SOMAIYA VIDYAVIHAR UNIVERSITY
K J Somaiya College of Engineering

Data Visualization

Scatter plot

Vaibhav P. Vasani

Assistant Professor

Department of Computer Engineering

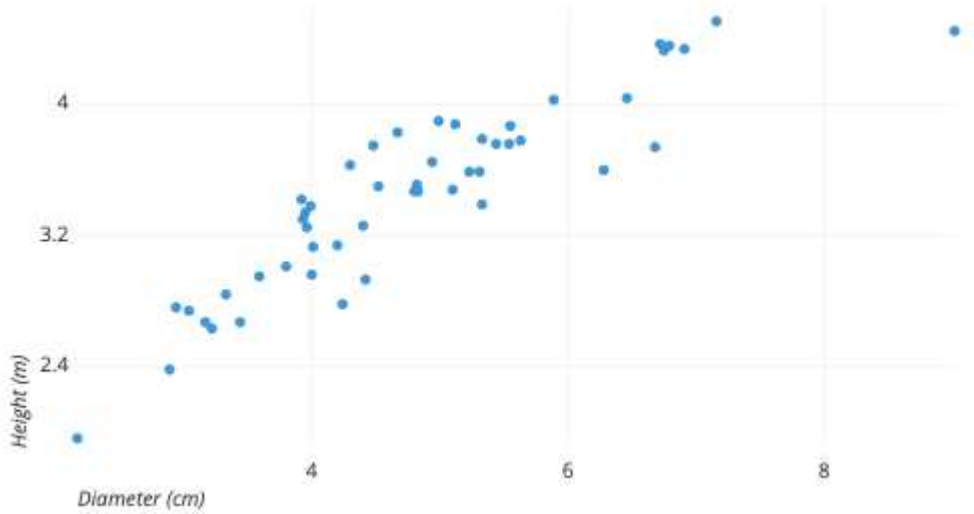
K. J. Somaiya College of Engineering

Somaiya Vidyavihar University

Scatter plot

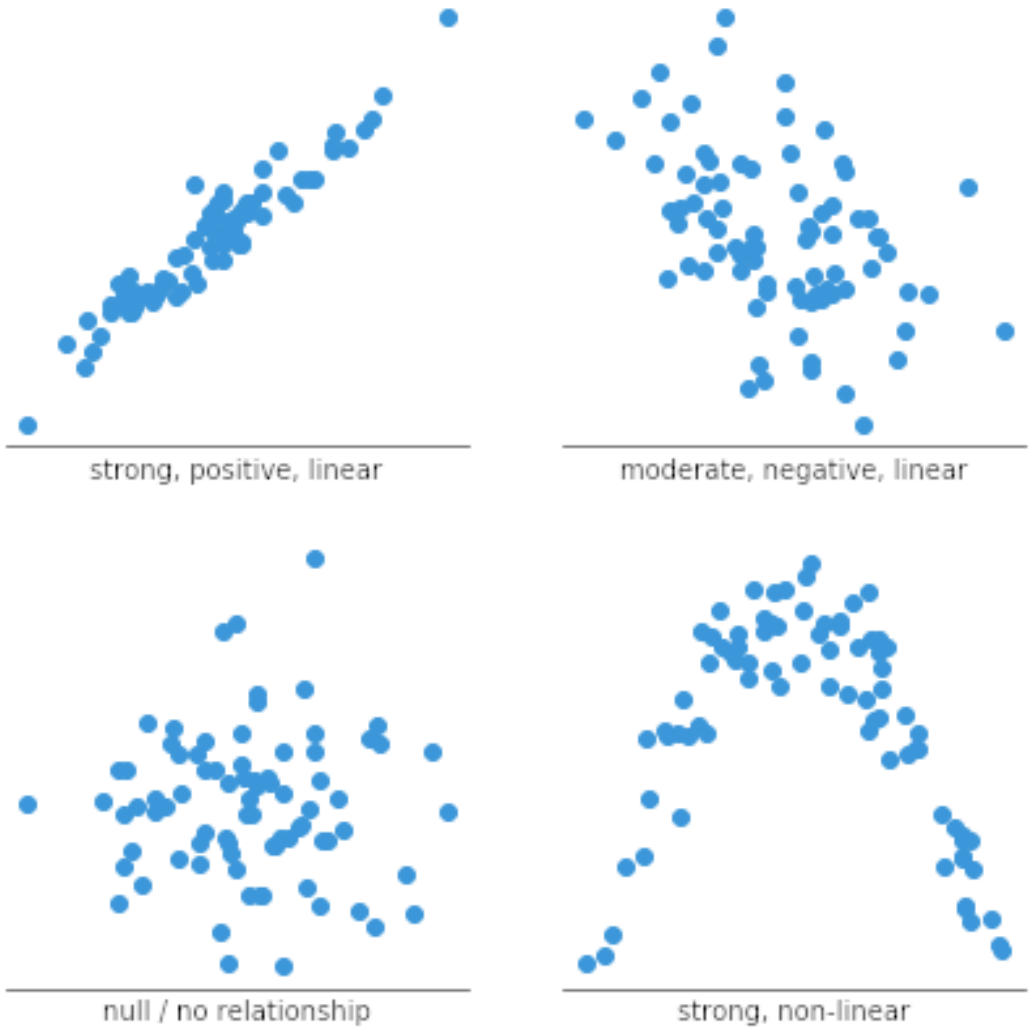
- A scatter plot (also called a scatterplot, scatter graph, scatter chart, scattergram, or scatter diagram)

