible method to compare the means is to perform three unpaired t-tests: one between setosa and versicolor, another between setosa and virginica, and finally between versicolor and virginica. Although the details are beyond the scope of the material, the probability of rejecting the null hypothesis that no significant difference in population means exists, when using multiple t-tests is 14%. Thus, a different approach is needed.

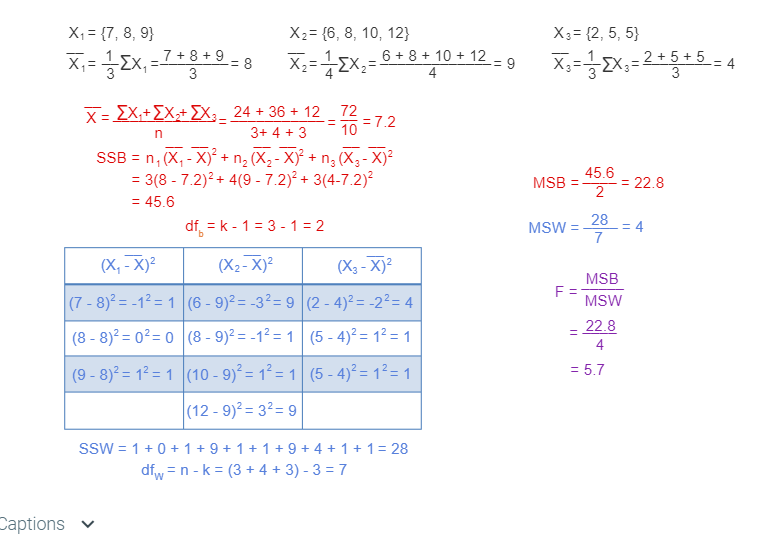
*Analysis of variance (ANOVA)* controls for the errors associated with comparing multiple population means. ***Analysis of variance (ANOVA)*** determines whether a statistically significant difference exists among the means of three or more populations. Equivalently, ANOVA tests for an association between a categorical predictor variable and a response variable. Ex: In the iris study, the predictor variable is the type of species and the response variable is sepal length. Data scientists and statisticians often refer to a categorical predictor variable as a *factor* and a possible value of a factor as a *level*. A factor can be a continuous variable partitioned into intervals commonly referred to as *bins*. Ex: The factor in the iris example, iris type, has three levels: setosa, virginica, and versicolor. If the factor only has two levels, then ANOVA is equivalent to a two-sample t-test with equal variances.

### **The F-statistic**

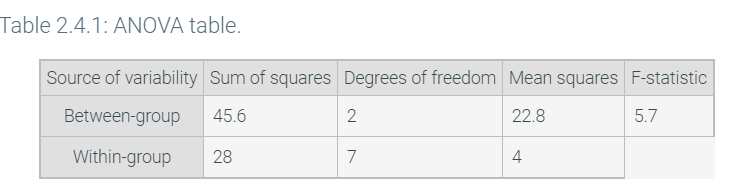
In an earlier section, spread was defined as a measure of variability in a population, usually measured in terms of variance. In a sample with multiple groups of categorical variables, variability can be partitioned into between-group variability and within-group variability. The ***F-statistic*** is the ratio of between-group variance to within-group variance.

A number of quantities are involved when calculating the F-statistic:





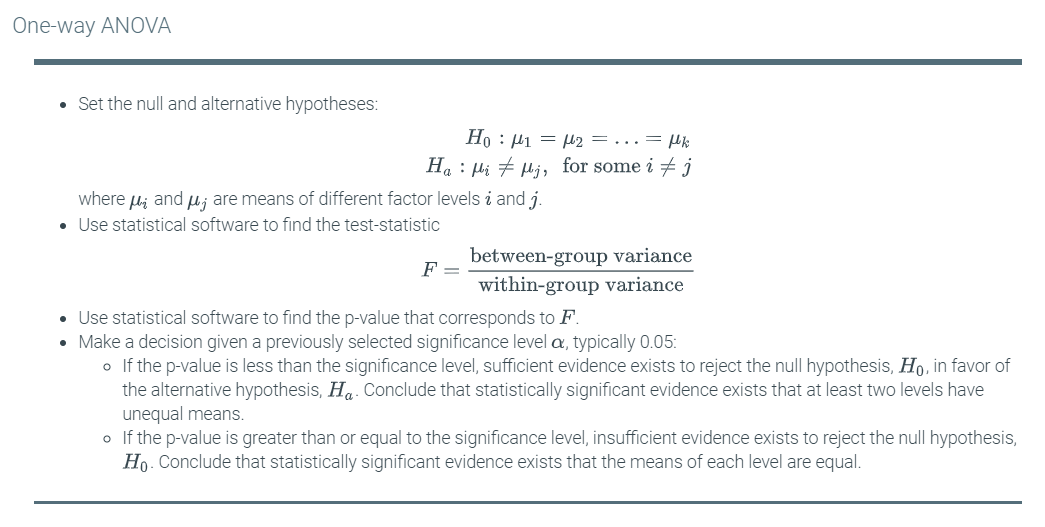
An ***ANOVA table*** displays the statistics used to perform hypothesis testing involving population means. The ANOVA table below summarizes the quantities given in the animation above.



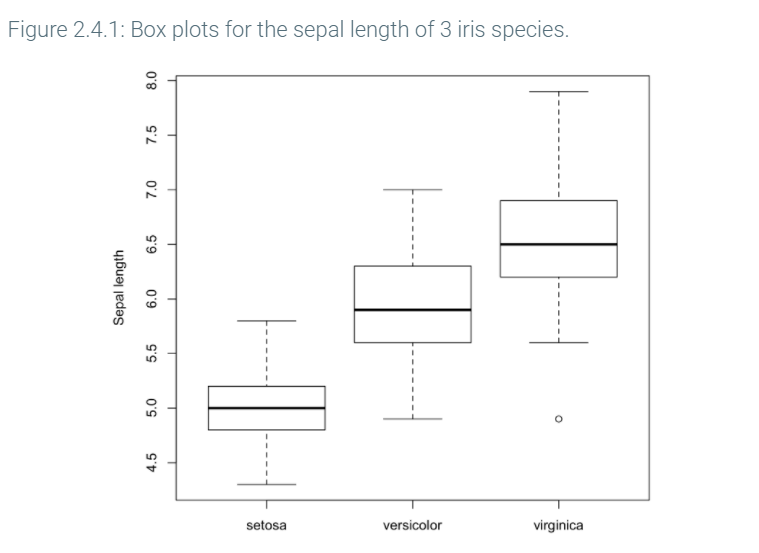
### **One-way ANOVA**

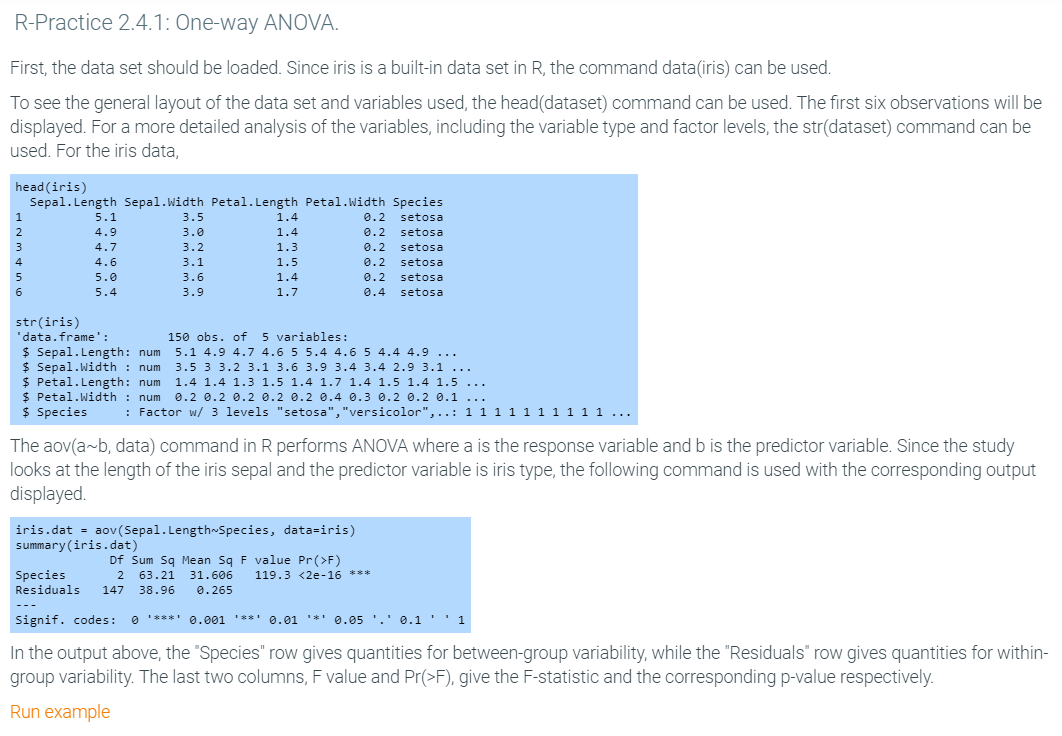
A model can certainly have multiple predictor variables. However, the material on ANOVA will only cover models with one predictor variable. A ***one-way ANOVA*** compares the means of three or more groups of one predictor variable.

The one-way ANOVA hypothesis test follows the same process as previously discussed tests. The null hypothesis for a one-way ANOVA is that all of the group means are equal. Caution should be exercised when stating the alternative hypothesis because the negation of the null hypothesis does not say that all group means are unequal. Instead, the alternative hypothesis should state that two groups with unequal means exist. The rest of the hypothesis test involves finding the F-statistic and the p-value to make a decision based on a significance level.



Returning to the iris study, the researcher wanted to check the possibility of running a classification algorithm based on iris sepal lengths. Among other things the researcher checks is whether the means of the sepal lengths for each of the factor levels are actually different. From the box plots shown below, the sepal length would appear to be different. To be certain, a one-way ANOVA test can be performed.



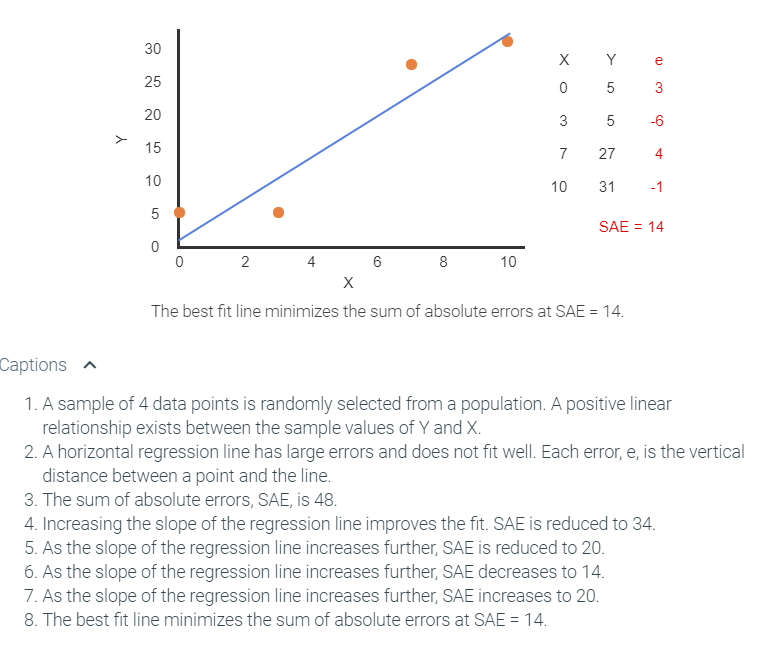


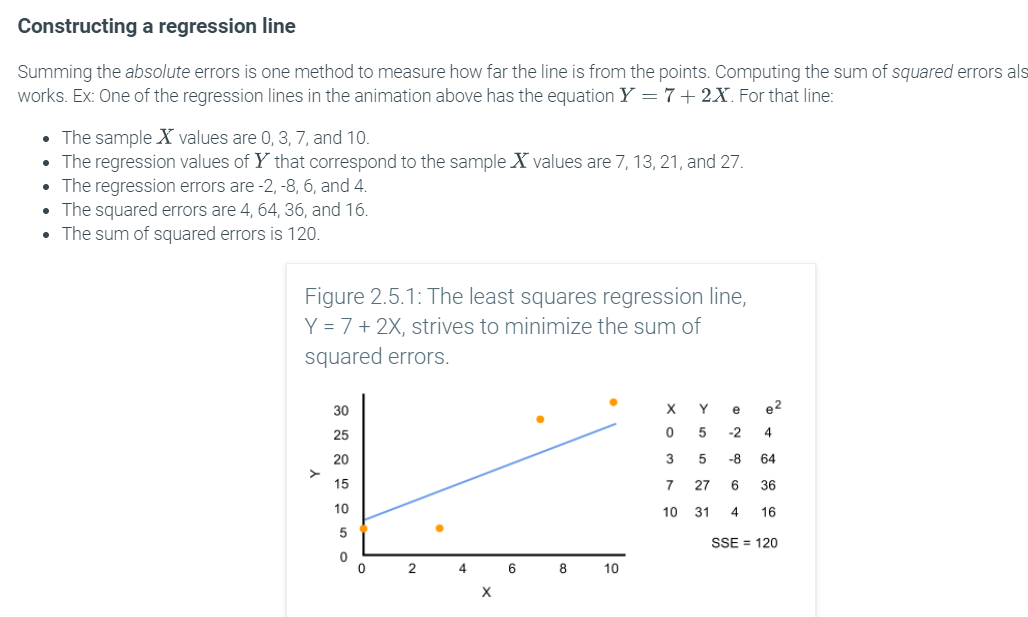
# 2.5 Linear regression

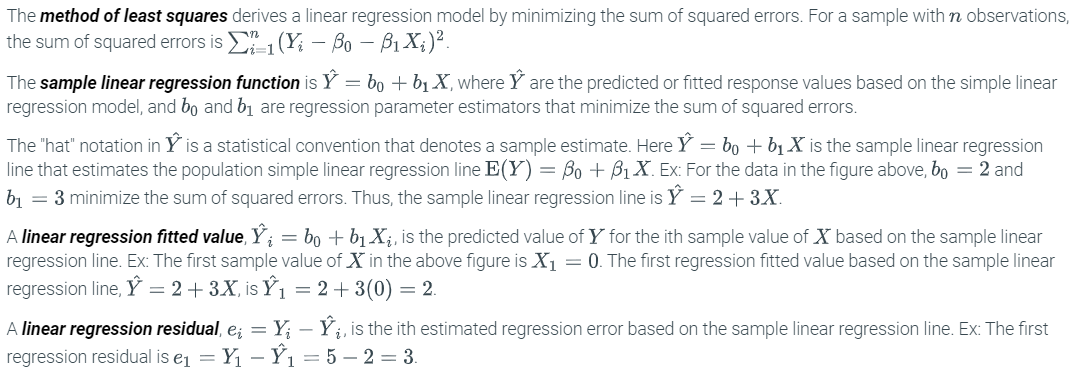
### **Linear regression model**

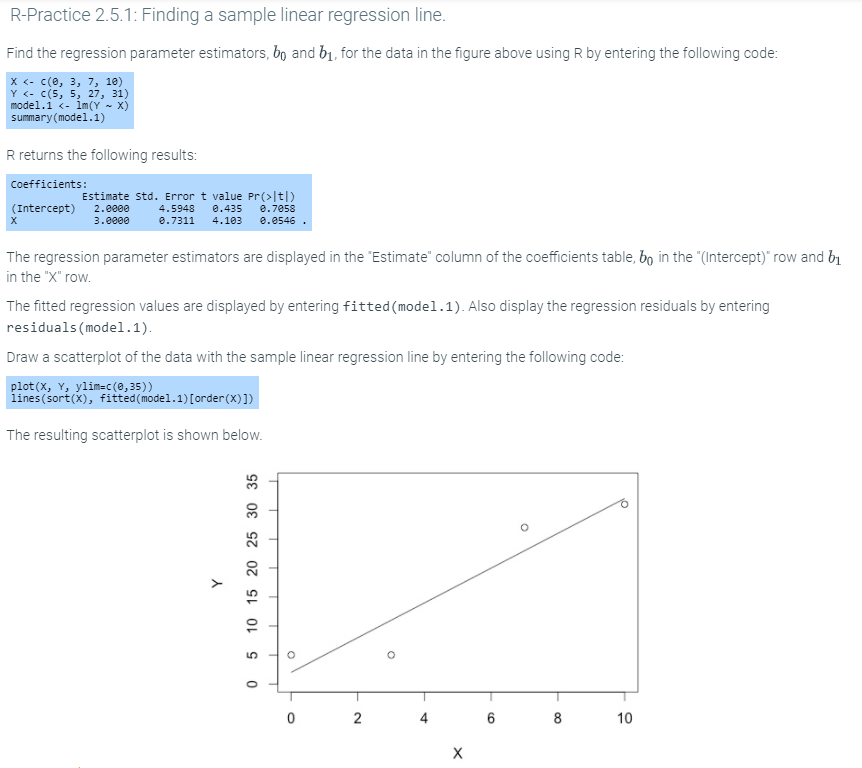
Many situations can be modeled by linear functions. A ***linear regression model*** is a linear function used to predict the value of one variable using the values of one or more variables. In a linear regression involving two variables, the ***response variable*** is the variable being modeled or predicted, while the ***predictor variable*** is the variable used to predict the response. Ex: A researcher wishes to construct a linear model that gives a student's grades given the number of hours spent studying. The predictor variable is the number of hours spent studying and the response variable is student grade.

