



Rensselaer

why not change the world?®

Data Analysis II

Ahmed Eleish

**Data Science – ITWS/CSCI/ERTH-4350/6350 Module 5, October 2nd,
2025**

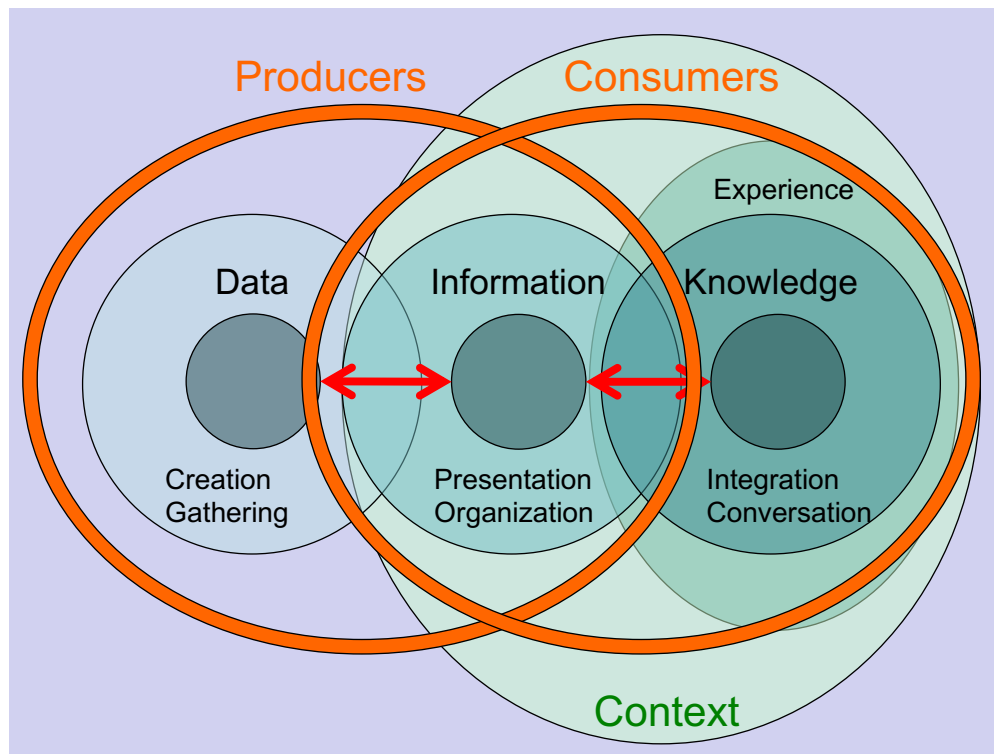


Contents

- Data Analysis I review
- Errors and uncertainty...
- Visualization as an information tool and analysis tool
- Visualization methods
- Use, citation, attribution and reproducibility



Data-Information-Knowledge Ecosystem



Review: Data Analysis I



Types of Data

Type of data	Level of measurement	Examples
Categorical	Nominal (no inherent order in categories)	Eye colour, ethnicity, diagnosis
	Ordinal (categories have inherent order)	Job grade, age groups
	Binary (2 categories – special case of above)	Results of some tests, e.g. positive/negative
Quantitative (Interval/Ratio) (NB units of measurement used)	Discrete (usually whole numbers)	Size of household (ratio)
	Continuous (can, in theory, take any value in a range, although necessarily recorded to a predetermined degree of precision)	Temperature °C/°F (no absolute zero) (interval) Height, age (ratio)

Mean and standard deviation

- The mean, m , of n values of the measurement of a property z (the average).

$$\bar{x} = \frac{1}{n} \left(\sum_{i=1}^n x_i \right) = \frac{x_1 + x_2 + \cdots + x_n}{n}$$

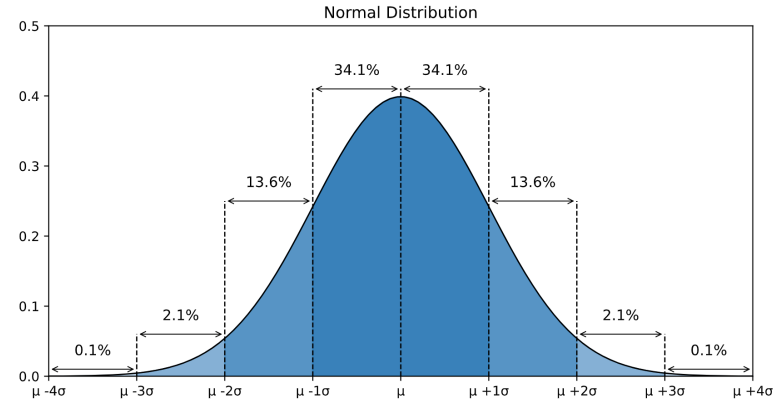
- The standard deviation s of the measurements is an indication of the amount of spread in the measurements with respect to the mean.

$$\sigma = \sqrt{\frac{1}{N} \sum_{i=1}^N (x_i - \mu)^2}, \text{ where } \mu = \frac{1}{N} \sum_{i=1}^N x_i.$$

- The quantity σ^2 is known as the variance of the measurements.

