https://chartio.com/learn/charts/what-is-a-scatter-plot/
https://calcworkshop.com/functions-statistics/line-best-fit/

https://towardsdatascience.com/understanding-boxplots-5e2df7bcbd51

In previous sections we discussed various charts. This section will cover plots (Some time also referred as charts only)

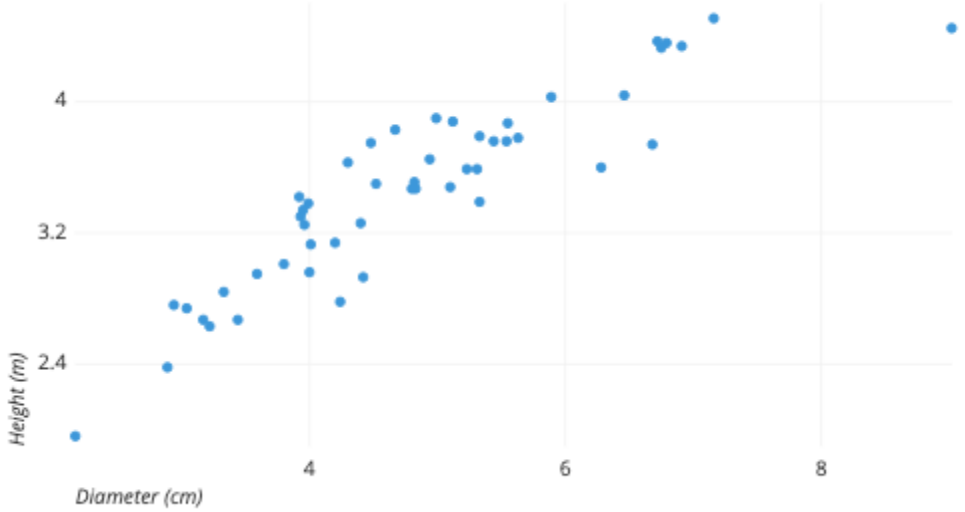
A. **Scatter plot (Scatter Chart)**
B. **Bubble Plot (Chart)**
C. **Box Plot**
D. **Time line Plot chart**

A. Scatter Plots

A scatter plot (Scatter Chart, Scatter Graph) uses dots to represent values for two different numeric variables. The position of each dot on the horizontal and vertical axis indicates values for an individual data point. Scatter plots are used to observe relationships between variables.