- A scatter plot 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.

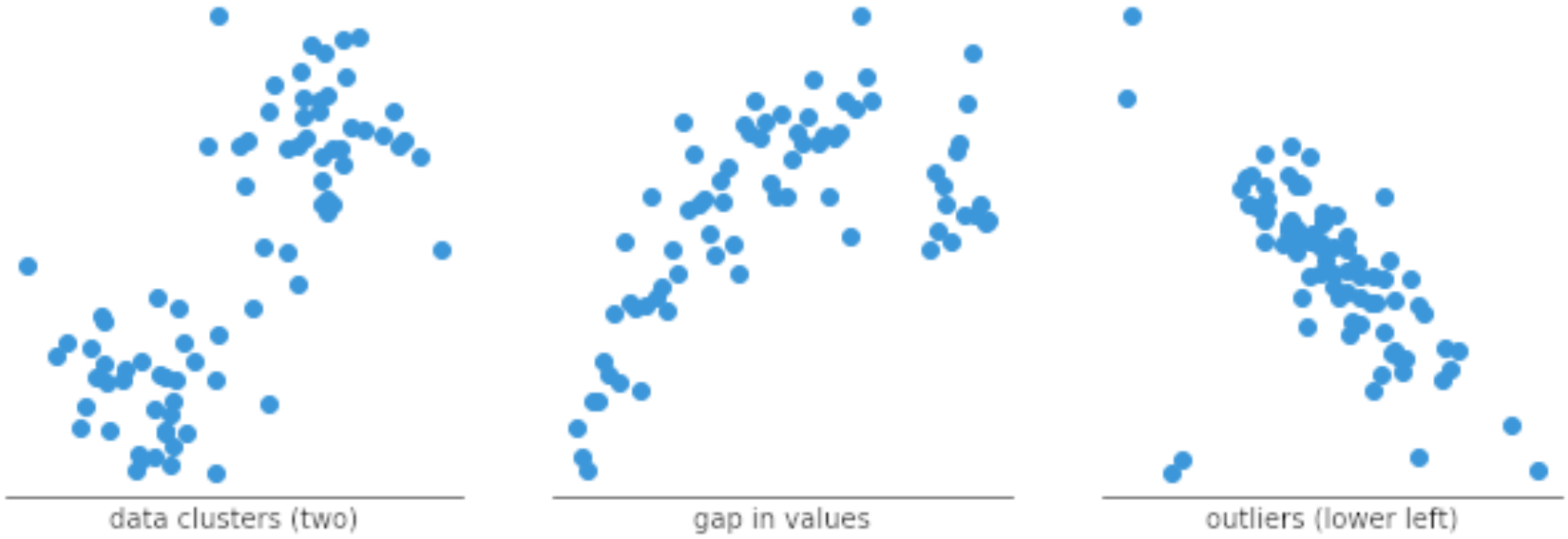

4
3.2
2.4
Height (m)
4
6
8
Diameter (cm)

- Scatter plots' primary uses are to 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.
- Identification of correlational relationships are common with scatter plots.