Normal Distribution

- The normal distribution implies tight bounds on the probability of lying far from the mean. 68% of the values must lie within one sigma (standard deviation) of the mean, and 95% within two times the sigma (standard deviation) and 99.7% lie within the three the sigma (standard deviation)



- Roughly 68.3% of the data is within 1 standard deviation of the average (from $\mu - 1\sigma$ to $\mu + 1\sigma$)
- Roughly 95.5% of the data is within 2 standard deviations of the average (from $\mu - 2\sigma$ to $\mu + 2\sigma$)
- Roughly 99.7% of the data is within 3 standard deviations of the average (from $\mu - 3\sigma$ to $\mu + 3\sigma$)

Image Credit: W3C school:

https://www.w3schools.com/statistics/statistics_normal_distribution.php



Visualizing Distributions

- How you visualize the distribution of a variable will depend on whether the variable is categorical or continuous.

A variable is *categorical* if it can only take one of a small set of values.

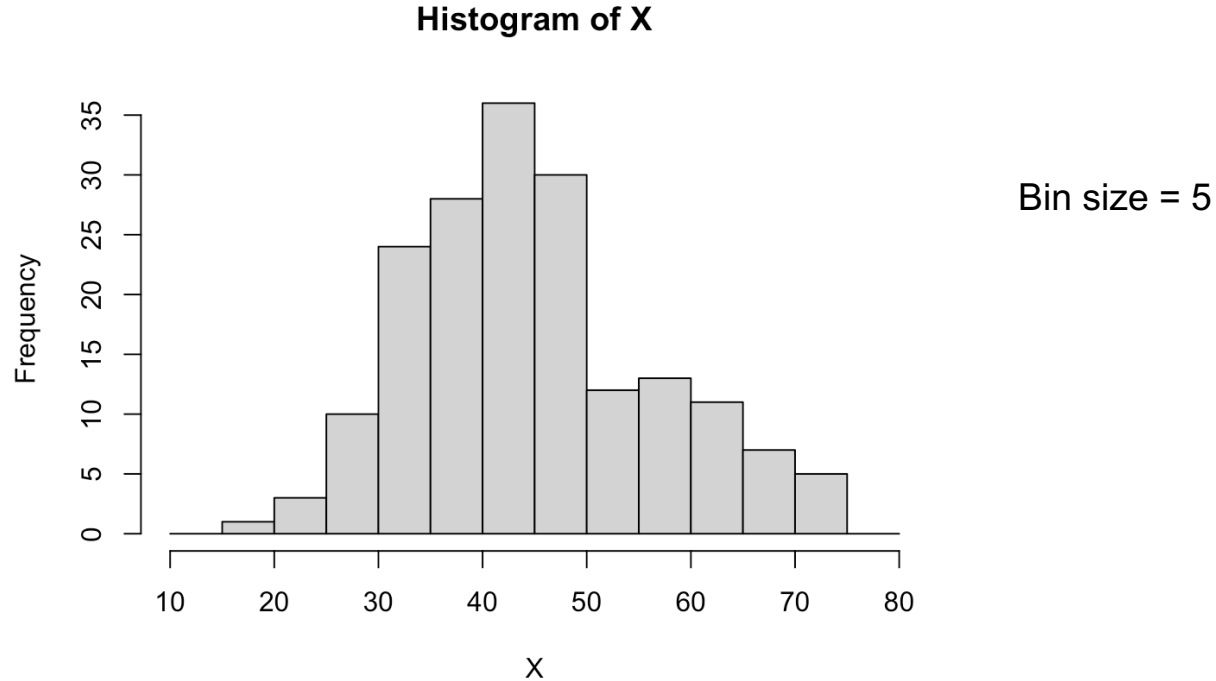
- To examine the distribution of a categorical variable, use a bar chart.

A variable is *continuous* if it can take any numeric value within an interval.

- To examine the distribution of a continuous variable, use a histogram.