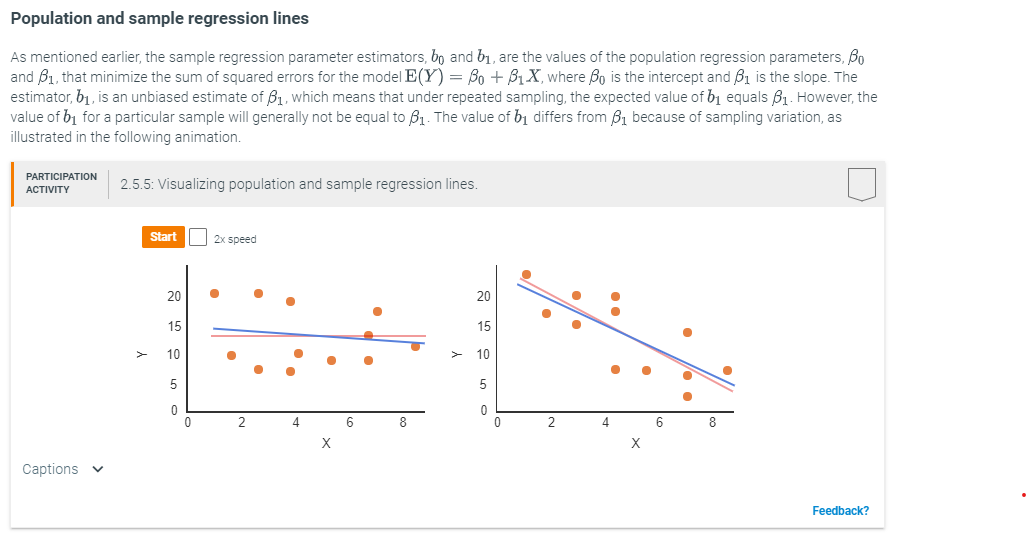
Plotting values of predictor and response variables can display all the information of a linear regression model. A ***scatter plot*** displays a set of predictor and response variable pairs as data points on a coordinate system. The predictor variable X goes on the horizontal axis and the response variable Y goes on the vertical axis. Assuming the data represents a population, the ***population linear regression function*** is given by E(Y)=β0+β1X, where E(Y) is the expected value of Y, and β0 and β1 are unknown population regression parameters. The difference between the true value Y and E(Y) is the ***absolute error*** ϵ. The regression parameters are unknown because typically the entire population dataset is not observable. Instead, a random sample of data points is generally available. Statistical inference is used to draw inferences about the population based on what can be observed in the sample. To apply statistical inference to make a prediction using regression requires estimating the population regression parameters β0 and β1 from sample data. Thus, given sample data points, finding the "best fitting" regression line for those data points provides estimates of β0 and β1. One method for estimating β0 and β1 minimizes the sum of the absolute errors for the sample points as illustrated in the animation below.

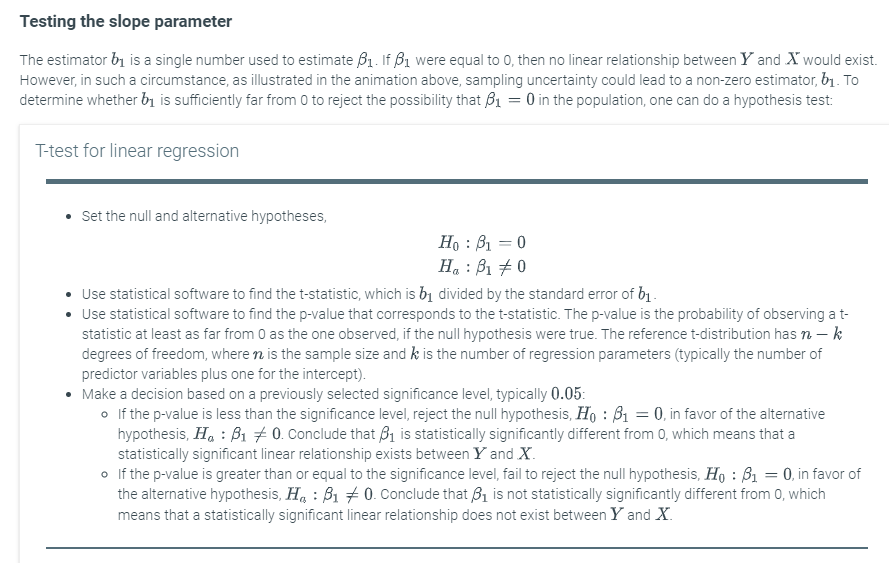


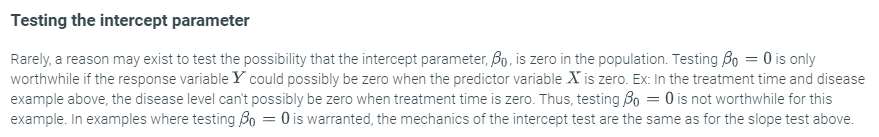








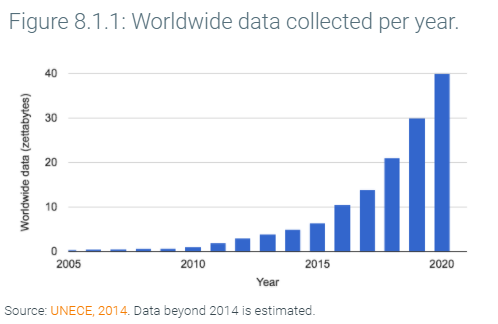


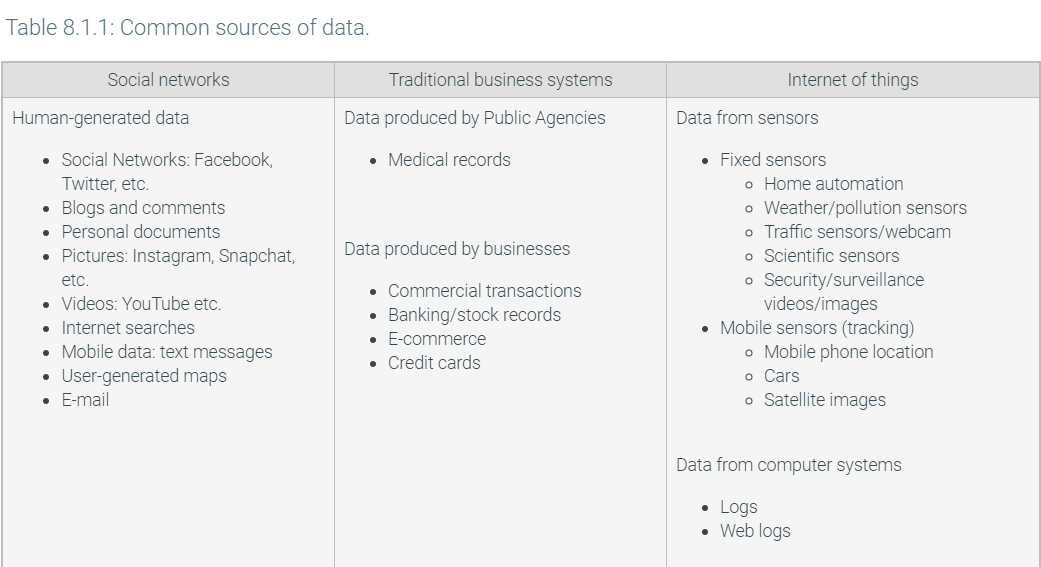


# 8.1 Introduction to data analytics

***Data*** is information, especially facts or numbers, usually collected or computed for purposes of analysis. Ex: The world population was about 300 million in the year 1000, 500 million in 1500, and 7 billion in 2000. ***Data analytics*** is the field of analyzing data to gain insight, draw conclusions, or make decisions. Ex: Analysis of the above data suggests the world's population is growing rapidly, which may influence various decisions like use of natural resources.

The amount of collected data has grown tremendously. A first reason is because computers became ubiquitous around the 1980's and can easily record data. More recently, the world-wide web became ubiquitous in the early 2000's, transforming how people do business, communicate, and recreate, in ways such that data is easily recorded and analyzed. Smartphones and tablets of the 2010's provide nearly continuous computer/web access. Plus, numerous items like streets, cars, and buildings have recently been equipped with sensors and cameras and allowing for more data collection. Some estimates are that 90% of all data ever collected was generated in just the past couple years. The figure below shows the worldwide data collected per year, in zettabytes. A ***zettabyte*** is about 1,000,000,000,000,000,000,000 bytes. The subsequent table lists common sources of data.





Example 8.1.1: Data analysis catches cheating teachers.

Standardized exams are commonly given to students in public schools. The average scores for a teacher's students are commonly used to evaluate a teacher or a school. A researcher performed data analysis to detect whether some teachers were cheating. For example, if a particular teacher's students answered the last 10 or so questions correctly more frequently than for another teacher, one might assume that the teacher filled in those last questions (correctly) for students who didn't complete the exams. Or, if a teacher's students did well above average one year, but those same students performed below average the year before and the year after, one might assume that the teacher gave students the answers.

In the book *Freakonomics*, Steven Levitt described analyses he performed on several years of exam data from Chicago public schools. He found that at least 5% of teachers were cheating. As a result of his data analysis, several teachers were fired, and cheating subsequently decreased.

[Levitt's paper describing the analysis](http://pricetheory.uchicago.edu/levitt/Papers/JacobLevitt2003.pdf)

[Two-minute video of Levitt discussing the analysis](https://www.youtube.com/watch?v=3Tu-ElppFRs)

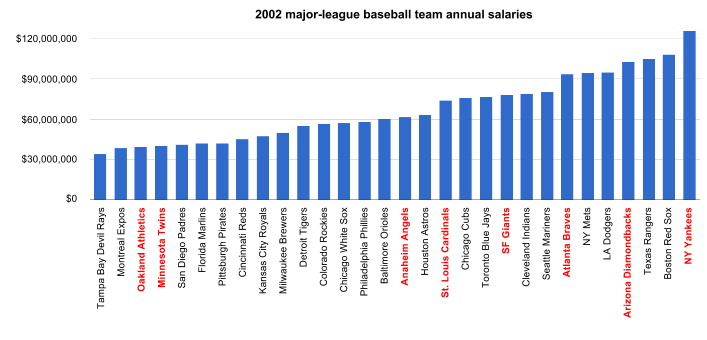
**Feedback?**

Example 8.1.2: Sports and data analytics.

In the early 2000's, the Oakland Athletics had one of the smallest budgets in professional baseball. The team leaders used data analytics to gain an edge. Traditionally, baseball players were sought based on widely-known factors like a player's batting average (how often the player got a hit), runs batted in (how many runs the player caused by making a hit), and similar numbers. Instead, through data analysis, the team leaders found less-popular factors were more important, like on-base percentage. The team thus hired players strong in those less-advertised factors and paid such players with lower salaries due to not being in high demand. The technique worked, and the Oakland Athletics made the playoffs, both in 2002 and 2003, despite having nearly the lowest salaries in the league.

Source: Oakland A's stadium ([Travis Wise](https://www.flickr.com/photos/photographingtravis/16666072878) / [CC-BY-SA-2.0](https://creativecommons.org/licenses/by-nd/2.0/) via Flickr)

This real story is the basis of the popular 2014 movie [Moneyball](http://en.wikipedia.org/wiki/Moneyball) starring Brad Pitt. Many teams, in baseball and other sports, have since adopted such data-analytic techniques. With the advent of computers, and thus more recording of data and more ability to analyze such data, data-analytic techniques are used in more arenas to gain insight and achieve better results. Ex: Online dating sites, stock market investing, language translation, and much more.



2002 playoff teams (in red above): NY Yankees, Anaheim Angels, *Oakland Athletics*, Minnesota Twins, Atlanta Braves, San Francisco Giants, Arizona Diamondbacks, St. Louis Cardinals.

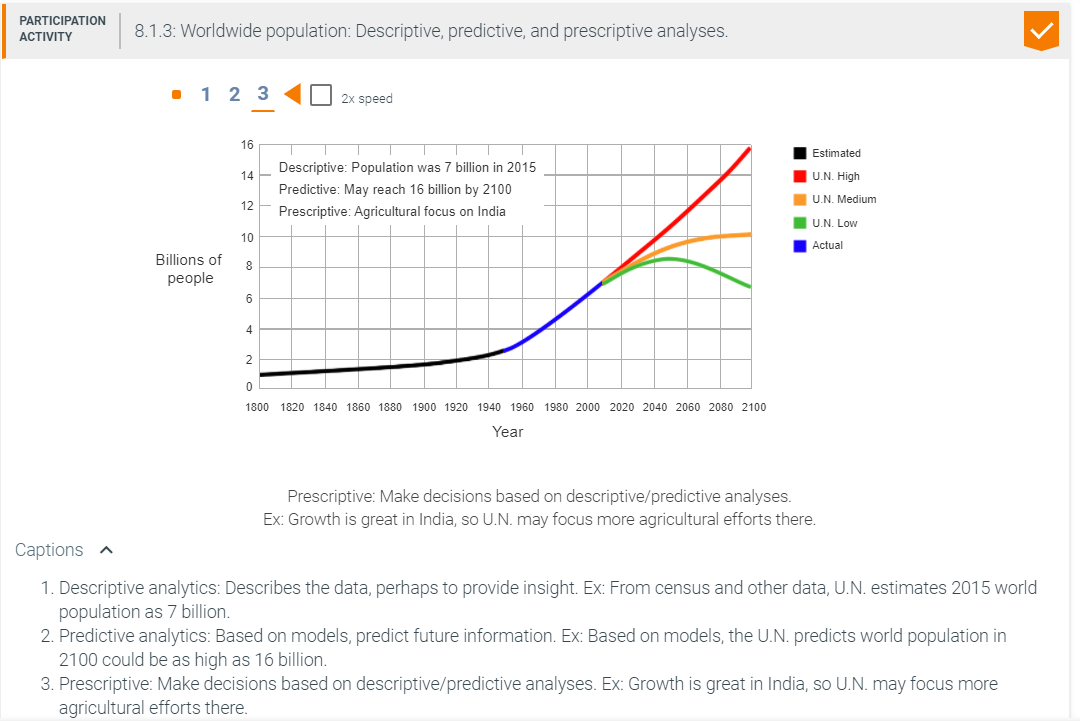
**Feedback?**

With so much data being collected today, one can imagine that data analytics is a growing field with increasing job opportunities. ***Big data*** is a term commonly used to refer to data analytics on large amounts of data, which is the form in which much data exists today. Articles summarizing jobs in big data are abundant, and summaries and predictions both describe large increases in job opportunities, such as this [2014 Forbes article on big data jobs](http://www.forbes.com/sites/louiscolumbus/2014/12/29/where-big-data-jobs-will-be-in-2015/).

Note: Past writers defined the term *datum* as a single item, and *data* as the plural of datum. Language evolves, however, and the use of the term datum is diminishing. This material follows the increasingly common usage of the term *data* for both the singular and plural.

Data analytics is sometimes viewed as having three aspects, especially in business.

* ***Descriptive*** data analytics seeks to describe data, providing insight and knowledge. Ex: Based on collected data, the world population in 2015 is about 7 billion.
* ***Predictive*** data analytics seeks to make predictions from data. Ex: Using models based on birth rates, death rates, medical care improvements, and other data, the United Nations predicts the world population will reach 11.2 billion in 2100.
* ***Prescriptive*** data analytics seeks to make decisions (prescriptions) based on data. Ex: Population predictions for specific countries help the United Nations decide where to focus agricultural development efforts.



# 8.2 Introduction to data visualization

### **Basics**

***Data visualization*** is the display of data in a format, such as a table or chart, that seeks to achieve a goal of conveying particular information to a viewer. Data presented in a text-only format often does not convey information well. Ex: Given this text-only data on 2013 median house prices in southern California counties, finding the price for a particular county is inconvenient: Los Angeles $405,000; Orange $661,000; Riverside $306,000; San Bernardino $192,000; San Diego $473,000; Ventura $464,000.

Instead, displaying the data visually as a table better conveys the information. A ***table*** displays data using rows and columns.