- 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.

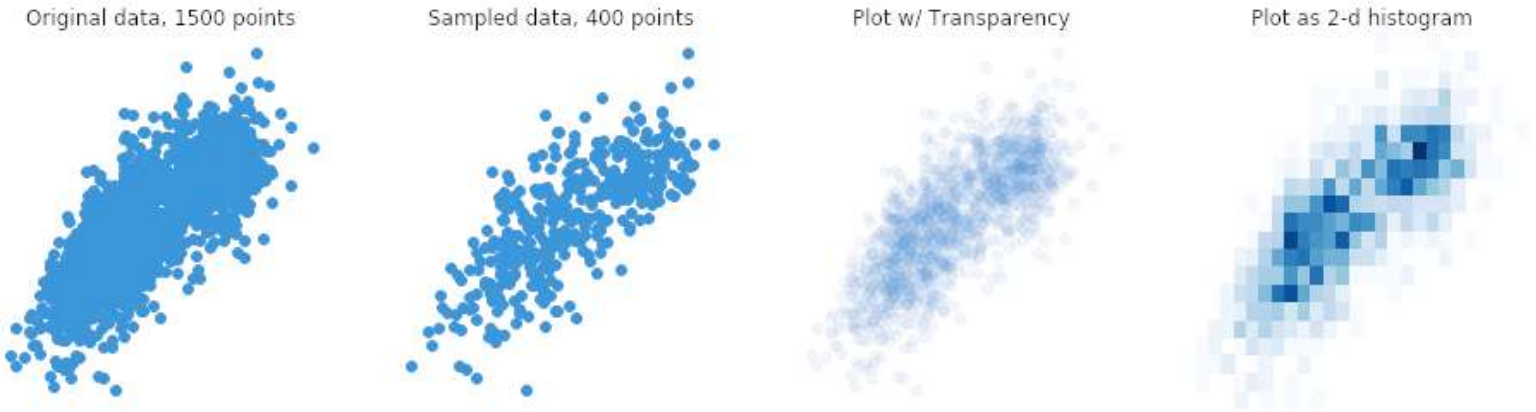

strong, positive, linear
moderate, negative, linear
null / no relationship
strong, non-linear

- A scatter plot can also be useful for identifying 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.


data clusters (two)
gap in values
outliers (lower left)

Common issues when using scatter plots

- Overplotting



Overplotting 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.

Interpreting correlation as causation

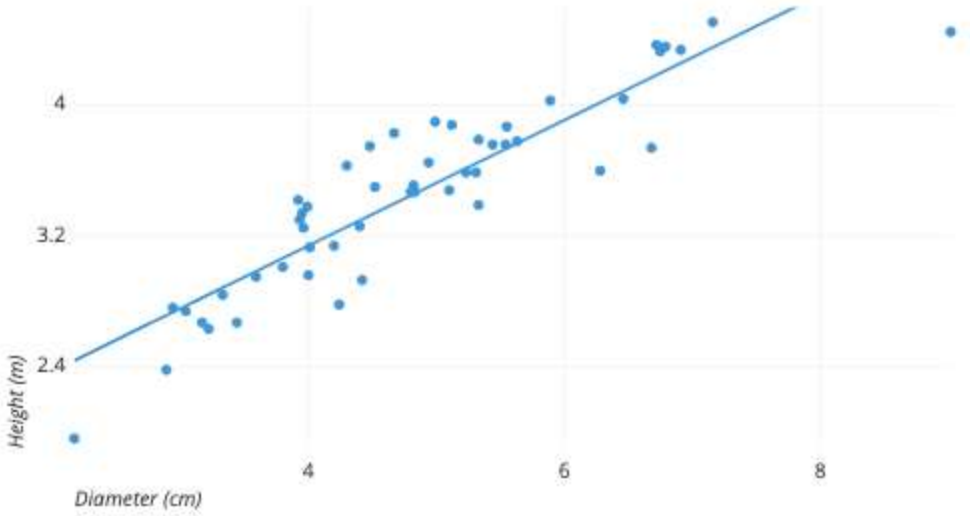
- This is not so much an issue with creating a scatter plot as it is an issue with its interpretation.
- Simply because we observe a relationship between two variables in a scatter plot, it does not mean that changes in one variable are responsible for changes in the other.
- This gives rise to the common 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.