Example of data structure

DIAMETER	HEIGHT
4.20	3.14
5.55	3.87
3.33	2.84
6.91	4.34
…	…

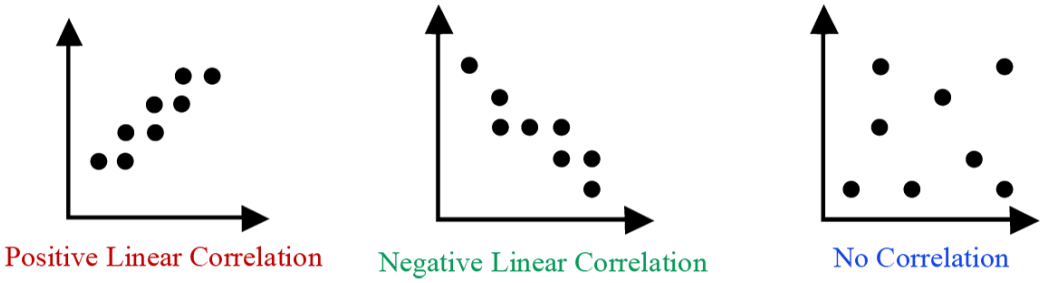


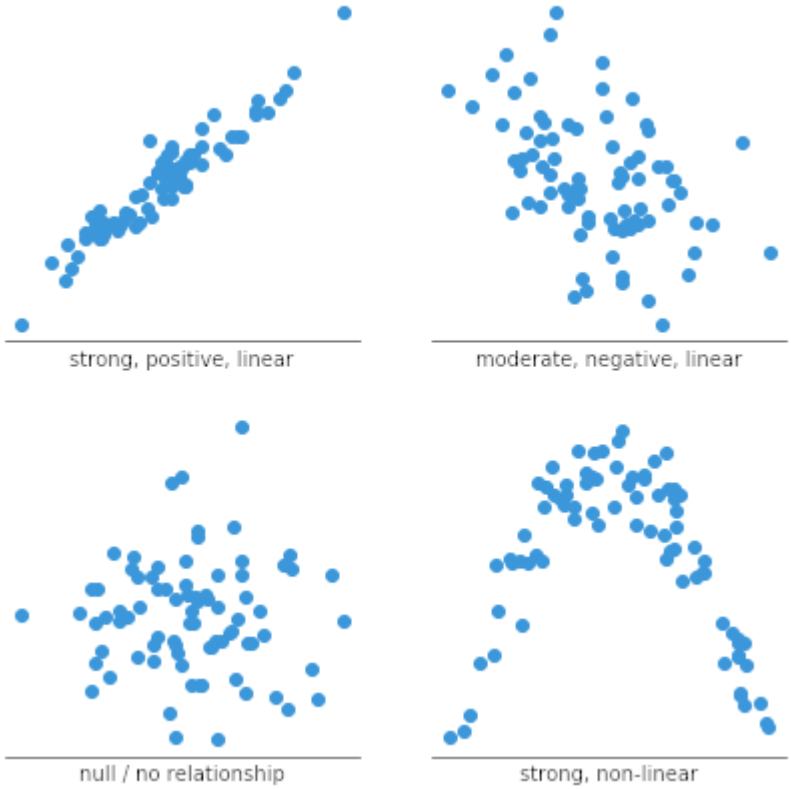
The example scatter plot above shows the diameters and heights for a sample of fictional trees. Each dot represents a single tree; each point's horizontal position indicates that tree's diameter (in centimeters) and the vertical position indicates that tree's height (in meters). From the plot, we can see a generally tight positive correlation between a tree's diameter and its height. We can also observe an outlier point, a tree that has a much larger diameter than the others. This tree appears fairly short for its girth, which might warrant further investigation.

Usage of Scatter plot

1. **Observe and show relationships between two numeric variables:** The dots in a scatter plot not only report the values of individual data points, but also patterns when the data are taken as a whole.

2. **Identification of correlational relationships**: In these cases, we want to know, if we were given a particular horizontal value, what a good prediction would be for the vertical value. You will often see the **variable on the horizontal axis denoted an independent variable, and the variable on the vertical axis the dependent variable**. Relationships between variables can be described in many ways: positive or negative, strong or weak, linear or nonlinear.





3. **To identify other patterns in data** : We can divide data points into groups based on how closely sets of points cluster together. Scatter plots can also show if there are any unexpected gaps in the data and if there are any outlier points. This can be useful if we want to segment the data into different parts, like in the development of user personas.

In order to create a scatter plot, we need to select two columns from a data table, one for each dimension of the plot. Each row of the table will become a single dot in the plot with position according to the column values.

Common issues when using scatter plots

1. Over plotting

When we have lots of data points to plot, this can run into the issue of over plotting. It is the case where data points overlap to a degree where we have difficulty seeing relationships between points and variables. It can be difficult to tell how densely-packed data points are when many of them are in a small area.