Table 8.2.1: Southern California median house prices by county (2013). A table may be more comprehensible than text.

|  |  |
| --- | --- |
| County | Median house price |
| Los Angeles | $405,000 |
| Orange | $661,000 |
| Riverside | $306,000 |
| San Bernardino | $192,000 |
| San Diego | $473,000 |
| Ventura | $464,000 |

**Feedback?**

As another example, the following data represents California median house prices from 2000-2010: 2000 $241,000; 2001 $262,000; 2002 $316,000; 2003 $372,000; 2004 $451,000; 2005 $523,000; 2006 $556,000; 2007 $560,000; 2008 $348,000; 2009 $275,000; 2010 $305,000. A table conveys the information better than text, but if the goal is to educate the viewer about the housing price "bubble" that grew and then burst in 2008, a chart is even better.

Table 8.2.2: California median house prices, 2000-2010. This table is better than text, but a chart would be even better.

|  |  |
| --- | --- |
| Year | California median house price |
| 2000 | $241,000 |
| 2001 | $262,000 |
| 2002 | $316,000 |
| 2003 | $372,000 |
| 2004 | $451,000 |
| 2005 | $523,000 |
| 2006 | $556,000 |
| 2007 | $560,000 |
| 2008 | $348,000 |
| 2009 | $275,000 |
| 2010 | $305,000 |

**Feedback?**

Figure 8.2.1: California median house prices, 2000-2010. A chart may convey certain information better than a table: The bubble growth and burst are clearly visible.



**Feedback?**

An ***information conveyance goal*** is the intent of presenting data to people. The best visualization depends on the information conveyance goal. If providing data to a company's executives to help inform a relocation decision to southern California, such data being the salary amounts needed to support employee home purchases, then the above table may enable convenient lookup for a particular considered county. In contrast, to help real estate agents understand the housing price bubble from 2000-2010, then the chart is better. *The data visualization task consists largely of choosing among the various methods for displaying data (tables, charts, etc.) to best satisfy an information conveyance goal.* Other sections discuss such methods.

### **Minimizing superfluous information**

Once a method is chosen, a key task of data visualization is to optimize for a particular information conveyance goal. A key optimization in data visualization is ***minimizing superfluous information***. The optimization can be stated as follows:

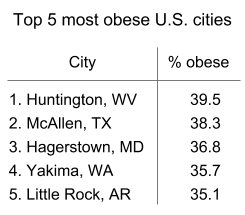
It's not done when nothing else can be added,

it's done when nothing else can be *removed*. Saint-Exupery quote

The below chart comes from the same data as the above chart, but has superfluous information. More data labels on axes and more gridlines exist than are really necessary. The chart title and legend are redundant, given the axis titles. Labeling each data point provides unneeded details. All the extra information detracts from the particular information conveyance goal.

Figure 8.2.2: The designer of this chart did not minimize superfluous information.

Figure 8.2.3: Simplified version of the above table achieved by minimizing superfluous information.



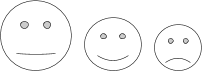
**Feedback?**

### **The brain's preattentive versus conscious processing**

Designing data visualizations is aided by some understanding of how the human brain processes sensory data (from sight, smell, hearing, taste, and touch) via a ***dual process*** approach: unconsciously and consciously. The unconscious processing, sometimes called ***preattentive*** or ***system 1***, is automatic and almost instant. The conscious processing, sometimes called ***system 2***, is much slower that the preattentive processing. The brain preattentively processes several visual features, such as size, color, and location; such fast processing likely yielded many survival advantages when humans lived as hunters/gatherers.

Good data visualizations make use of a human's preattentive processing, leading to better and faster understanding. As an example, one might *quickly* glance at the below figure, then continue reading.

Figure 8.2.4: Preattentive processing detects many visual features almost immediately.



**Feedback?**

One may note how quickly the brain noticed the relative sizes of the faces, from largest to smallest. The figure depicts data from a [Forbes 2013 survey](http://www.forbes.com/sites/susanadams/2013/10/10/unhappy-employees-outnumber-happy-ones-by-two-to-one-worldwide/) of U.S. workers: 50% are not engaged, 30% are happy, and 20% are unhappy. (In this case, the brain quickly detects the size ordering, but doesn't precisely process the relative size ratios).

Extensive past research on visual processing shows that humans preattentively process certain graphical features, like length, 2D location, width, size, intensity, color, and shape. Furthermore, length and 2D location are processed with high quantitative precision, meaning, for example, that humans preattentively determine that one bar is half as long as another, or placed half as high. Width, size, and intensity provide less quantitative precision, meaning, for example, that humans preattentively notice a bar is less wide than another, but whether the bar is a half or a third as wide is harder to detect.

Other features like color and shape are quickly processed, but of course have no intuitive quantitative meaning. Color and shape are quickly recognized as forming groups, which is useful when drawing charts.

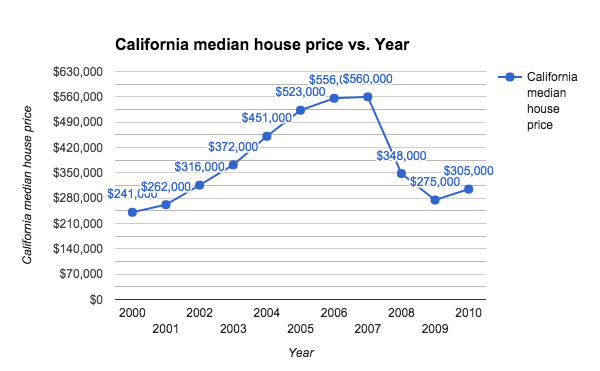
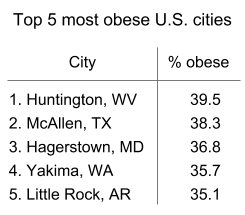


Figure 8.2.3: Simplified version of the above table achieved by minimizing superfluous information.



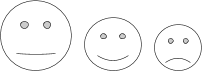
**Feedback?**

### **The brain's preattentive versus conscious processing**

Designing data visualizations is aided by some understanding of how the human brain processes sensory data (from sight, smell, hearing, taste, and touch) via a ***dual process*** approach: unconsciously and consciously. The unconscious processing, sometimes called ***preattentive*** or ***system 1***, is automatic and almost instant. The conscious processing, sometimes called ***system 2***, is much slower that the preattentive processing. The brain preattentively processes several visual features, such as size, color, and location; such fast processing likely yielded many survival advantages when humans lived as hunters/gatherers.

Good data visualizations make use of a human's preattentive processing, leading to better and faster understanding. As an example, one might *quickly* glance at the below figure, then continue reading.

Figure 8.2.4: Preattentive processing detects many visual features almost immediately.

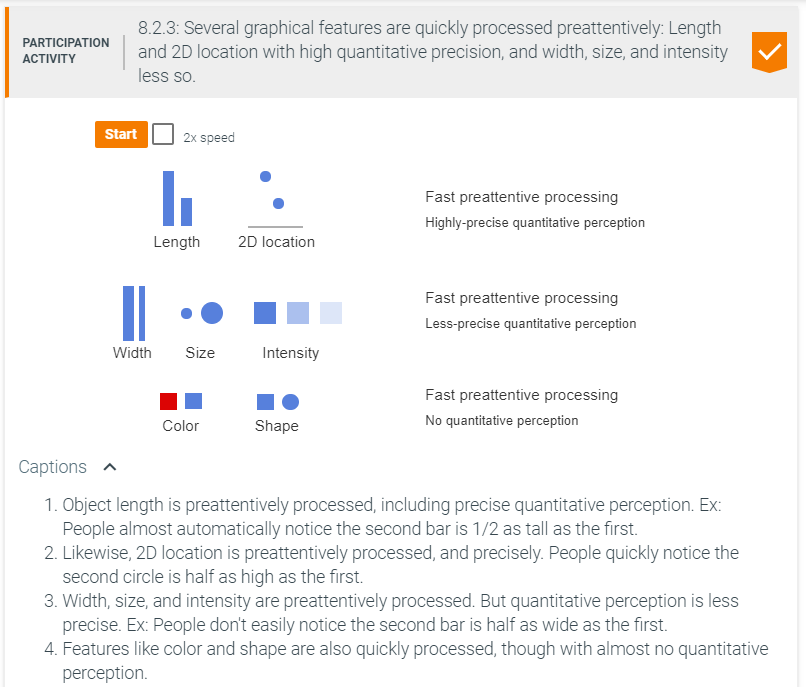


**Feedback?**

One may note how quickly the brain noticed the relative sizes of the faces, from largest to smallest. The figure depicts data from a [Forbes 2013 survey](http://www.forbes.com/sites/susanadams/2013/10/10/unhappy-employees-outnumber-happy-ones-by-two-to-one-worldwide/) of U.S. workers: 50% are not engaged, 30% are happy, and 20% are unhappy. (In this case, the brain quickly detects the size ordering, but doesn't precisely process the relative size ratios).

Extensive past research on visual processing shows that humans preattentively process certain graphical features, like length, 2D location, width, size, intensity, color, and shape. Furthermore, length and 2D location are processed with high quantitative precision, meaning, for example, that humans preattentively determine that one bar is half as long as another, or placed half as high. Width, size, and intensity provide less quantitative precision, meaning, for example, that humans preattentively notice a bar is less wide than another, but whether the bar is a half or a third as wide is harder to detect.

Other features like color and shape are quickly processed, but of course have no intuitive quantitative meaning. Color and shape are quickly recognized as forming groups, which is useful when drawing charts.



# 8.3 Types of data

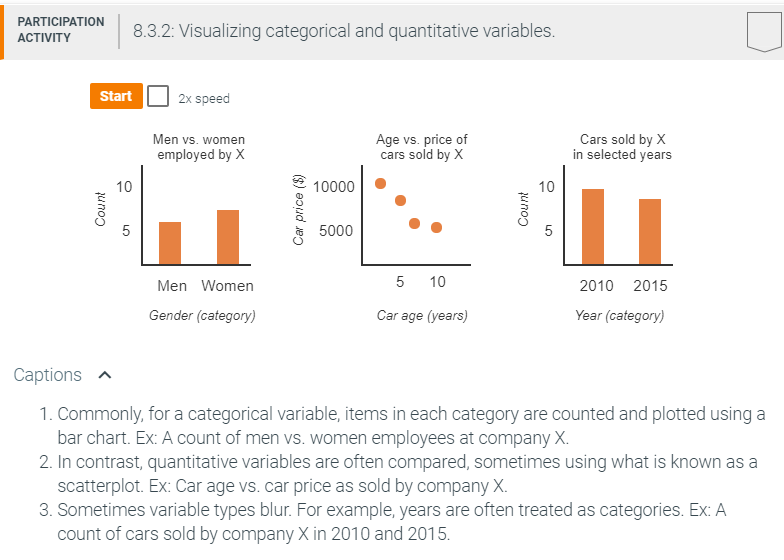
Data is typically represented using variables. A ***variable*** is an item that can have different ("varying") values. Ex: A person's age is a variable and can have the value 10, 33, 99, or other values. Variables are often considered as being of two possible types:

* A ***quantitative variable*** can take on a numeric value (quantitative data) that can be measured and ordered. Ex: A person's age, the outside temperature, and a meal's price are quantitative variables. Example numeric values are an age of 33 or 99 years, a temperature of 40 or 45 degrees, and a price of 12 or 15 dollars.
* A ***categorical variable*** can take on the value (usually a label) of one of several categories. Ex: A person's gender, seasons, and U.S. companies are categorical variables. Gender can be male or female, seasons can be fall, winter, spring, or summer, and U.S. companies can be Wal-Mart, McDonalds, UPS, etc. A categorical variable is often called a ***qualitative variable*** (known by qualities, rather than quantities).

Most numbers represent quantitative data, but exceptions exist. Ex: A person's phone number is a number but is not quantitative data; a phone number isn't measured, nor ordered; people don't say: "Joe's phone number is greater than Mary's." In general, if adding the numbers makes sense, the variable is likely quantitative, else categorical. (People may add ages but don't add phone numbers.)

A reason for distinguishing variable types is that each type is handled differently in data analytics. Ex: A categorical variable typically involves counting the instances of each category, often then depicted with a bar chart or pie chart. But a quantitative variable is commonly plotted versus another quantitative variable, often depicted with a scatter plot or line chart. Those chart types are described in other sections.

Sometimes the distinction between variable types is blurred. For example, years are clearly quantitative but are often treated as categorical. Similarly, the months (January, February, March, ...) are mostly categorical but can be represented as numbers (1, 2, 3, ...) and measured, ordered, and even added. Despite such blurring, the general distinction is still useful.



Two types of categorical variables are often distinguished:

* A ***nominal variable***'s categories have no ordering, existing in name only, like apples, oranges, and grapes. ("Nominal" means "in name only").
* An ***ordinal variable***'s categories have an ordering, like disagree, neutral, and agree.

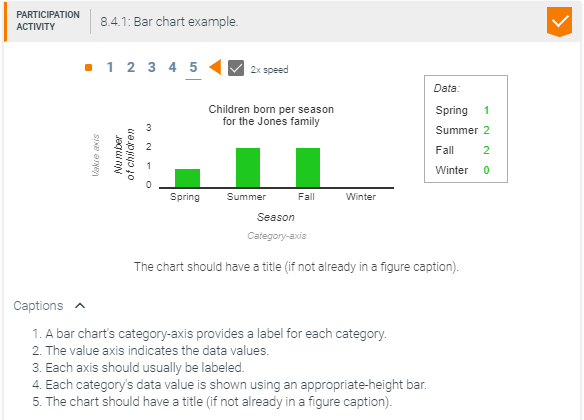
The difference is sometimes relevant. Ex: On a chart, the ordinal variables would almost always be sorted along the x-axis, listed as "small medium large" rather than arbitrarily as "small large medium."

Two types of quantitative variables are often distinguished:

* A ***continuous variable***'s values are infinite along a continuum of values within a range, typically real numbers. Continuous variables usually represent measurements, like height (0.00104 meters) or temperature (98.6 degrees).
* A ***discrete variable***'s values are finite within a range, typically integers. Discrete variables usually represent countable items, like people in a family (5) or cars in a city (502,434). Generally, if "number of" can be added to the beginning, the variable is discrete, like "number of people in a family", but not "number of height". Note: "Discrete" means separate or distinct, not to be confused with "discreet" which means careful or unobtrusive.

# 8.4 Bar charts

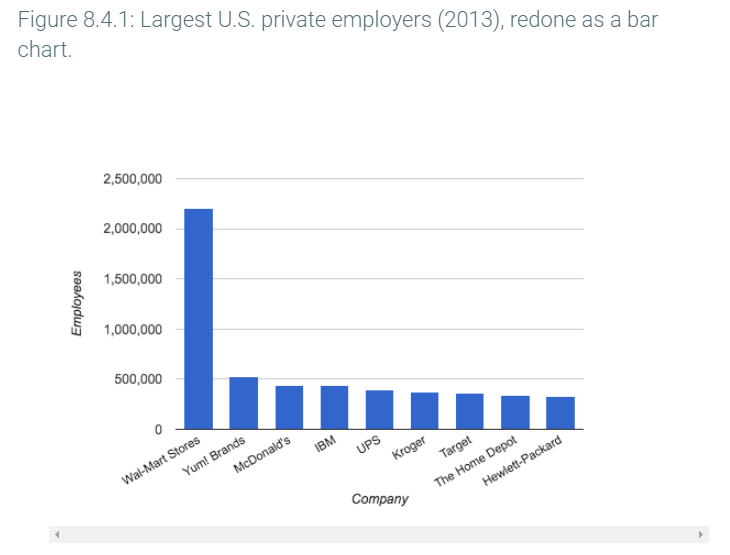
A ***bar chart*** depicts data values (usually counts) for a categorical variable, using rectangular bars having lengths proportional to category values. The chart is drawn using two axes: a category-axis and a value axis (the counts).



A bar chart is especially useful to intuitively convey *relative* differences among categories, due to the brain's ability to quickly (preattentively) process differences in length. Ex: Given the following data on largest U.S. private employers and an information conveyance goal of showing Wal-Mart's dominance, the subsequent bar chart intuitively conveys the goal.

Table 8.4.1: Largest U.S. private employers (2013).

|  |  |
| --- | --- |
| Company | Employees |
| Wal-Mart Stores | 2,200,000 |
| Yum! Brands | 523,000 |
| McDonald's | 440,000 |
| IBM | 434,246 |
| UPS | 399,000 |
| Target | 361,000 |
| Kroger | 343,000 |
| The Home Depot | 340,000 |
| Hewlett-Packard | 331,800 |



Categories are commonly ordered along the category-axis. Above, the nominal variable Company's categories were ordered by each category's data value, highest (Wal-Mart's 2,200,000) to lowest (Hewlett Packard's 331,800). If instead the categories represented years (1970, 1980, etc.) or some other measure of time, such an ordinal variable's categories would be ordered with time increasing to the right.