Read Yourself Slide

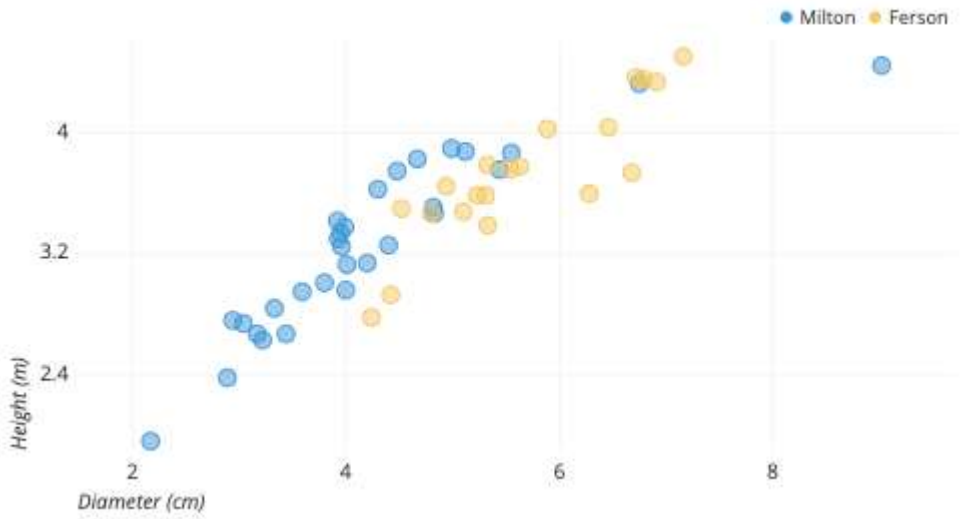
- For example, it would be wrong to look at city statistics for the amount of green space they have and the number of crimes committed and conclude that one causes the other, this can ignore the fact that larger cities with more people will tend to have more of both, and that they are simply correlated 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.
-