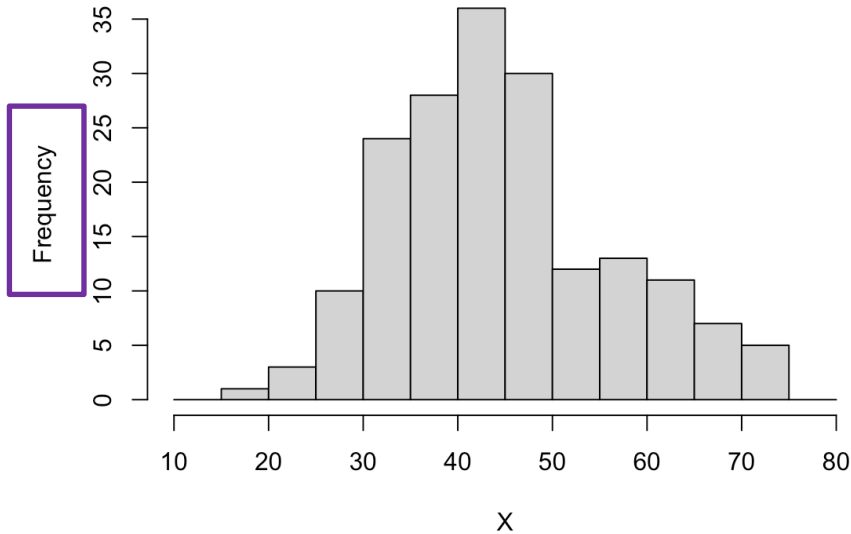
Resource: R for Data Science, *Garrett Grolemund, Hadley Wickham*, Chapter 5, <https://r4ds.had.co.nz/>

Grouped Frequency Distribution aka binning



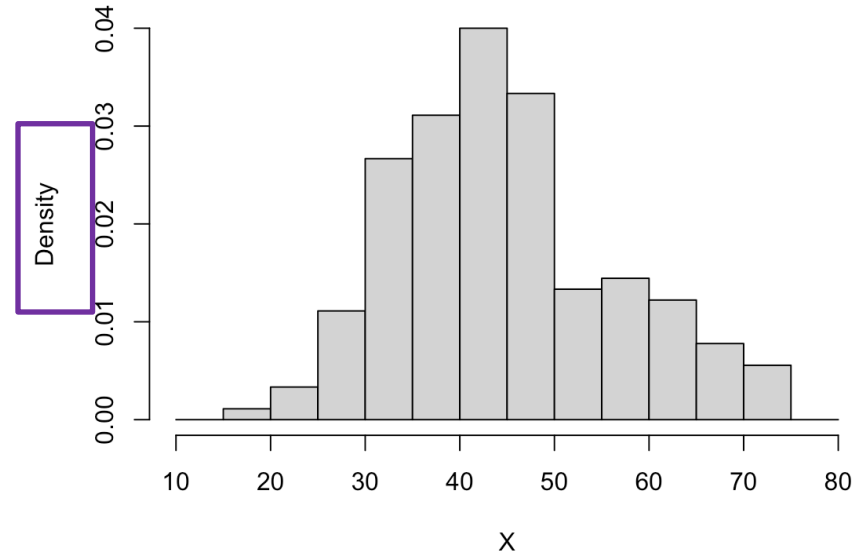
Frequency vs. Density

Histogram of X



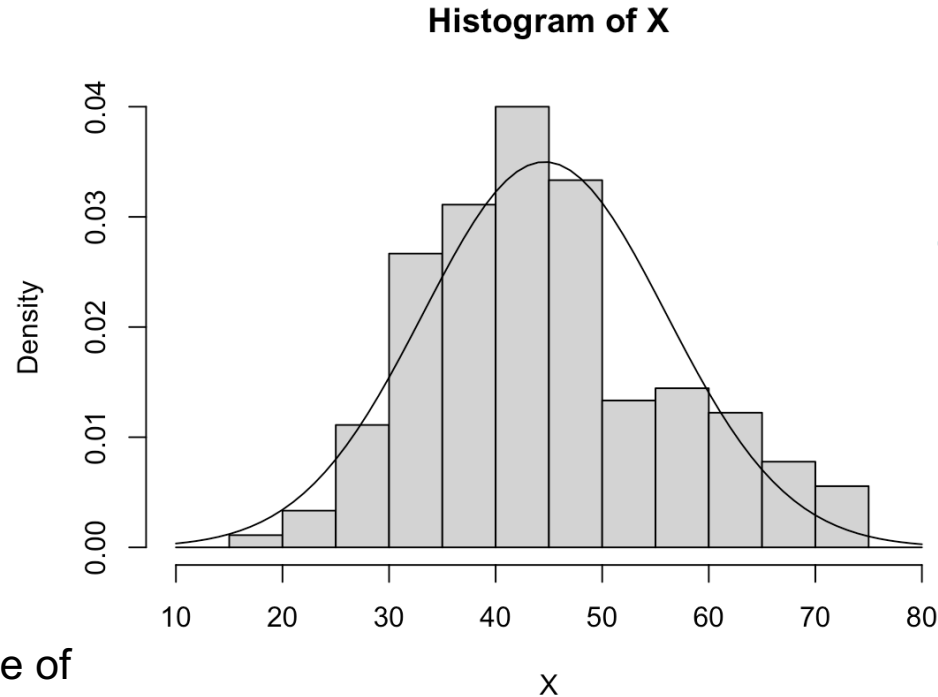
- 36 observations where $40 < x < 45$
- $36/180$ (total) = 0.2 or 20%