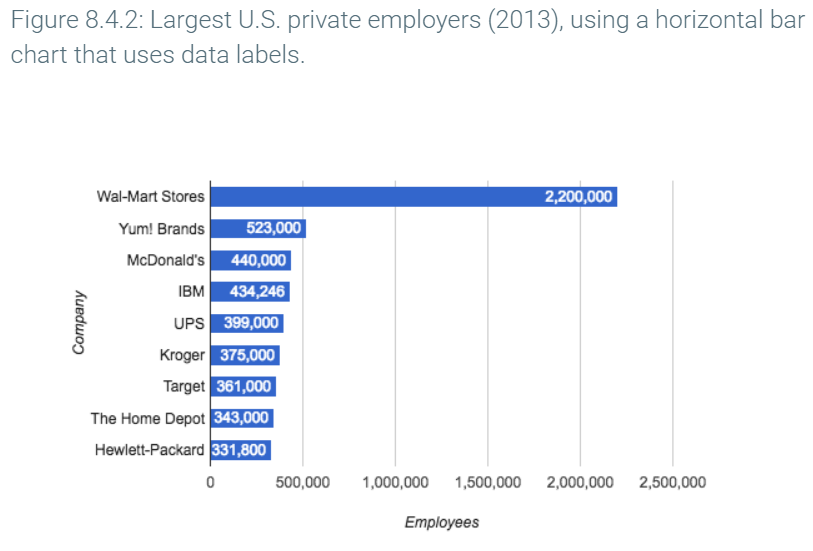
Each listed category has a ***category label***, such as "Kroger" and "Target" above. If labels don't fit when written horizontally, the labels can be rotated, such as rotated 30 degrees as above. Rotations of 60 or 90 degrees are also common.

The appropriate increment for the value axis is important for readability. Small increments clutters the chart with too many values: Above, an increment of 100,000 would yield 25 values. Too few, like increments of 1,000,000 above, can make visually estimating a category's value difficult. Above, increments of 500,000 leads to reasonably-estimable values (Wal-Mart can be seen to have a value of about 2,200,000) without clutter.

Grid lines help the viewer estimate the value for a category. If easier estimation is desired, additional grid lines can be drawn between number increments (but kept minimal to reduce clutter). If precise values need to be conveyed, data values known as ***data labels*** can be shown next to the bars, or even inside the bars, as in the chart further below. However, precise values are not typically needed if the information goal is to show relative differences among categories.

A bar chart can be drawn vertically (as above) or horizontally (as below). A horizontal bar chart is useful for long labels, like "Wal-Mart Stores", which need not be written at an angle as was done above. A horizontal bar chart is also useful when numerous categories exist because the categories increase the height rather than width, and due to the nature of paper and computers, width is usually more limited while height is less limited.

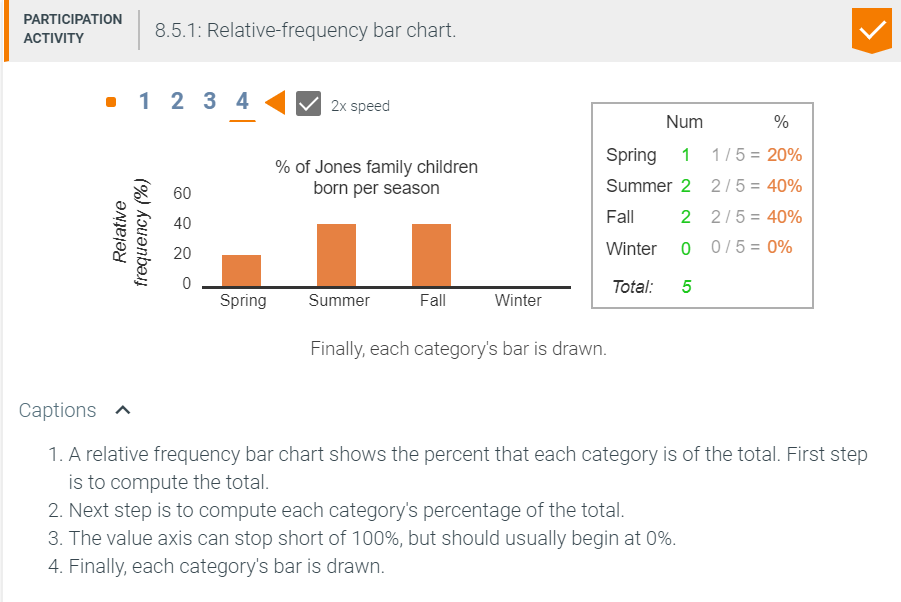
In contrast, a vertical chart is often preferred due to "height" intuitively representing amount. This preference is especially the case when negative values are shown (which would appear going downwards).



A note on terminology: Some authors and tools use the term "bar chart" to refer exclusively to a horizontal bar chart. In that case, a ***column chart*** is a term used for a vertical bar chart. However, the term bar chart is widely used for vertical charts by many respected authors and tools. Thus, this material uses the term bar chart for either orientation, prepending the word horizontal or vertical where appropriate.

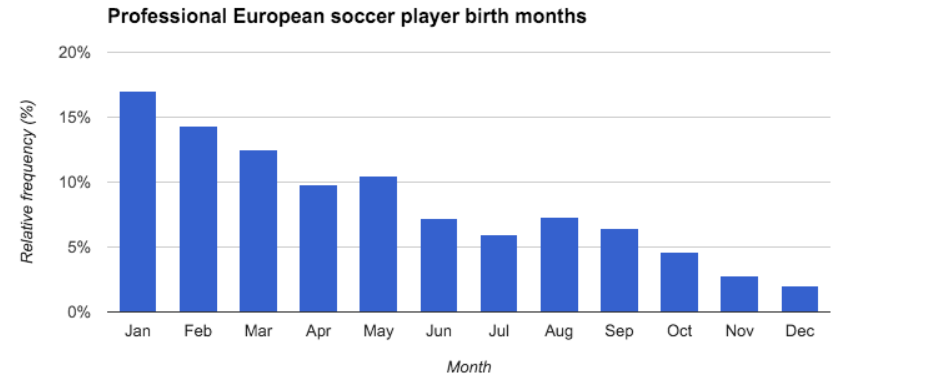
# 8.5 More bar charts

A basic bar chart's value axis provides the raw data value for each category. Instead, a ***relative-frequency bar chart*** shows each category's portion of the total data, typically as a percentage. The data total is first computed, then the percentage for each category is computed, and finally those percentages are drawn as a bar chart.



Example 8.5.1: European soccer player birth months.

The following relative-frequency bar chart depicts the birth months for professional European soccer players. Most players were born in the first few months of the year. A possible explanation is the January 1 cutoff date for youth soccer leagues, meaning kids born in January are the oldest on their teams, while kids born in December are the youngest. Older kids are likely to be better players initially, causing coaches to give them more attention and playing time, and also causing those kids to enjoy playing and thus practicing more. (Source: [Freakonomics](http://freakonomics.com/2011/11/02/the-disadvantages-of-summer-babies/)).



The above example illustrates the power of data analytics and data visualization. Having understood such data, many parents now choose to postpone their child's entry onto a sports team so that the child is not the youngest on the team. In fact, similar data exists for schoolkids: Kids born just after cutoff dates tend to be more successful in school (getting more attention from teachers, and causing those kids to feel smarter and enjoy school more), and thus many parents now delay their child's school enrollment by a year. (Source: [The New Yorker](http://www.newyorker.com/tech/elements/youngest-kid-smartest-kid)).

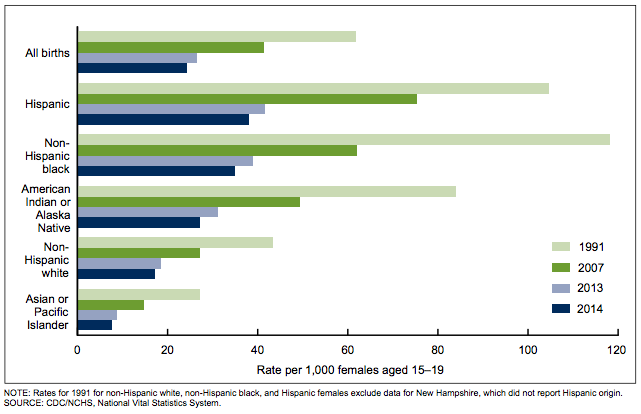
A ***grouped bar chart*** depicts two or more groups on a single bar chart, with each group using a different colored (or shaded) bar. A ***legend*** indicates what group each color represents in a chart. Ex: The below bar chart shows the number of men vs. women in the U.S. workforce over time. The categories are decades (1970, 1980, ...), the category values are number of people, and the two groups are men and women.

Because the categories represent time (decades), a vertical chart is preferred so that time proceeds to the right.

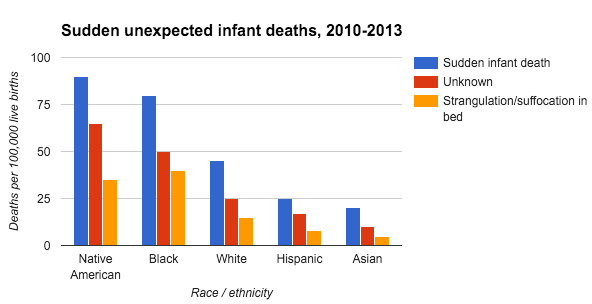


Example 8.5.2: Case study: CDC teen birth rate data.

Below is a grouped bar chart published by the U.S. Centers for Disease Control on teenage birth rates, with categories being race, and groups being the years 1991, 2007, 2013, and 2014. The chart appears in a section titled "Birth rates for teenagers declined to historic lows in 2014", which clearly suggests the information conveyance goal being the decrease.



A ***stacked bar chart*** is a grouped bar chart where the bars are stacked on each other. A stacked bar chart is useful for showing each category's *total*, while still showing the breakdown of groups within each category (though the relative sizes of each group in a category becomes harder to see due to not being side-by-side). The following shows a stacked bar chart and grouped bar chart for the same data.



The power of descriptive data analytics is readily seen in the above chart. The data immediately compels people to want to understand why Hispanics and Asians have such low infant death rates, to perhaps help educate other people of how to reduce rates.

The concept of a relative frequency chart is commonly applied to a stacked bar chart. Such a chart clearly shows how a particular group's proportion of the total changes across categories (such as across years).

Figure 8.5.3: Massachusetts state spending on healthcare versus all other state spending, using a relative frequency stacked bar chart.

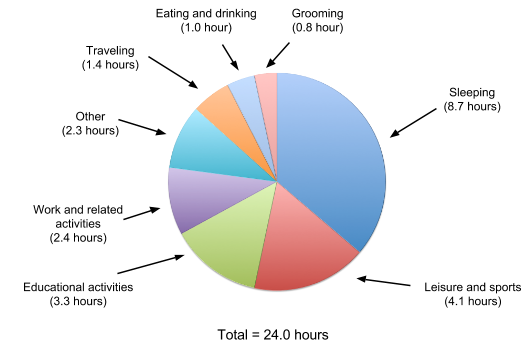
Source: [Kaiser Family Foundation, State Health Facts](http://kff.org/statedata/)

# 8.6 Pie charts

A ***pie chart*** shows relative frequency for categories using a circle, with each category shown as a slice of appropriate size. The appearance is one of a sliced pie, leading to the chart's name. Because the brain interprets length differences more precisely than size differences, people often prefer to create bar charts. However, pie charts remain common, perhaps in part because the brain finds curved shapes more pleasant than rectangular shapes.

The following example shows a pie chart whose main information conveyance goal is to indicate the number of hours that college students spend on various activities, with a second goal to show the relative times. The data labels (in hours rather than %'s) help ensure the main goal is achieved, and the relative slice sizes achieve the second goal.

Figure 8.6.1: U.S. college student weekday average time use.

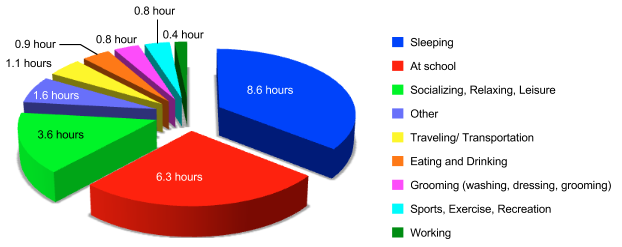
Source: [U.S. Bureau of Labor Statistics, American Time Use Survey, 2010-2014](http://www.bls.gov/tus/charts/students.htm)

**Feedback?**

Pie charts are commonly misused. Below is a pie chart that shows U.S. high-school student time use. The use of a legend to represent the categories interferes with the intuitive processing that graphics can provide, instead requiring the user to look back and forth to determine which slice corresponds to which activity. Some colors may be hard to distinguish, such as the two variations of blue used for the Sleeping and the Other categories below. (Furthermore, the reason for adding the 3D effect is questionable).

Thus, if a pie chart is to be used, many recommend avoiding a legend, instead associating the category labels directly with slices as above.

Figure 8.6.2: U.S. high school student weekday average time use. A pie chart with a legend hinders intuitive processing.

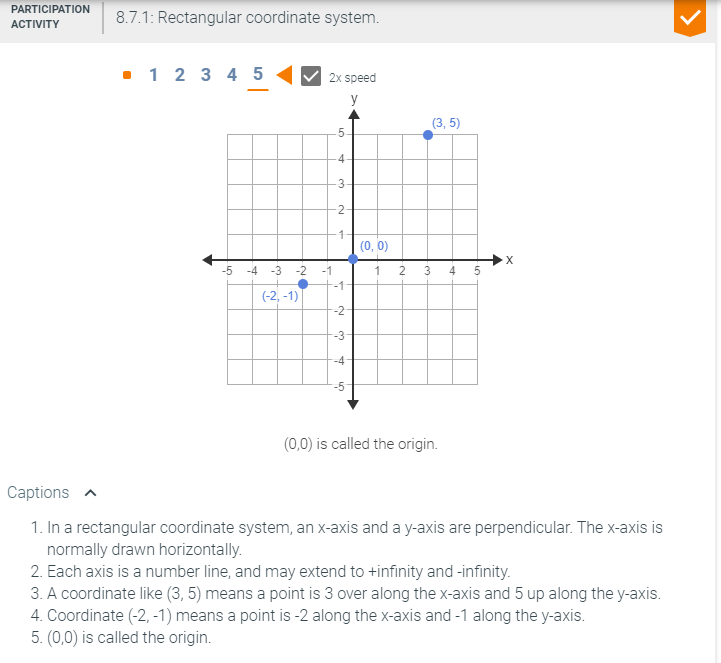


# 8.7 Scatter plots

A ***scatter plot*** (or ***scatter chart***) depicts the relationship between two (usually quantitative) variables, on a rectangular coordinate system, each axis corresponding to one variable.

### **Rectangular coordinate system**

The ***rectangular coordinate system*** consists of two perpendicular lines usually labeled x and y, called ***axes***, such that a coordinate (x, y) indicates a point's distance along each axis. The intersection of axes is called the ***origin*** and is at point (0, 0). The x-axis is drawn horizontally, and the y-axis



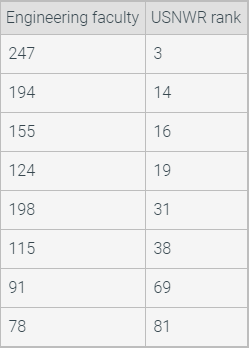
The rectangular coordinate system is sometimes called the ***Cartesian coordinate system***, named for 17th century mathematician Rene Descartes. A Cartesian coordinate system may involve two, three, four, or more axes. Two are easy to display, three are harder, but four or more are not readily visualizable.

### **Creating a scatter plot**

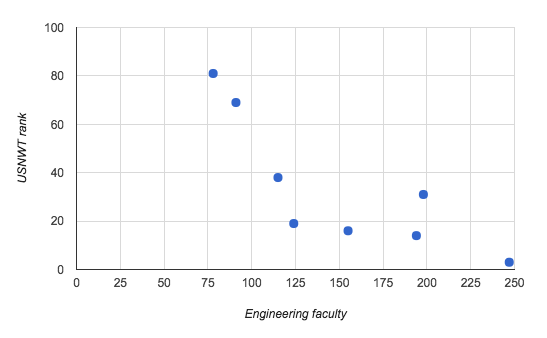
Given data consisting of value pairs for two variables, each pair becomes a coordinate that is plotted on the rectangular coordinate system. Commonly, the variables only have non-negative values, and thus only the upper-right quadrant is drawn.

Example 8.7.1: Number of engineering faculty versus school rank.

To inform a decision of whether to hire new engineering faculty, a dean collected data showing number of faculty versus engineering school rank (using the U.S. News and World Report ranking, lower rank is better), for eight University of California campuses (UC Berkeley, UCLA, etc.). A table of the data (2014) is shown below. For example, the school with 247 engineering faculty is ranked #3 in the country, while the school with only 78 engineering faculty is ranked #81.



Below is a scatter plot showing engineering faculty size vs. engineering school rank (lower rank is better) for the eight UC campuses (2014). Each row in the above table becomes a coordinate, leading to the following scatter plot.