Common scatter plot options

- Add a trend line

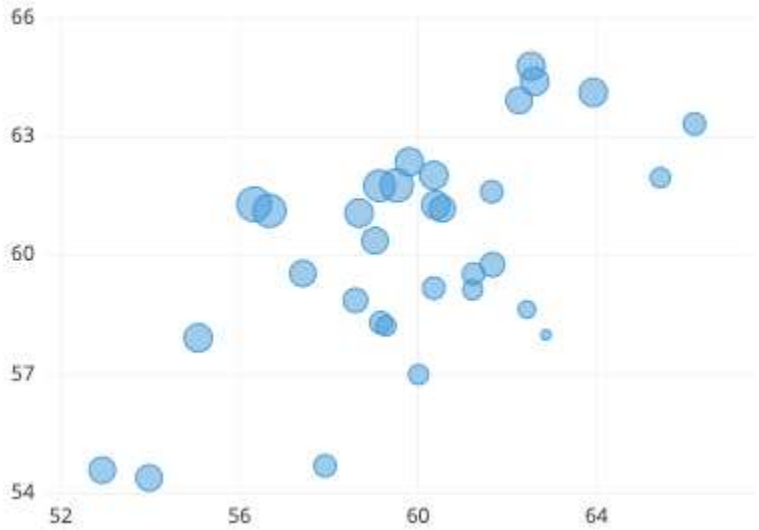


Categorical third variable



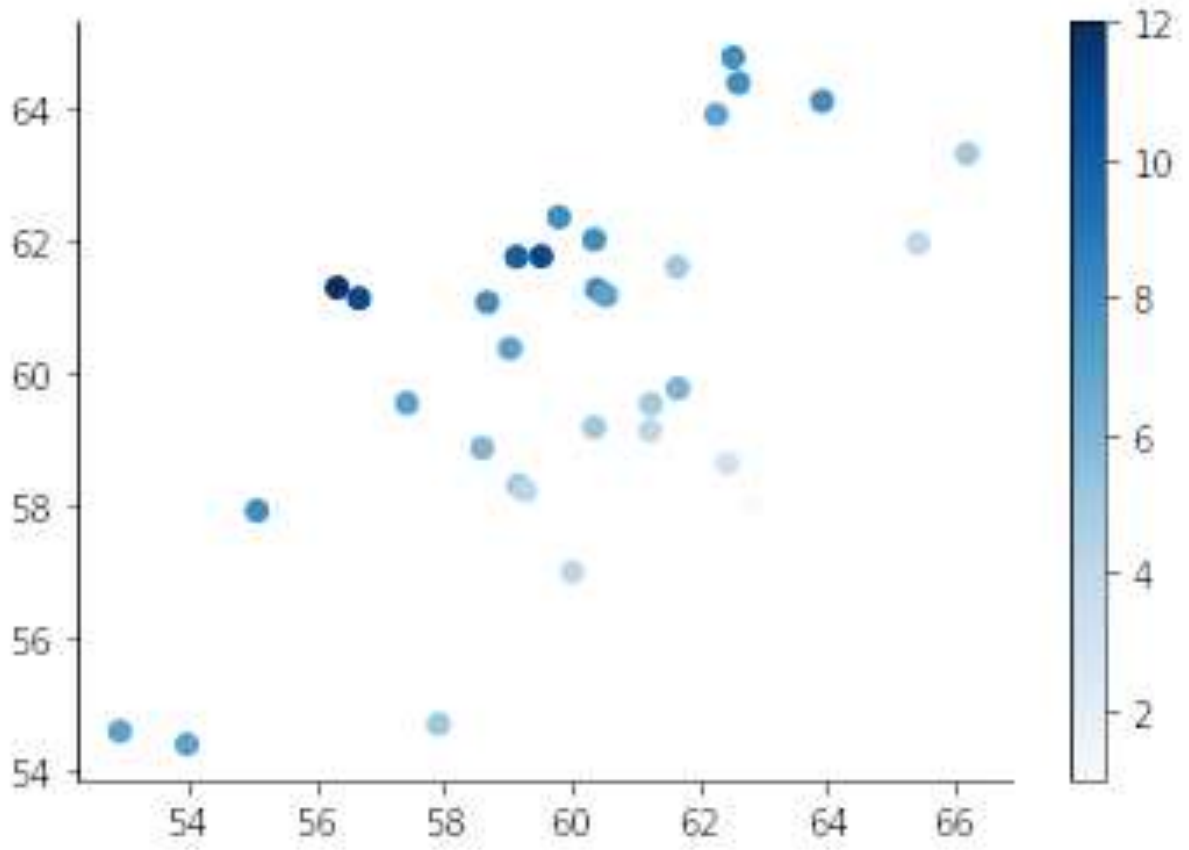
- One other option that is sometimes seen for third-variable encoding is that of shape. One potential issue with shape is that different shapes can have different sizes and surface areas, which can have an effect on how groups are perceived. However, in certain cases where color cannot be used (like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