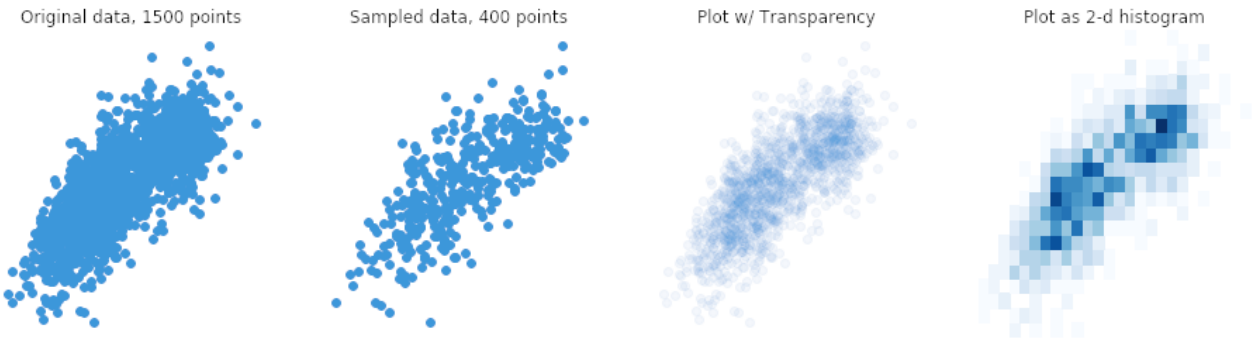
Solution for Over Plotting

i. Sample only a subset of data points: a random selection of points should still give the general idea of the patterns in the full data.

ii. Change the form of the dots, adding transparency to allow for overlaps to be visible,

iii. Reduce point size so that fewer overlaps occur.

iv. Use different chart type like the Heat map (where color indicates the number of points in each bin. Heat maps in this use case are also known as 2-d histograms)



2. Interpreting correlation as causation

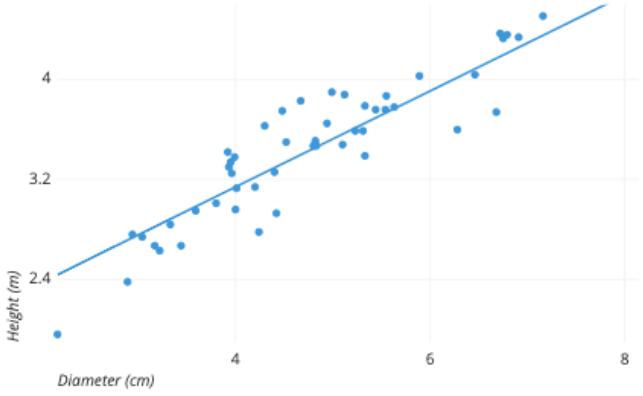
This is not so much an issue with creating a scatter plot as it is an issue with its interpretation. Simply because we observe a relationship between two variables in a scatter plot, **it does not mean that changes in one variable are responsible for changes in the other.** This gives rise to the common phrase in statistics that correlation does not imply causation. It is possible that the observed relationship is driven by some third variable that affects both of the plotted variables, that the causal link is reversed, or that the pattern is simply coincidental.

For example, it would be wrong to look at city statistics for the amount of green space they have and the number of crimes committed and conclude that one causes the other, this can ignore the fact that larger cities with more people will tend to have more of both, and that they are simply correlated through that and other factors. If a causal link needs to be established, then further analysis to control or account for other potential variables effects needs to be performed, in order to rule out other possible explanations. There are option to enhance the interpretation.

- **Common scatter plot options**

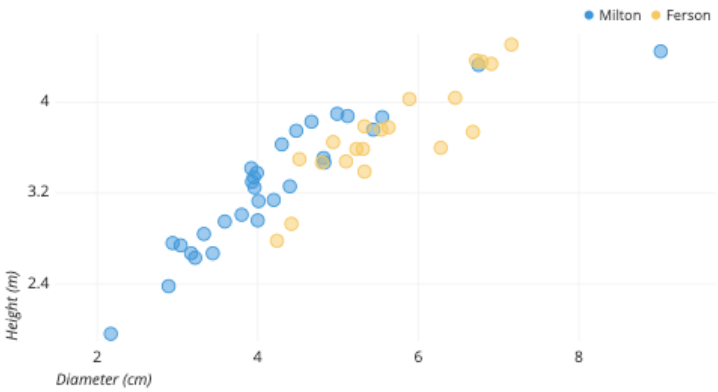
  i. Add a trend line

  When a scatter plot is used to look at a predictive or correlational relationship between variables, it is common to add a trend line to the plot showing the mathematically best fit to the data. This can provide an additional signal as to how strong the relationship between the two variables is, and if there are any unusual points that are affecting the computation of the trend line.