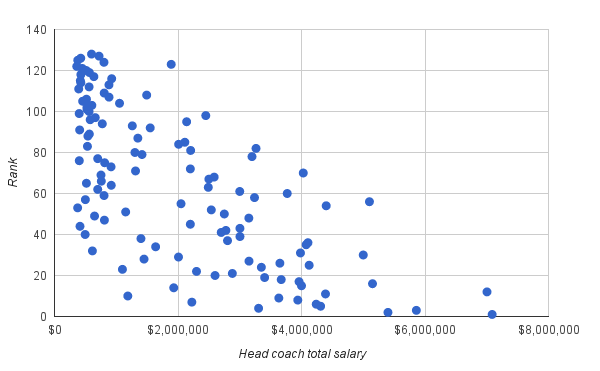


The scatter plot clearly shows the relationship between number of faculty and rank. The data suggests, but does not prove, that increasing the number of engineering faculty may be important to improving rank.

**Feedback?**

A scatter plot often has numerous data points that are "scattered" about the rectangular coordinate system, leading to the name "scatter plot". Below is a scatter plot showing all 128 college football team rankings (lower is better; #1 is best) and the total salary for each team's head coach ([Source 1](https://www.teamrankings.com/college-football/ranking/predictive-by-other), [Source 2](http://sports.usatoday.com/ncaa/salaries/)). (Yes, college head coaches often have multi-million-dollar salaries). A viewer quickly sees that more than half of coaches earn over $1 million, that many earn 3 or 4 million, and that most poorly-ranked team's coaches earn just a few hundred thousand dollars. The viewer can also see that several teams, in the lower left, seem to be getting a great bargain.

Figure 8.7.1: College football rankings vs. head coach salary.



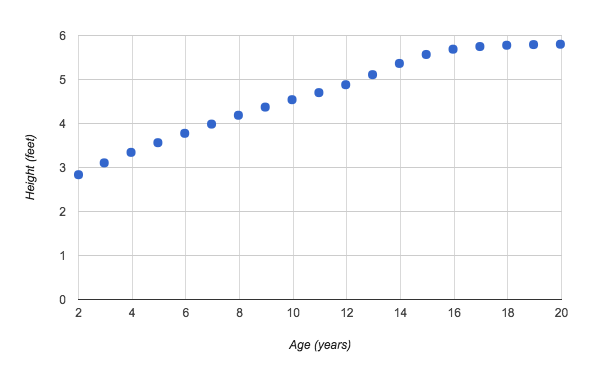
### **Independent and dependent variables**

Commonly, a scatterplot seeks to visualize how one variable depends on another. Above, the dean was interested in showing how rank might depend on the number of faculty. To distinguish dependent and independent variables, one might consider to what extent a viewer might ask: If one variable's value is 5 (or some other value), how does that affect the other variable? The variable that a user "sets" is the ***independent variable***, while the variable that is then determined based on that set value is the ***dependent variable***. The independent variable is usually plotted on the x-axis, and the dependent on the y-axis. Viewers are accustomed to moving over on the x-axis with a desired independent value, then up the y-axis to see the dependent value.

Below is an example showing how height depends on age for males aged 2 to 20. Age is the independent variable (e.g., one might ask "If a male is age 12, what height might he be?"), so is plotted on the x-axis.

Example 8.7.2: Height vs. age for males aged 2 to 20.

Parents commonly wish to known how their child's height or weight compares to other children. The following scatter plot shows the median height for U.S. males aged 2 to 20.

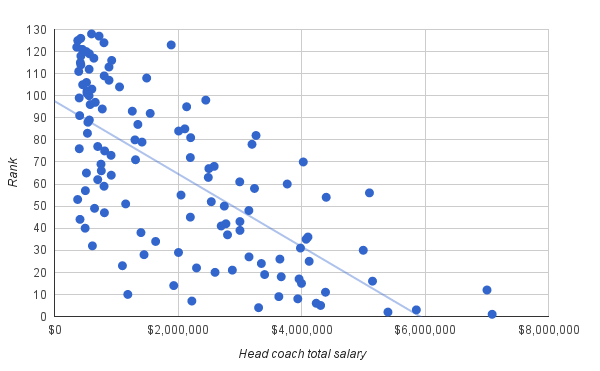


### **Trend line**

When two variables' data show some relationship but the relationship is loose, a trend line can be added to visualize that relationship. A ***trend line*** is a straight or curved line in a scatter plot showing the general relationship of two variables. The first figure below shows a trend line for the earlier coach salary vs. team rank data, and the second figure shows a trend line for the earlier engineering faculty vs. rank data. The line eases the task of answering a question like "What rank can be expected for 100 faculty?". The viewer just moves over 100 on the x-axis, then up until reaching the line, leading to an estimated rank of 50.

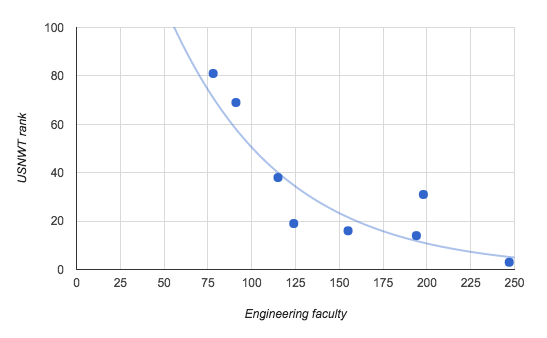
If a relationship is linear, a straight trend line is usually best. Sometimes a relationship is better represented by a curved trend line, such as an exponential curve. The best-fitting line (straight or curved) for a scatter plot is called a ***regression line***.

Figure 8.7.2: Linear trend line added to earlier scatter plot for head coach salary vs. team ranking.



**Feedback?**

Figure 8.7.3: Curved (exponential) trend line added to earlier scatter plot of engineering faculty vs. rank.



# 8.8 Line charts

A ***line chart*** (or ***line graph***) depicts data trends by using straight lines to connect subsequent data points in a scatter plot. The straight lines show the general direction that data changes over time. Because trends involve time, line charts commonly use a time metric for the horizontal axis. Ex: Given the following data on Apple stock prices from March 2015 to March 2016, the subsequent line chart shows how Apple's stock price changes (vertical axis) as each month passes (horizontal axis).

Table 8.8.1: Apple stock prices.

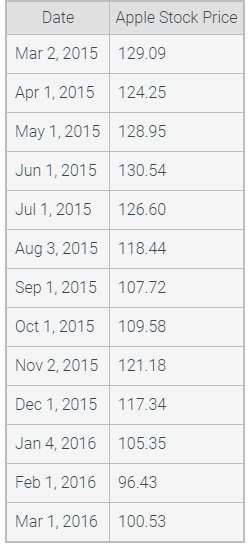
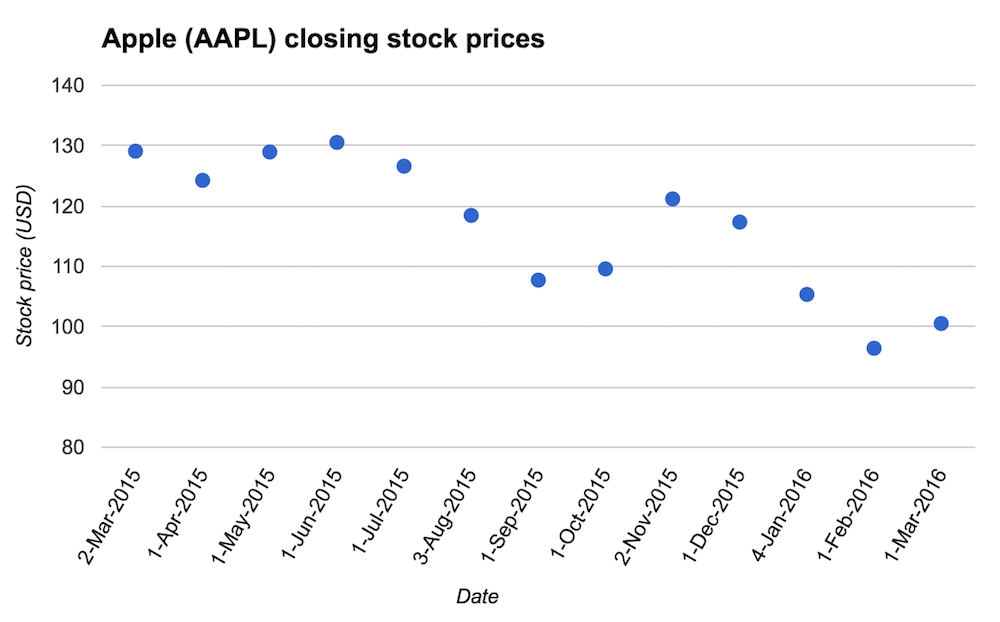
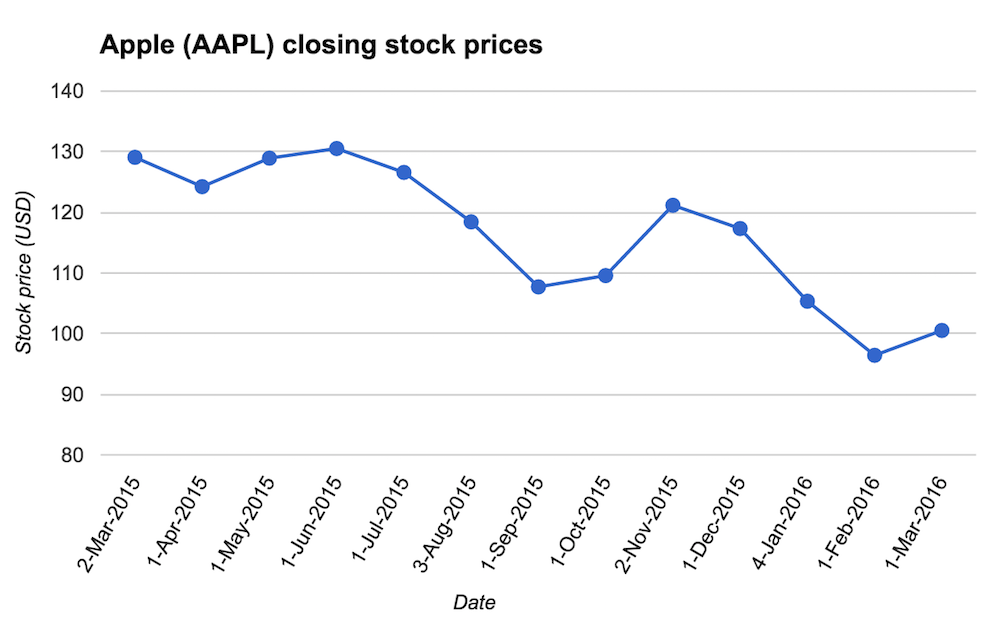


Figure 8.8.1: Apple stock charts (March 2015 - March 2016) with and without lines.





**Feedback?**

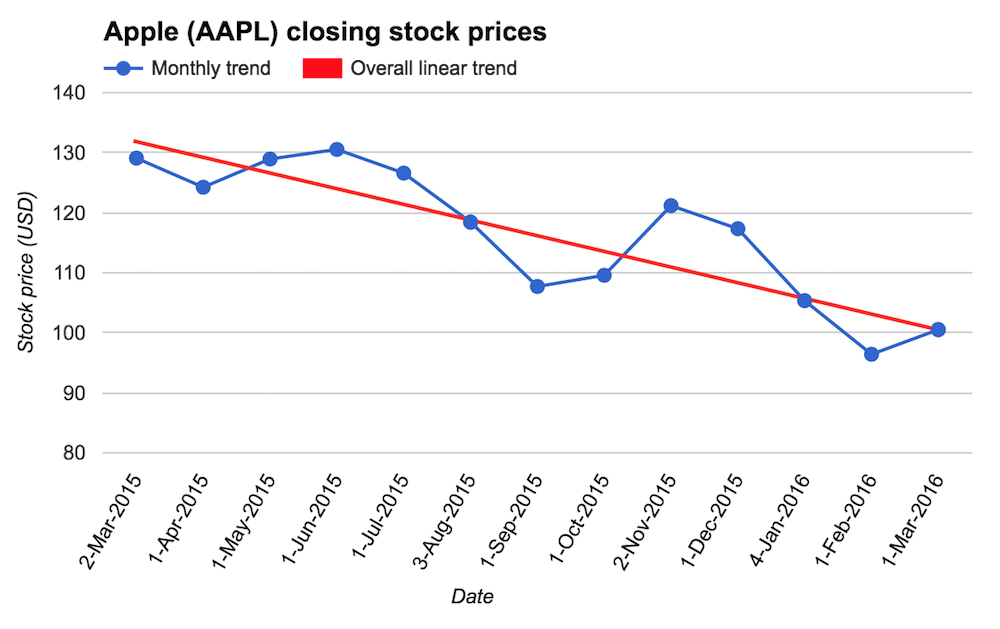
The main benefit of a line graph's lines is to help the brain quickly determine if values are increasing, decreasing, or remaining constant between data points. Steeper lines indicate more rapid increases or decreases, while flatter lines indicate little change between data points. Ex: The line graph above clearly shows that the steepest increase in the stock value was from October to November, which may lead investors to research what happened to Apple in October 2015.

Lines also help convey that values exist between data points. Ex: Although the Apple line chart shows two consecutive data points for July 1 and August 3, the stock price took on many values in between those dates. The line connecting July 1 and August 3 does not represent real data, but rather, a basic trend of the data change between data points.

A ***linear trend line*** is a straight line that depicts the general direction data changes from the first to last data point, often added to summarize the entire chart. A good linear trend line is typically computed using various techniques such as linear regression (discussed elsewhere), and is not merely a connecting of the first and last points.

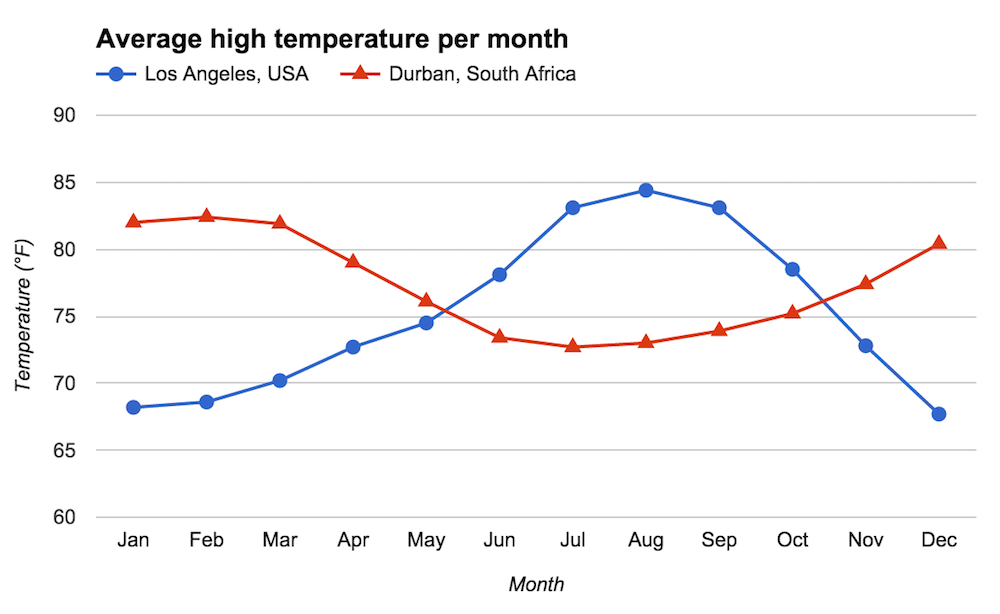
In the Apple stock line chart below, a linear trend line is added in red and starts slightly above the first data point. While the stock price had two large increases from March 2015 to March 2016, the linear trend line clearly shows the stock tended to decrease during that time.

Figure 8.8.2: Apple stock prices (March 2015 - March 2016) line chart with overall trend line.



Multiple data sets are commonly shown in one line chart to highlight differences. Each data set is distinguished by different color, data point shape, and/or line style, as noted by a legend. Ex: The line chart below shows how Los Angeles and Durban (South Africa) temperatures differ per month, due to being on opposite sides of the equator. The chart also shows how Los Angeles has more extreme temperature swings than Durban.

Figure 8.8.3: Line chart showing multiple data sets: average high temperatures for Los Angeles, USA, and Durban, South Africa.



# 8.9 Tables

### **Table benefits: Precise values, different magnitudes**

A ***table*** presents data as rows and columns. Compared to charts, one of a table's benefits is showing precise values. A second benefit is showing values with very different magnitudes.

Perhaps the most common table format provides corresponding values for two (or more) variables. A table's ***header*** row provides variable names, under which each subsequent row provides corresponding values. One or more summary rows may provide totals, means, or other summaries for some variables.

The table below lists an Age variable's values from 0 to 17, with corresponding Expenditure values for each age. The information conveyance goal is more so to indicate the expenditure values (especially the total), and less so to point out the trend as a child ages. Thus, an analyst may choose to use a table over a chart.