Numeric third variable

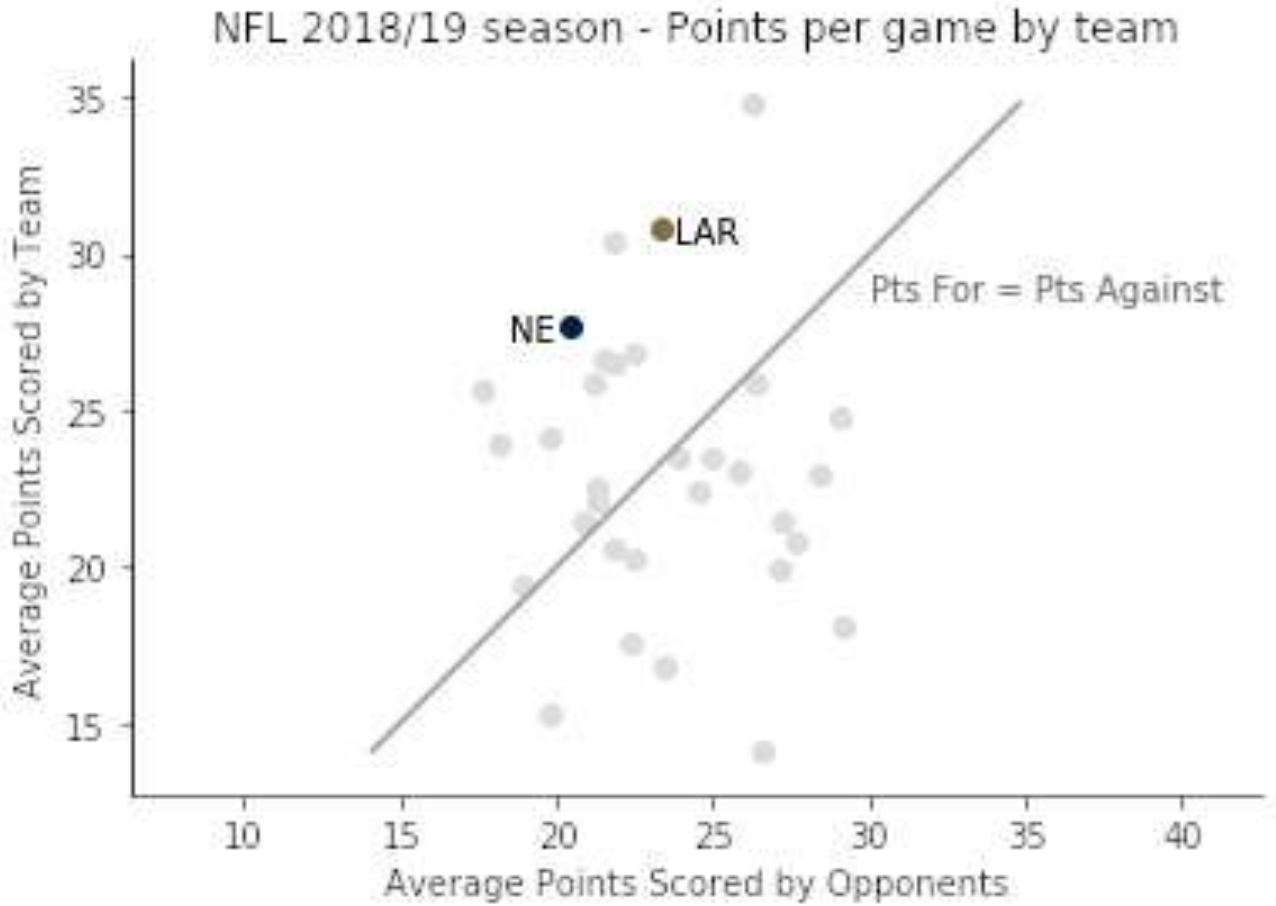


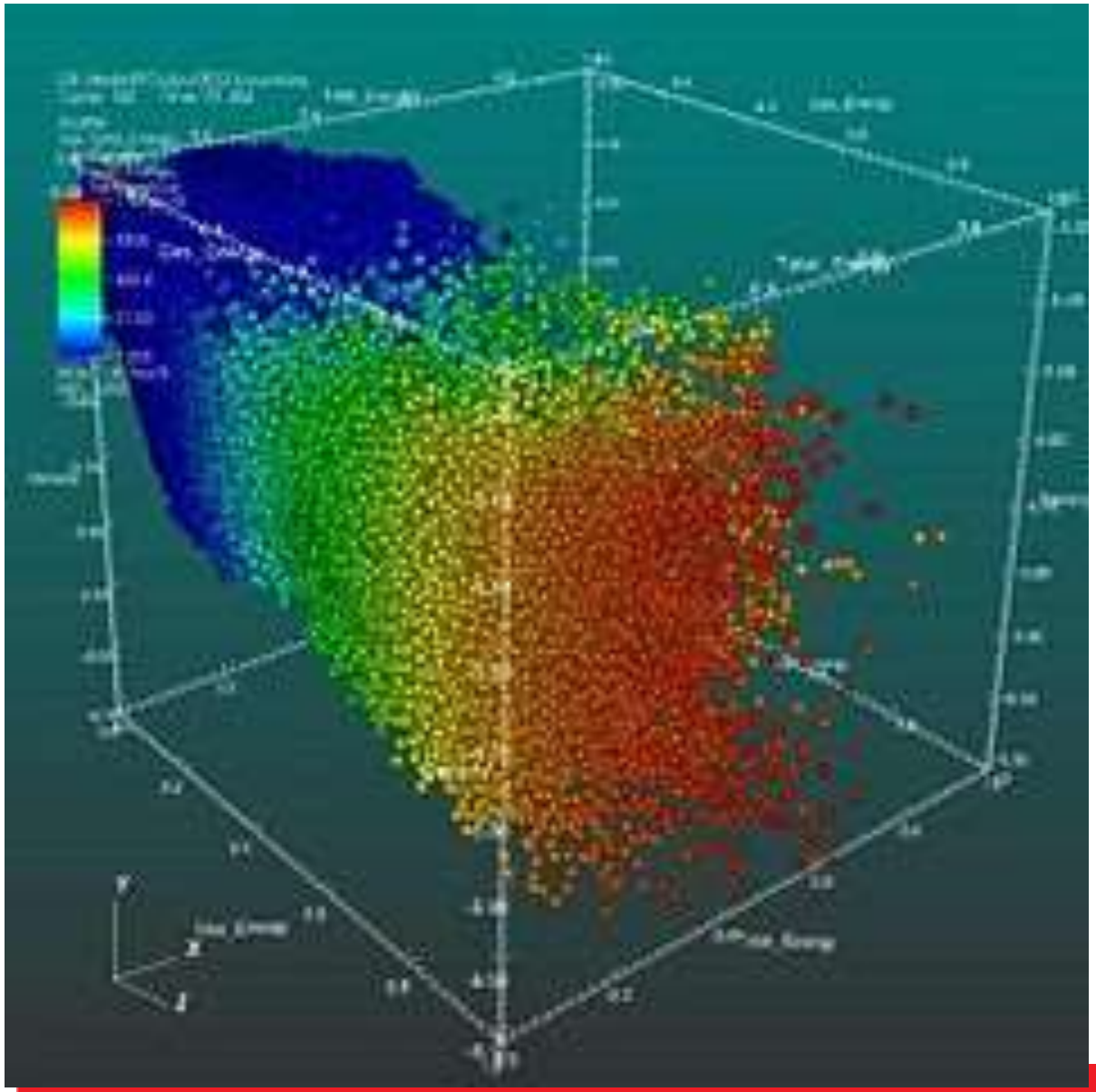
- Hue can also be used to depict numeric values as another alternative. Rather than using distinct colors for points like in the categorical case, we want to use a continuous sequence of colors, so that, for example, darker colors indicate higher value.