  ii. Categorical third variable

  A common modification of the basic scatter plot is the addition of a third variable. Values of the third variable can be encoded by modifying how the points are plotted. For a third variable that indicates categorical values (like geographical region or gender), the most common encoding is through point color. Giving each point a distinct hue makes it easy to show membership of each point to a respective group.



Coloring points by tree type,

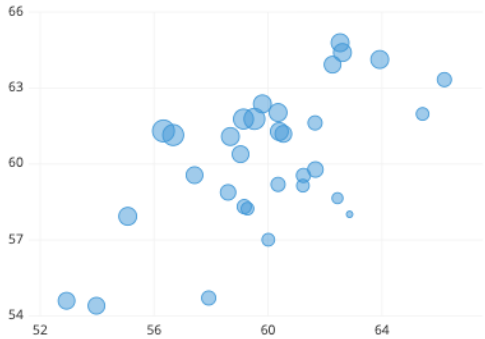
From Graph Fersons (yellow) are generally wider than Miltons (blue), but also shorter for the same diameter.

iii. Different shapes as Third variable:

One other option that is sometimes seen for third-variable encoding is that of shape. One potential issue with shape is that different shapes can have different sizes and surface areas, which can have an effect on how groups are perceived. However, in certain cases where color cannot be used (like in print), shape may be the best option for distinguishing between groups.

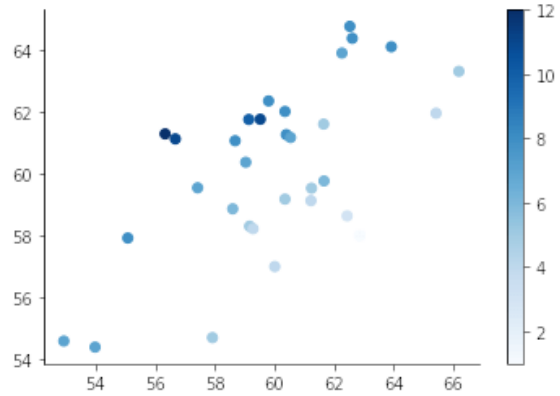
iv. Numeric third variable

- **Size:** For third variables that have numeric values, a common encoding comes from changing the point size. A scatter plot with point size based on a third variable also known as bubble chart. Larger points indicate higher values. (Detail of Bubble chart in following discussion)



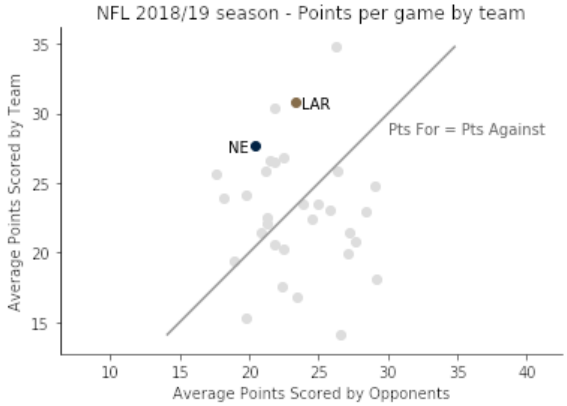
- **Hue:** Rather than using distinct colors for points like in the categorical case, we want to use a continuous sequence of colors, so that, for example, darker colors indicate higher value. Note

that, for both size and color, a legend is important for interpretation of the third variable, since our eyes are much less able to discern size and color as easily as position.