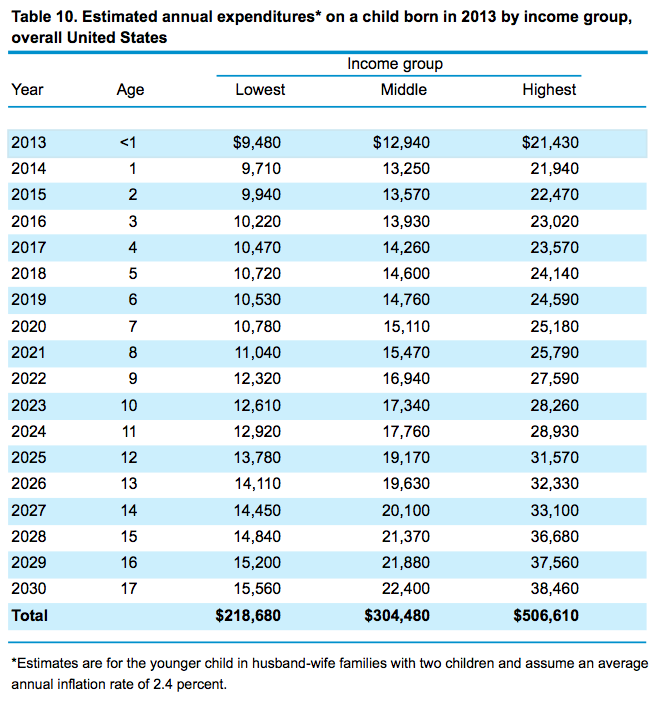
### **Other benefits: Many variables**

A third benefit of tables is showing corresponding values for a larger number of variables than typically possible on a chart.

The table below shows annual expenditures on a child, but for three different income groups. A grouped bar chart would have had 18 items on the x-axis of Age, with 3 bars per item, for 48 bars total. So many bars could be overwhelming. A scatterplot can show corresponding values for two variables, but more is tricky. The table also includes a Year variable to help the viewer predict future expenses, which was easily added to the table. A line chart might do fine showing the data using three lines for each of Lowest, Middle, and Highest income groups, though the x-axis would need to have two values per increment and would need a way of showing the total.

A fourth benefit is that a user can copy-paste data from a table into a spreadsheet, enabling the user to do calculations, generate plots, etc.

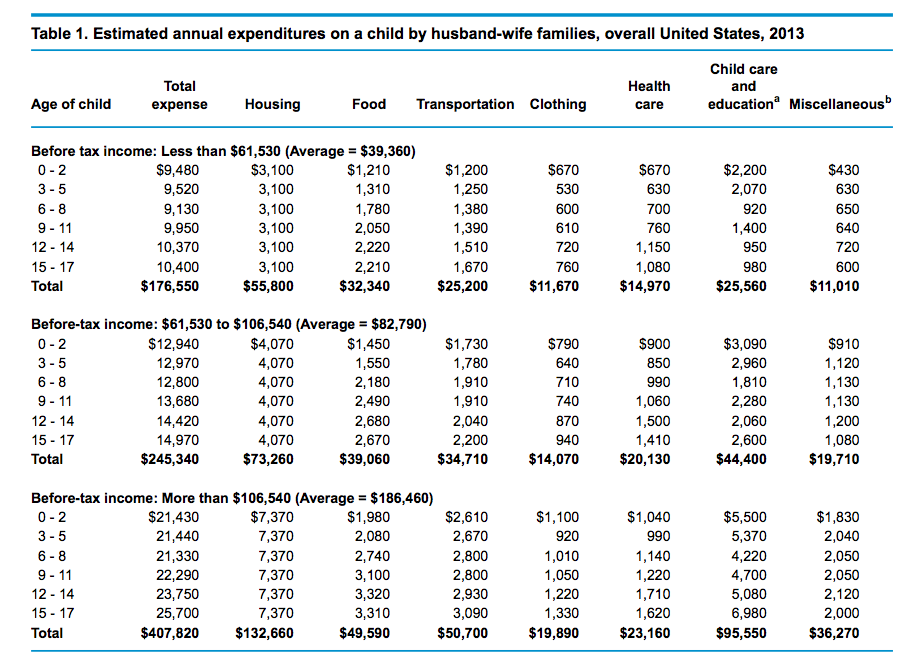
Figure 8.9.1: A table easily displays more variables than does a chart.

Source: [U. S. Department of Agriculture, 2013](http://www.cnpp.usda.gov/sites/default/files/expenditures_on_children_by_families/crc2013.pdf)

**Feedback?**

A fifth benefit of tables is the ability to group variables, group values, and introduce text throughout to aid the reader. The table below illustrates.

Figure 8.9.2: A table with groupings and text: Annual expenditures on a child.

Source: [U. S. Department of Agriculture, 2013](http://www.cnpp.usda.gov/sites/default/files/expenditures_on_children_by_families/crc2013.pdf)

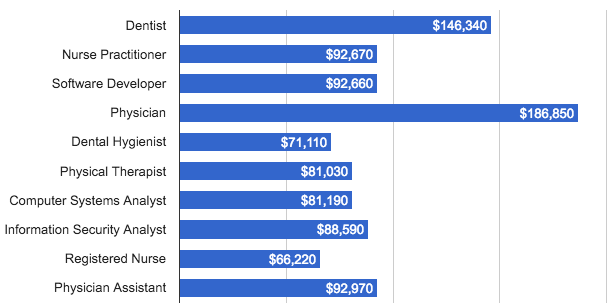
# 8.10 Data and chart guidelines

Several guidelines can help in creating effective charts.

### **Using appropriate precision**

***Precision*** is the number of digits treated as significant in a number. Ex: Pi can be written as 3.14 to 3 digits of precision, or as 3.14159 to 6 digits of precision. A common mistake is to include more precision than necessary to achieve a desired information conveyance goal. The figure below rounds to the nearest 10, which is better than rounding to the nearest 1, but even better might have been rounding to the nearest 100 (or even 500's).

Figure 8.10.1: Median salary of the 10 best jobs of 2015 from U.S. News.

Source: [U.S. News, 2015](http://money.usnews.com/money/careers/slideshows/the-25-best-jobs-of-2015)

**Feedback?**

A related mistake is to include more precision than necessary on a plot's axes, as in the below figure which shows y-axis values as 10.0%, 20.0%, ..., rather than just 10%, 20%, ... This mistake has become common due in part to Microsoft Excel's chart tool inheriting by default the precision of the displayed axis values from the data's precision. A good data presenter remembers to adjust the axis settings for plots.

Figure 8.10.2: Unnecessary precision in the y-axis labels, due to Excel's chart tool's default.

[Source: Quora 2012](https://www.quora.com/What-percentage-of-airline-revenue-is-made-from-business-and-first-class-tickets)

Of course, the converse of the above, namely excessive rounding, can be a problem. Excessive rounding may lose important information. In the salary data above, presenting the salaries as 145,000, 90,000, 90,000, 185,000, etc., loses too much information. In deciding the amount of precision to present, the key is to find the right balance.

### **Shifting y-axis values**

To make differences in values more visible, presenters sometimes skip lower values on the y-axis, shifting the higher values down. On a bar chart, such adjustment distorts the bar lengths, which misleads the viewer into believing the differences are greater than in reality. Such adjustment is especially common when the presenter wants the differences to be perceived as large, such as when criticizing a person or organization (like a country, as below). Such distortion is often inappropriate. Usually, the y-axis should not skip values. The following animation illustrates how skipped y-axis values exaggerate the differences in male/female birth ratios for various countries (Source: [The World Bank](http://databank.worldbank.org/data/reports.aspx?source=gender-statistics) ).

**PARTICIPATION ACTIVITY**

8.10.2: A common error is to adjust the y-axis to accentuate differences. Keeping the axis standard is usually preferable.

**12**

2x speed

1.0

1.1

1.2

North America

Latin America & Caribbean

European Union

Arab World

China

India

Male/female birth ratio, 2012

1.0

North America

Latin America & Caribbean

European Union

Arab World

China

India

Male/female birth ratio, 2012

1.0

1.0

1.0

1.1

1.1

1.2

*Heights don't accurately*

*depict relative differences*

Usually, the y-axis scale should be kept standard, so that bar lengths accurately depict values. Data labels can be added to precisely convey differences.

##### Captions

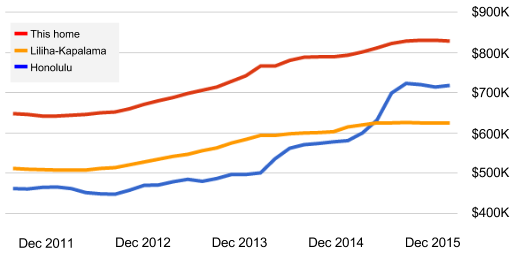
keyboard\_arrow\_up

1. Presenters sometimes skip y-axis values to make differences more visible. But such skipping misleads the viewer into seeing the differences as larger.
2. Usually, the y-axis scale should be kept standard, so that bar lengths accurately depict values. Data labels can be added to precisely convey differences.

**Feedback?**

Skipping of y-axis values may be acceptable when relative heights/lengths are not important, but rather conveying actual values are the main goal. Ex: The popular house-pricing website Zillow shows a house's price-estimate history, along with median prices for the region and city, skipping lower y-axis values so the prices are more visible. The figure below illustrates. The viewer is mostly interested in the house's prices, and not the relative differences with the region and city prices, which are just provided for convenient reference. Skipping lower y-axis values is likely appropriate in this case.

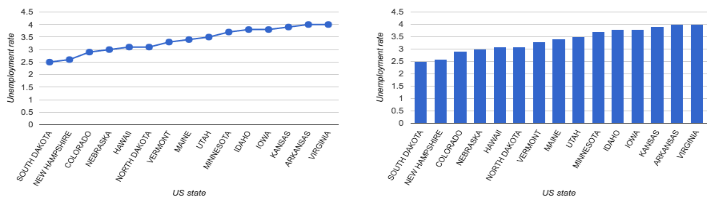
Figure 8.10.3: Skipping y-axis values is OK if actual values, and not relative differences, are the goal.



### **Line charts are not for nominal categorical data**

A line chart should not be used for nominal categorical data. Lines suggest some relation from one item to the next, but nominal variables have no ordering so can have no such relation. Ex: The plot below on the left inappropriately shows lines, even though no relationship exists between South Dakota and New Hampshire (the left two points), for example. Showing just dots for each data point would be OK. A bar chart, as on the right, is also common.

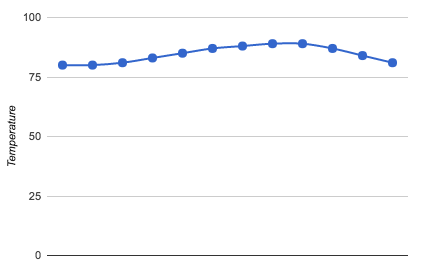
Figure 8.10.4: A line chart is not appropriate for categorical data on the horizontal axis.



### **Always including axis labels and units**

Common mistakes are to forget to label an axis or to forget to provide units. Without such information, a viewer cannot appropriately interpret the information.

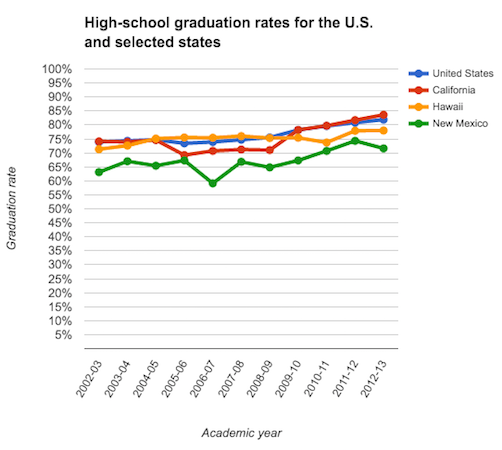
Figure 8.10.5: Average monthly temperatures in Hawaii, with some labels/units missing.



### **Removing unnecessary items**

A good step (perhaps the last step) in designing a visualization is to remove unnecessary items. Ex: In the figure below, the chart title is redundant with the figure caption, so can be removed. Also, the 5% increments are too detailed and can be removed as well. Even the dots at the data points might be removed.

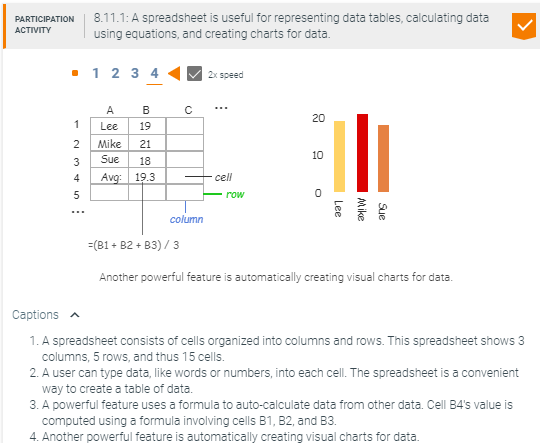
Figure 8.10.6: U.S. and selected state public high school graduation rates, 2002-2013. Unnecessary info exists.



# 8.11 Spreadsheets

### **Basics**

A ***spreadsheet application*** is a common computer application for representing tables of data like text or numbers, for using formulas to calculate data from other data (like sums or averages), and for creating visual charts from data. A spreadsheet contains ***cells*** organized into columns labeled A, B, ... and rows labeled 1, 2, ...; a spreadsheet user can type data in each cell.



This section demonstrates spreadsheets using Google Sheets. Most features behave similarly in Microsoft Excel.

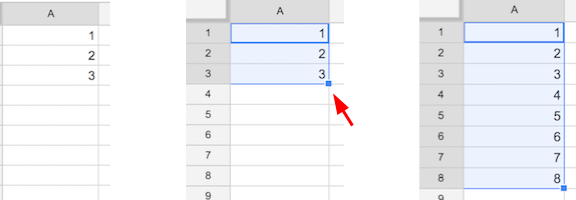
### **Autofill**

***Autofill*** is a useful spreadsheet feature that recognizes a pattern and fills additional pattern values.

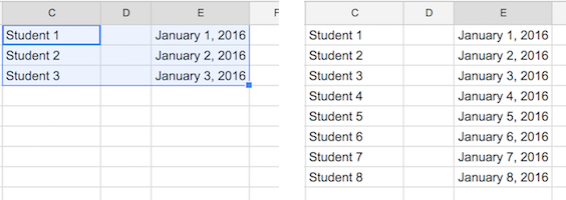
Spreadsheet Practice 8.11.1: Autofill.

The following table, captured in Google Sheets, specifies the start of a pattern.

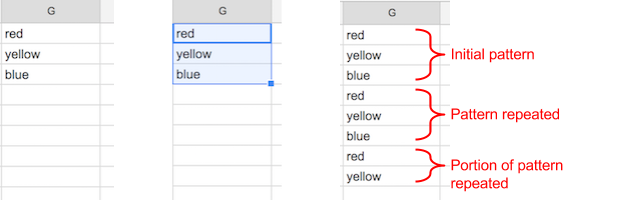
Autofill cells: Enter the first few values of a pattern. Highlight the cells. Drag the blue box, appearing in the bottom right, to the desired number of cells.



Patterns may include text, numbers, or dates, among others.



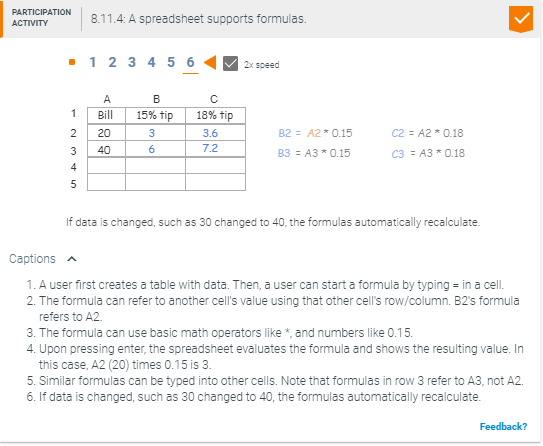
If a pattern is not recognized, the autofill feature repeats the pattern provided.



Spreadsheets can be quite clever in detecting patterns. Ex: Starting with two cells having Jan and Feb, Google Sheets will fill in Mar, Apr, May, etc. (Excel doesn't detect the pattern, though).

### **Spreadsheet formulas**

An important spreadsheet feature allows a user to type a ***formula*** in a cell to compute that cell's value based on other cells' values. A formula begins with = followed by a math expression using operators like +, -, \*, /, and parentheses (). Another cell's value can be included using that cell's row letter and column number, as in: = 3 \* A1.



### **Spreadsheet functions**

A spreadsheet ***function*** is a predefined formula that supports common tasks such as computing the average, minimum, or maximum of a group of cells. Spreadsheets commonly support a number of functions encompassing engineering, statistics, finance, and other applications. Users can also define custom functions, which is an advanced topic that is not discussed in this section.