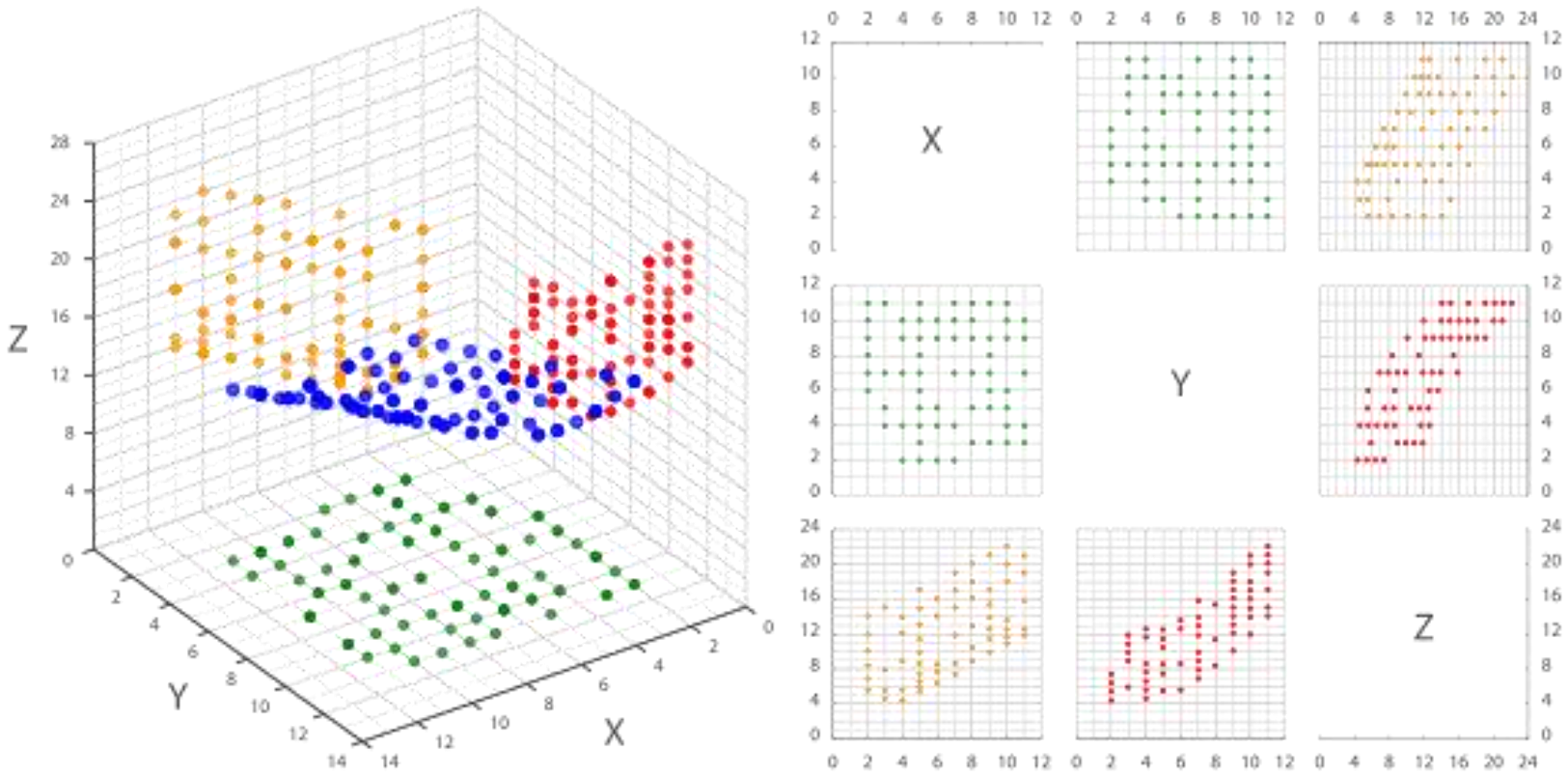
- Note that, for both size and color, a legend is important for interpretation of the third variable, since our eyes are much less able to discern size and color as easily as position.
-


64
62
60
58
56
54
54
56
58
60
62
64
66
12
10
8
6
4
2

Highlight using annotations and color






Z
Y
X
X
Y
Z

Question ?