v. Highlight using annotations and color

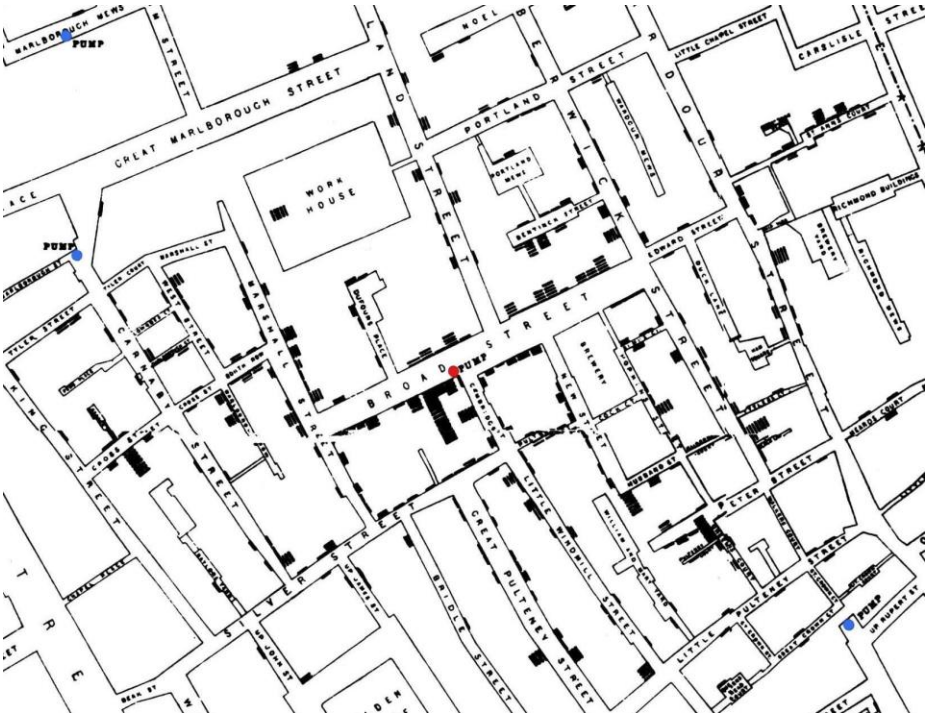
If you want to use a scatter plot to present insights, it can be good to highlight particular points of interest through the use of annotations and color. Desaturating unimportant points makes the remaining points stand out, and provides a reference to compare the remaining points against.



- ***Other related representation of scatter plot***

i. Scatter Map

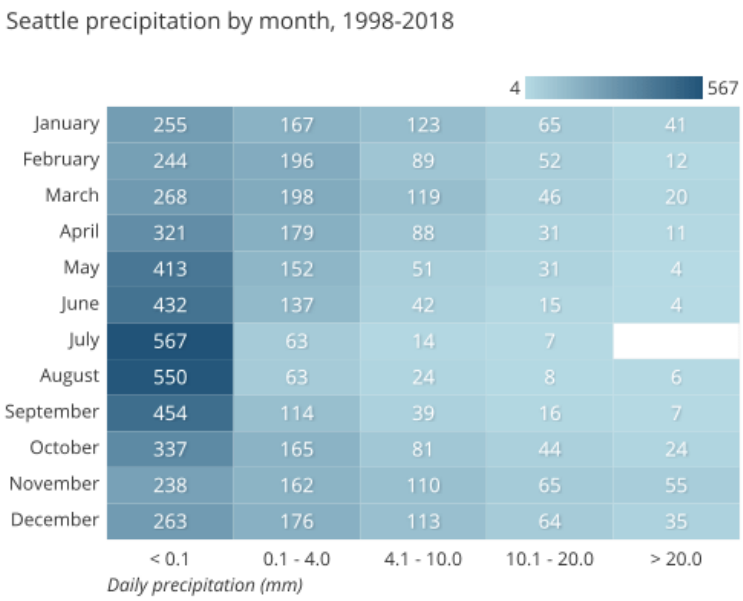
When the two variables in a scatter plot are geographical coordinates – latitude and longitude – we can overlay the points on a map to get a scatter map (aka dot map). This can be convenient when the geographic context is useful for drawing particular insights and can be combined with other third-variable encodings like point size and color.