Table 8.11.1: Common spreadsheet functions.

|  |  |  |
| --- | --- | --- |
| Function name | Description | Function syntax |
| ABS | Returns the absolute value of a number. | ABS(value) |
| AVERAGE | Returns the numerical average value in a dataset, ignoring text. | AVERAGE(value1, [value2, …]) |
| COUNT | Returns a count of the number of numeric values in a dataset. | COUNT(value1, [value2, …]) |
| MAX | Returns the maximum value in a numeric dataset. | MAX(value1, [value2, …]) |
| MIN | Returns the minimum value in a numeric dataset. | MIN(value1, [value2, …]) |
| MODE | Returns the most commonly occurring value in a dataset. | MODE(value1, [value2, …]) |
| SUM | Returns the sum of a series of numbers and/or cells. | SUM(value1, [value2, …]) |

**Feedback?**

Above, the ***function syntax*** defines how the function is used, and specifies the function's name and accepted arguments. A ***function's arguments*** are surrounded by parentheses and specify the data that the function operates on. Arguments may be numbers, cells, a range of cells, or a combination thereof. The [ ] arguments are optional.

To call a function in a spreadsheet, = is followed by the function's name and then arguments separated by commas. Ex: =SUM(A1, A2, A3) calculates sum of cells A1, A2, and A3. The ***range operator*** (***:***) defines a reference to a group of cells. Ex: =SUM(A1:A4, B10) calculates the sum of cells A1, A2, A3, A4, and B10.

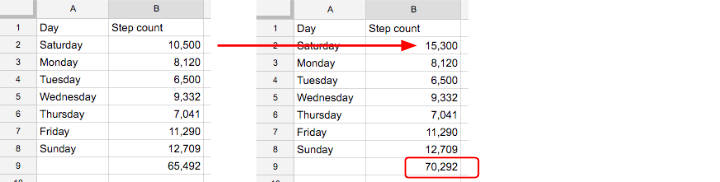
Spreadsheet Practice 8.11.2: SUM function usage.

The following table, captured in Google Sheets, lists a person's 7-day step count. The SUM function can be used to calculate the step count for the week.

Computing sum: Enter =SUM() into cell B9 and place the cursor inside of the parentheses. Click on cell B2 and drag to B8, then press enter. Instead of clicking and dragging, SUM's argument can also be defined directly by entering =SUM(B2:B8).



The SUM function automatically recalculates the sum if any of the data is changed.



### **Creating basic charts**

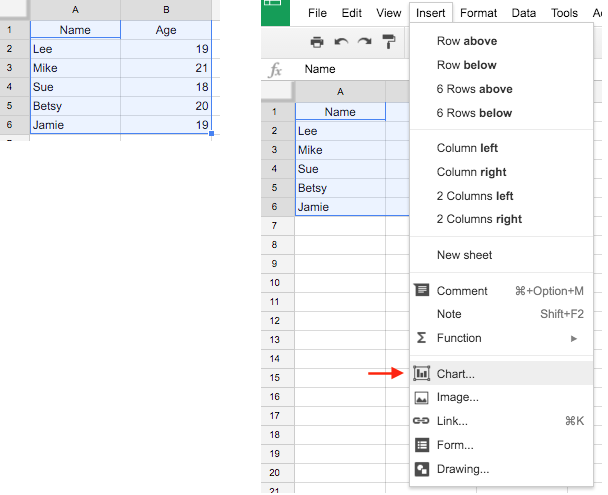
Automatically creating visual charts for data is a powerful spreadsheet feature.

Spreadsheet Practice 8.11.3: Creating a bar chart.

The following table, captured in Google Sheets, contains the data used to create a table.

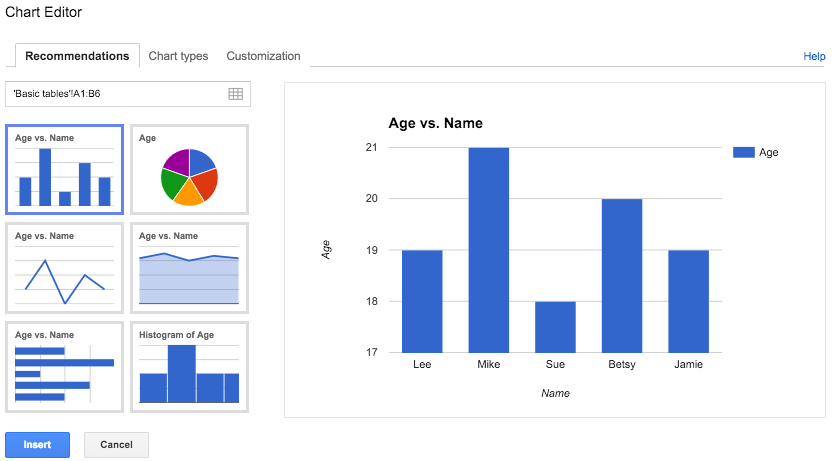
Create a table: Highlight the cells containing the chart labels and data. The first row contains the category and value labels. Each subsequent row contains a student's name and age.

Select Insert, and then Chart.

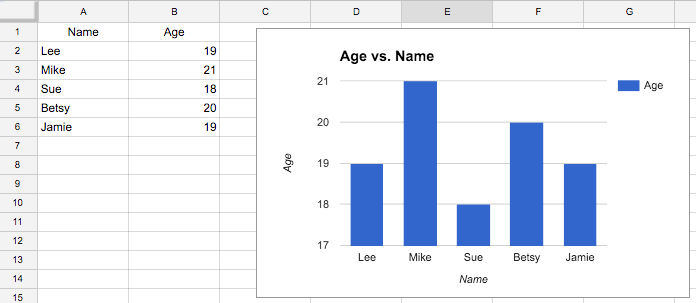


The Chart Editor provides an interface to customize various chart features such as chart type, chart title, font, coloring, and more. Upon selecting bar chart, the tool creates several items by default, like the title, legend, axis labels, axis numbers, etc.

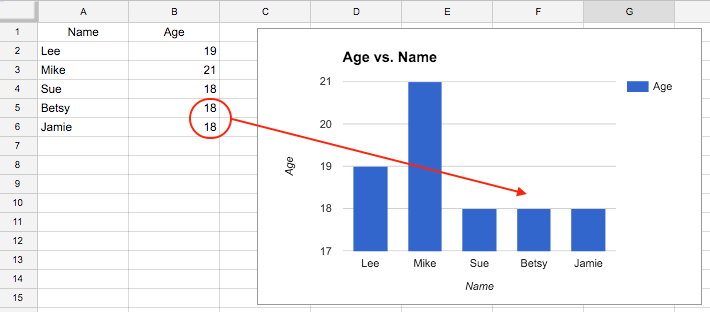
Click Insert.



Each row containing data is shown as a bar in the bar chart.



The bar chart is automatically updated if any of the data is changed.



A user commonly should change many of the default items. Above, the y-axis might be adjusted to start from 0 (to visually depict the relative ages accurately), and the legend and title might be removed (to remove unnecessary items).

# 8.12 Spreadsheet plotting

### **Creating a scatter plot**

Scatter plot data consists of value pairs for two variables, where each pair becomes a coordinate that is plotted on a rectangular coordinate system. If variable values are non-negative, then typically only the upper-right quadrant is shown.

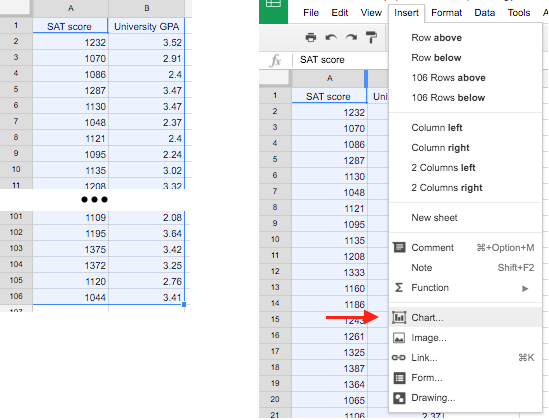
Spreadsheet Practice 8.12.1: Creating a scatter plot.

The following table, captured in Google Sheets, contains the data used to create a scatter plot. The table lists the SAT score (Math and Verbal) and the corresponding GPA for 105 graduating students (Source: [Online Statistics Education](http://onlinestatbook.com/2/case_studies/sat.html)). The first row contains the category labels. Each subsequent row contains a value pair for SAT score and University GPA.

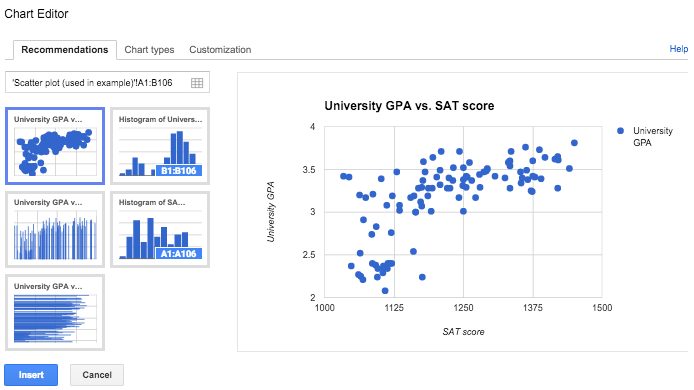
Create a scatter plot:

1. Highlight the cells containing the chart data.

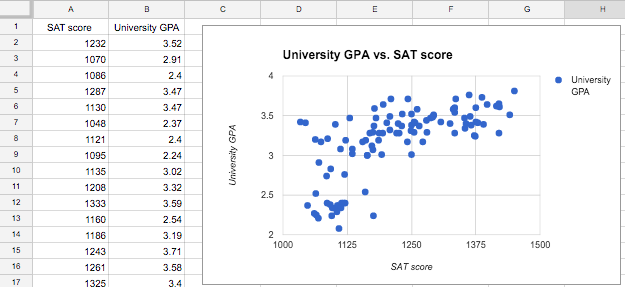
2. Select Insert, and then Chart.



3. Click Insert. (Note: A scatter plot type is already selected by default.)



Each row containing data is shown as a dot in the scatter plot.

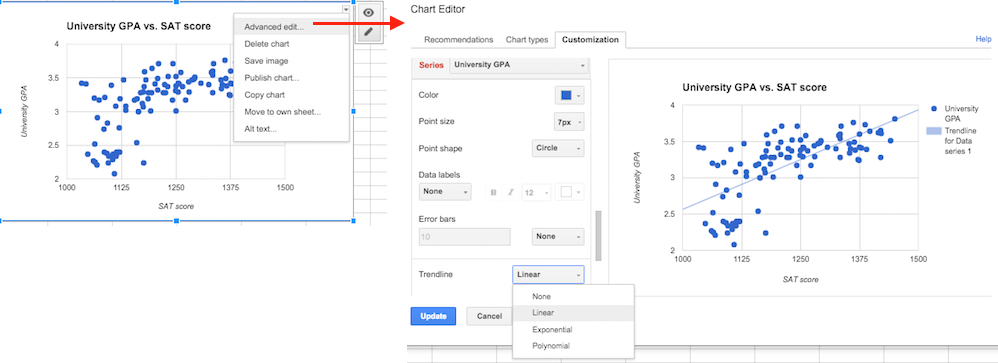


Add a trend line:

4. Click on the scatter plot, then click on the downwards facing triangle icon in the upper right.

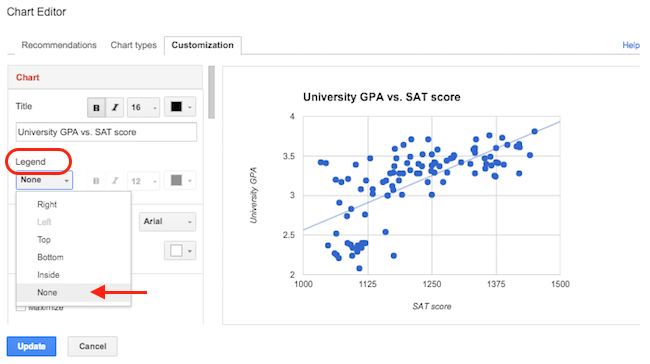
5. Click on Advanced edit ... to launch the Chart Editor.

6. A drop down Trendline option is available at the bottom of the Customization tab. Click on Linear, and then Update.



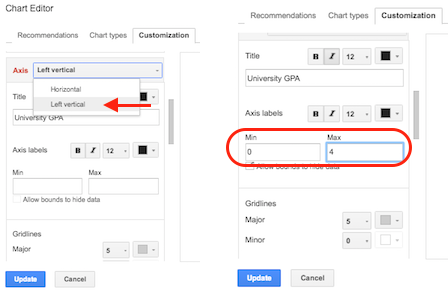
Remove legend and change y-axis labels:

7. A drop down Legend option is available under the Customization tab. Select None.

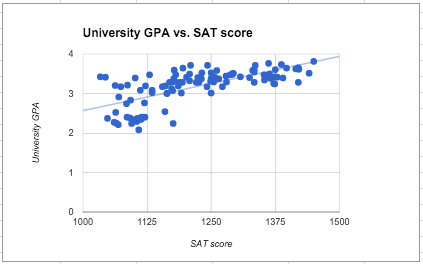


8. Select Left vertical under the Axis drop down menu to display the y-axis options.

9. Type 0 in the Min text field and 4 in the Max text field. Click Update.



The resulting scatter plot's y-axis begins at 0 and the legend no longer appears to the right.



**Feedback?**

A trend line is added to help visualize the general relationship of two variables. The Chart Editor supports a linear, exponential, or polynomial trend line. Once a trend line is added, additional configuration options such as the line color, opacity, and thickness appear in the Customization tab.

Figure 8.12.1: Additional trend line options in Google Sheets.



### **Creating a line chart**

A line chart uses straight lines to connect subsequent data points in a scatter plot. The straight lines can help visualize the general direction that data changes as x values increase, which is especially useful when the x-axis represents time.

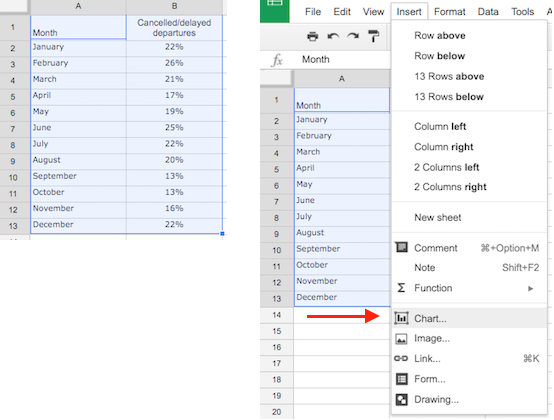
Spreadsheet Practice 8.12.2: Creating a line chart.

The following table, captured in Google Sheets, contains the data used to create a line chart. The table provides the percentage of cancelled or delayed flights for each month in 2016 (Source: [U. S. Department of Transportation](http://goo.gl/Aqn464)). The first row contains the category and value labels. Each subsequent row contains the month and the percentage of cancelled or delayed flights for that month.

Create a line chart:

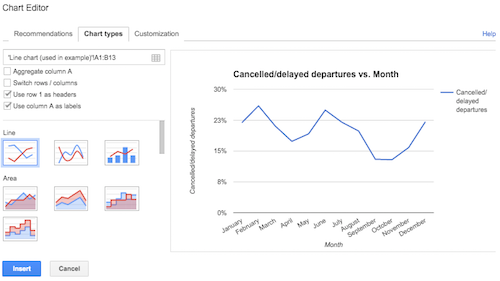
1. Highlight the cells containing the chart data.

2. Select Insert, and then Chart.

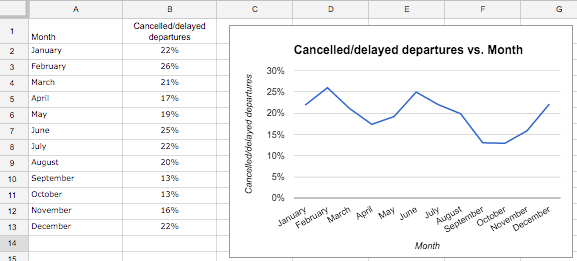


3. Click on the Chart type tab to expose all chart types supported by the Chart editor.

4. Click on the first line chart icon to select a line graph that connects subsequent values with a straight line, then click on Insert.



A straight line connects neighboring data values. The default y-axis labels are not intuitive and can be changed to 5% increments. Similarly, the default legend does not aid in comprehension and can be hidden.