A famous example of scatter map is John Snow's 1854 cholera outbreak map, showing that cholera cases (black bars) were centered around a particular water pump on Broad Street (central dot). Original: Wikimedia Commons

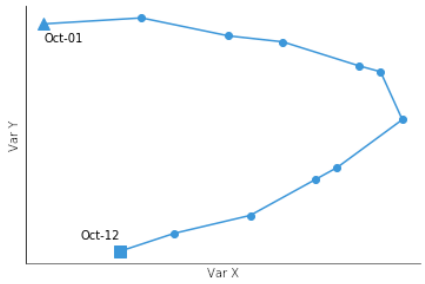
ii. Heat map

Heat map can be a good alternative to the scatter plot when there are a lot of data points that need to be plotted and their density causes over plotting issues. The heatmap can also be used in a similar fashion to show relationships between variables **when one or both variables are not continuous and numeric**. If we try to depict discrete values with a scatter plot, all of the points of a single level will be in a straight line. Heatmaps can overcome this over plotting through their binning of values into boxes of counts.



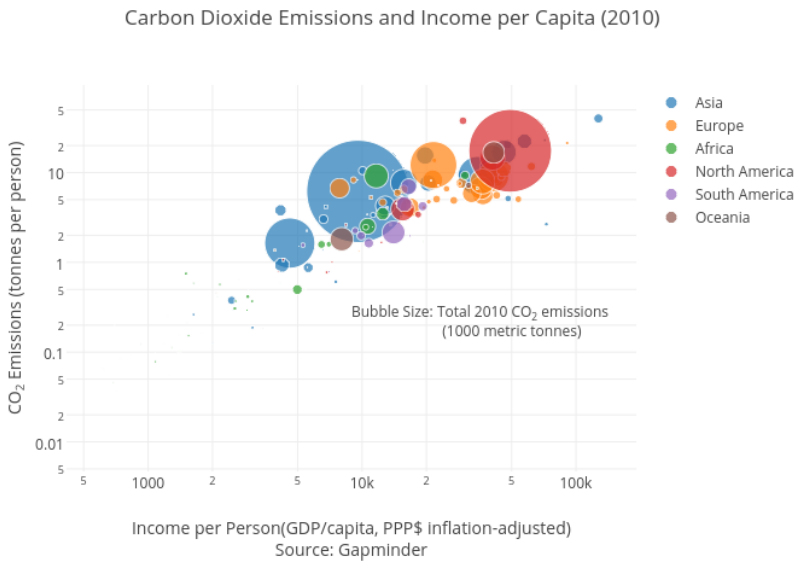
iii. Connected scatter plot

If the third variable we want to add to a scatter plot indicates timestamps, then one chart type we could choose is the connected scatter plot. Rather than modify the form of the points to indicate date, we use line segments to connect observations in order. This can make it easier to see how the two main variables not only relate to one another, but how that relationship changes over time. If the horizontal axis also corresponds with time, then all of the line segments will consistently connect points from left to right, and we have a basic line chart.



B. Bubble Plot (Chart)

A **bubble chart** is a scatter plot that includes a third variable. This third variable is represented as the size of the data point, creating the bubble. Adding another variable can help your data tell a more complete story. This bubble chart uses the same data as the scatter plot, but now the dot size is proportional to the total carbon emissions of the country. The fourth variable shown here is continent, represented with color.



This chart can lend insight that wasn't available with the scatter plot alone. For example, although Qatar has the highest CO2 emissions per person, the entire country doesn't contribute nearly as much to global CO2 as China or the United States. If your bubbles range in size a lot, it might be hard to see the smallest bubbles, and the largest bubbles might obscure the surrounding data. You can make some adjustments on the mode tab under traces.