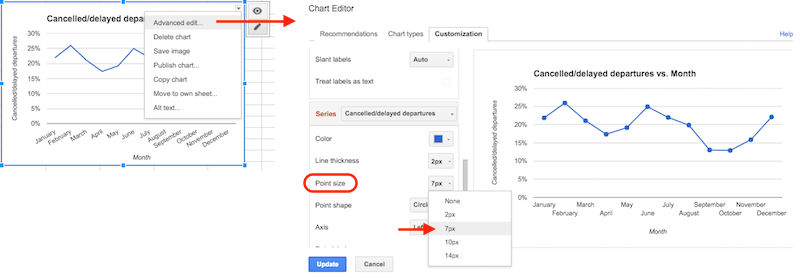


Add data points:

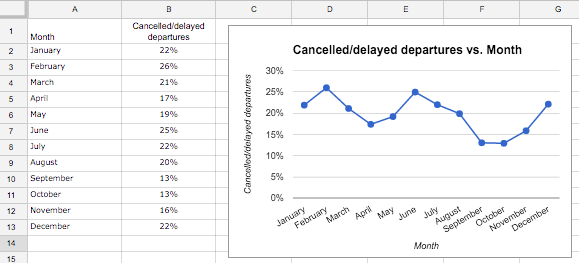
5. Click on the line chart, then click on the downwards facing triangle icon in the upper right.

6. Click on Advanced edit ... to launch the Chart Editor.

7. A drop down Point size option is available in the Customization tab. Click on 7px, and then click Update.



Each data point is shown as a dot on the line chart.



### **Creating a pie chart**

A pie chart is a common chart type used to illustrate the relative frequency for categories using a circle. Each category is shown as a slice, where the size of a slice is proportional to the size of the data as a whole. Data values must be positive numbers and can contain decimal values or percentages.

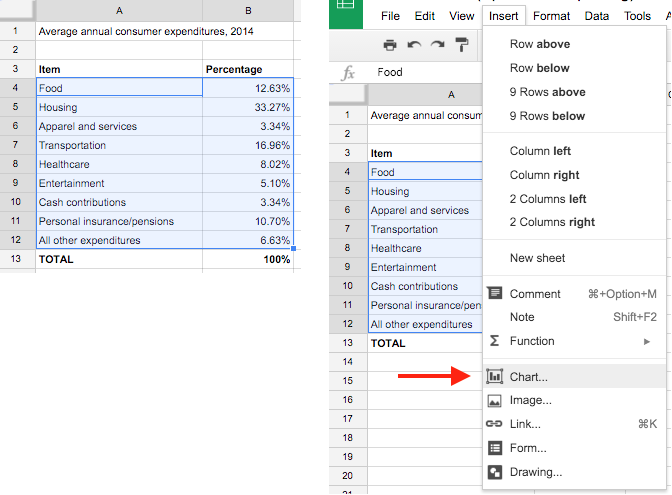
Spreadsheet Practice 8.12.3: Creating a pie chart.

The following table, captured in Google Sheets, contains the data used to create a pie chart. The table provides a breakdown of average consumer expenditures for various categories such as food, housing, and healthcare (Source: [Bureau of Labor Statistics, 2014](http://www.bls.gov/news.release/pdf/cesan.pdf)). The first row contains the category and value labels. Each subsequent row contains a spending category name and the corresponding percentage spent by a consumer in that category.

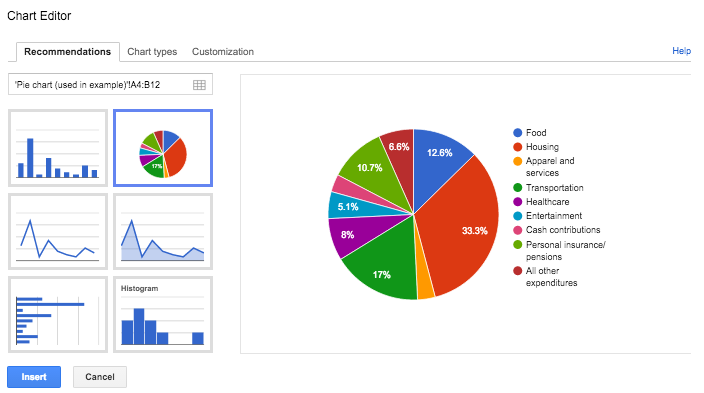
Create a pie chart:

1. Highlight the cells containing the chart data. (Note: The last row contains the total percentage and is omitted from the highlighted cells, as the total should not be included as a slice in the pie chart.)

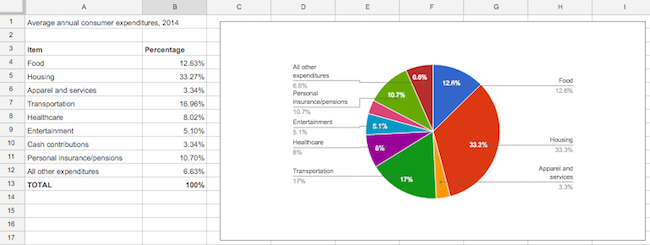
2. Select Insert, and then Chart.



3. Select the pie chart type from the Recommendations tab. Click Insert.



Each of the selected rows appear as a slice in the pie chart. The default legend to the right does not aid in comprehension. The chart legend option can be changed from "Right" to "Labeled" to provide direct labeling of each slice.

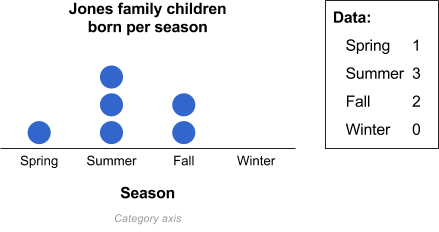


# 8.13 Dot plots

### **Basic dot plots**

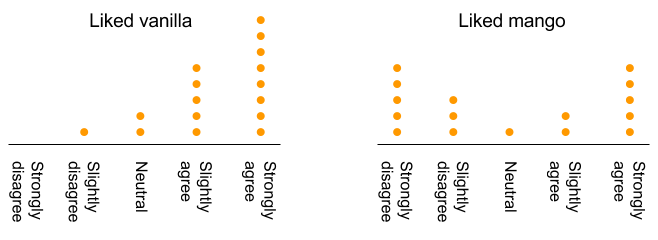
A ***dot plot*** uses multiple dots to indicate a category's value. Ex: If category Summer has value 3, the dot plot would show 3 dots. The plot commonly only labels the category-axis. Dot plots are suitable when category values can be counted with a quick glance.

Figure 8.13.1: Dot plot example.



Example 8.13.1: Survey results.

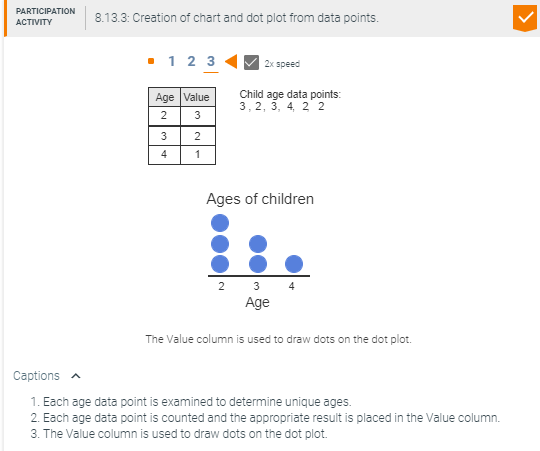
A dot plot can quickly convey much information about data. For example, a company may ask 16 customers to complete a survey, indicating agreement with the statement "I liked the mango frozen yogurt", with options "Strongly disagree", "Slightly disagree", "Neutral", "Slightly agree", and "Strongly agree". The agreement variable is an ordinal categorical variable. (Good practice in surveys use such an agreement statement and scale, known as a [Likert scale](https://en.wikipedia.org/wiki/Likert_scale)). With a dot plot summary, the user quickly (preattentively) gains information about how well liked was the product, and how spread was the response. Below, the observer quickly notices that most customers liked the vanilla yogurt, while customers were split on the mango yogurt.



**Feedback?**

### **Dot plots when category labels are numbers**

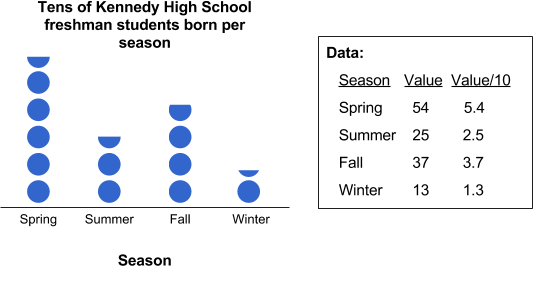
Sometimes category labels are numbers. Because values are also numbers, categories and values may be easily confused. Ex: Categories may be children's ages like 2, 3, and 4, and values may be the number of children in each category like 3, 2, and 1. To avoid confusing category labels and values, a simple chart can be made before drawing the dot plot. Each instance of a value is known as a ***data point***. Ex: 9, 8, 9, 5 has four data points. A category's value is the number of data points matching that category.



### **Condensing dots**

The brain has trouble quickly processing information on a dot plot with more than about 10 dots in a category. In such a case, a designer may condense points by dividing the number of data points in each category by a constant number. Ex: Every 10 data points may be drawn as 1 dot. If dividing yields a fraction, in some situations partial dots may be drawn, while in other situations rounding may be preferred.

Figure 8.13.2: Condensing data points.



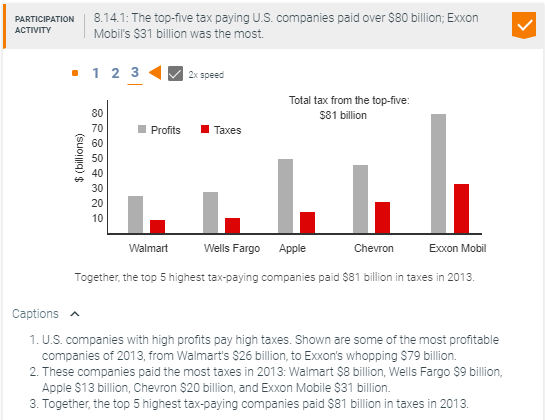
# 8.14 Animations

An ***animation*** is a dynamic figure, meaning a figure whose parts may appear, disappear, and/or move. While past data visualization has focused on static figures for paper, the ubiquity of computers and the web make animations more common, via videos, presentation applications like Microsoft PowerPoint or Google Slides, or animated web pages. In fact, the reader may notice that this material itself makes extensive use of animations.

Key principles that guide the design of animations for data visualization include:

* Fading-in distinct parts individually, explaining each part, to walk the viewer towards a goal.
* Using motion to catch the viewer's attention for key points.
* Minimize fading-out, so that the final figure has meaning when viewed statically.
* Avoid gratuitous animating; each fade-in and motion should have a purpose. Other items can be static.

Below is a static view of a bar chart whose information conveyance goal is to show that profitable companies contributed substantial taxes to the U.S. government: $81 billion in 2013. Upon pressing Start, a first step presents profits for those companies. Then, a second step presents taxes paid. Finally, a third step shows each company's taxes being added to a total of $81 billion. Walking through the steps may aid the reader in understanding the chart. ([Source: USA Today](http://www.usatoday.com/story/money/personalfinance/2013/03/17/companies-paying-highest-income-taxes/1991313/)).



# 

# 

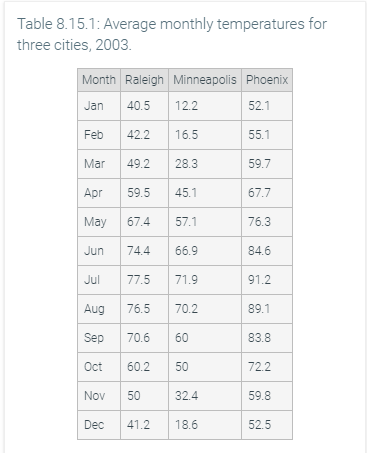
# 

# 

# 

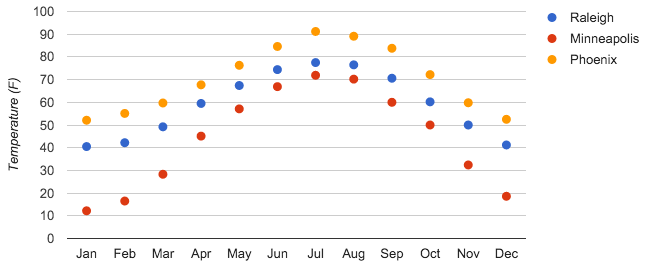
# 8.15 Data visualization: Case study

The table below provides average monthly temperature data for three U.S. cities (Source: [SAS tool example](http://www.u.arizona.edu/~pjones/edp548/overheads/sas/graphs.html)



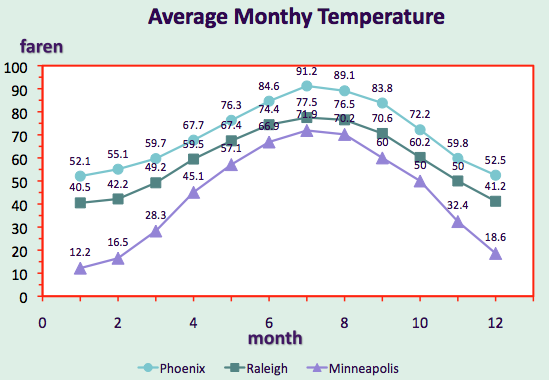
An analyst sought to depict the data using a plot. The analyst might consider a scatter plot, as shown below.

Figure 8.15.1: Scatter plot for three cities' temperature data.



Another analyst believes the scatter plot's individual data points are visually distracting, so decides to use a line chart. Also, the analyst believes the viewer cares about precise temperature values, so adds data labels. The analyst creates the line chart below.

Figure 8.15.2: Average monthly temperature for three cities. Superfluous information distracts from the chart's information conveyance goal.



A third analyst believes a hybrid line chart / table would be better. The analyst believes the line chart above has several problems: (1) Overly complex: data labels and data point symbols overwhelm the chart, (2) the x-axis should indicate months by names, not numbers, (3) the legend requires the viewer to jump between the chart and the legend.

The analyst creates the hybrid line chart / table below. (1) The line chart is simplified, clearly showing trends and relative values. (2) The x-axis uses months by names. (3) Each line is directly labeled, avoiding a legend. Also, the analyst put a table directly below to provide the viewer with precise values. The chart and table cleverly use the month names both for the x-axis labels and the table column headers.

Figure 8.15.3: Average monthly temperature for three cities. A line graph and table are used to illustrate temperature trends while providing data values.

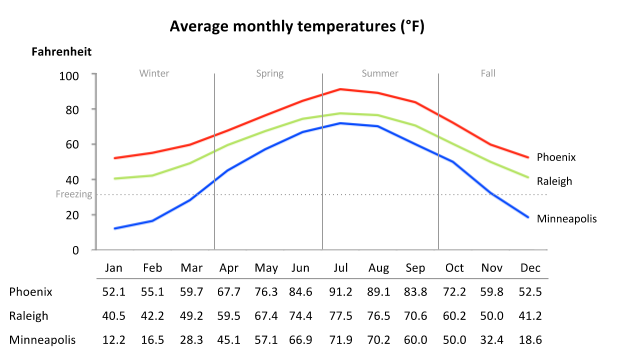


**Feedback?**

The analyst thinks more about what viewers would be interested in, and adds lines and labels to indicate seasons and freezing point, as below.

Note that the analyst went beyond what a spreadsheet can auto-generate, for the above and below figures.

Figure 8.15.4: Average monthly temperature for three cities. The chart is annotated with seasons and a freezing point reference.



# 8.16 Dashboards

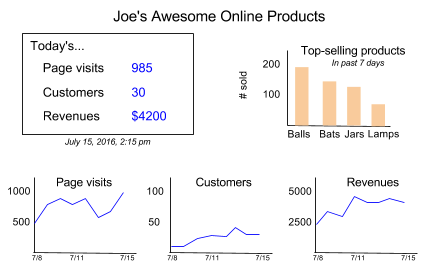


In the context of an organization, a ***dashboard*** displays an organization's key real-time data. ***Real-time data*** is data that was generated recently, such as in the past seconds, minutes, hours, or days. The term dashboard comes from cars, where the key real-time data includes speed, fuel level, temperature, etc.

For an organization, the dashboard typically displays the key real-time data for that organization's operations. Ex: For a web-based business, key real-time data may include that day's number of webpage visits, customers, and revenues. Historical data is commonly shown too, such as daily data for the past week. Non-real-time data typically isn't displayed, such as number of employees at the business (which changes infrequently). Non-key data, even if real-time, also isn't displayed, such as number of purchases using credit vs. debit card.

A dashboard combines data visualization techniques, like tables, line charts, bar charts, etc. A dashboard designer must first determine what is the key real-time data to present, and then how to best visualize that data.

Figure 8.16.1: Dashboard for a web-based business.



Example 8.16.1: Stock market dashboard.

Many people commonly see dashboards relating to financial markets. The dashboard below show's CNN's page summarizing the day's market activity. The table at the top provides values for the main three stock markets, focusing on the % change, but also showing the actual values. Below that is a bar chart for the most popular stock of the day, showing one (extremely narrow) bar per minute, each bar indicating the stock's value. (Care must be taken to notice that the vertical axis skips values). The dashboard uses red text to indicate decreases and green text for increases, to provide a quick sense of which items are going up or down.