Histogram of X



- density at bar = 0.04 where $40 < x < 45$
- area of bar = $0.04 * 5$ (width of bar) = 0.2 or 20%

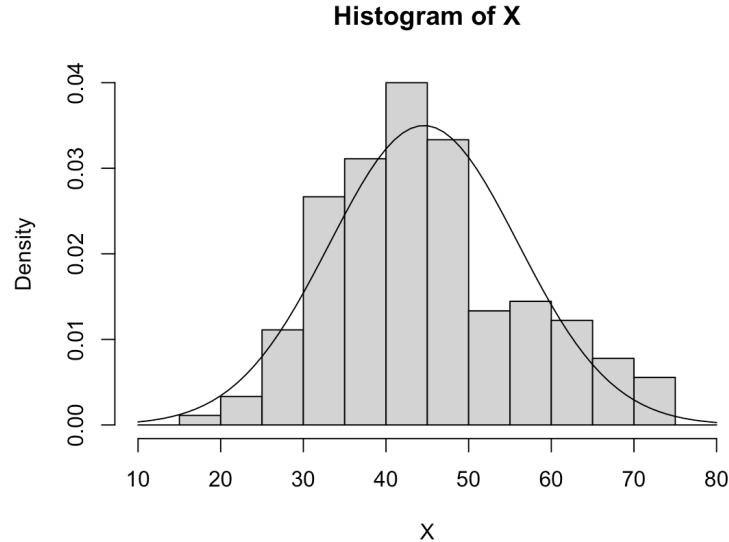
Empirical vs. Theoretical



Is variable X
normally
distributed??

- Probability density curve of normal distribution overlayed

Empirical vs. Theoretical



Is variable X
normally
distributed??

No!!

Shapiro-Wilk normality test

data: X

W = 0.96964, p-value = 0.0005824

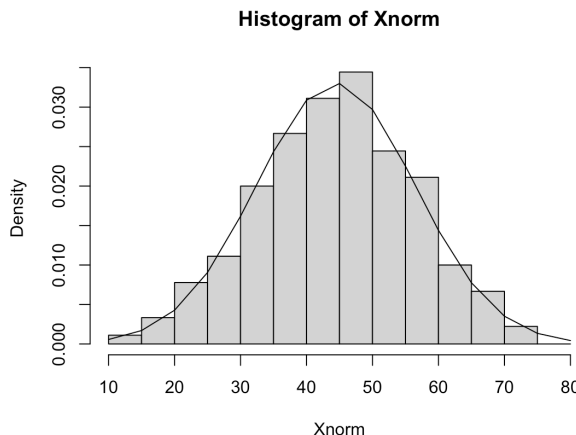
Less than 0.05!

Empirical vs. Theoretical

Let's generate 180 numbers drawn at random from a normal distribution with the same mean and sd as X

**Is variable X
normally
distributed??**

Yes!!



Shapiro-Wilk normality test

data: Xnorm

W = 0.99517, p-value = 0.8308

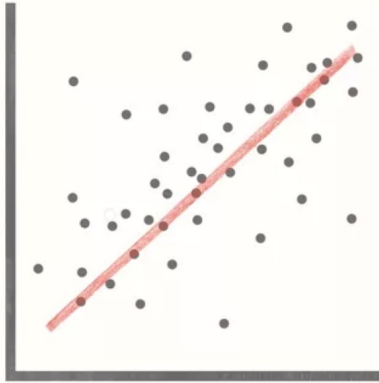
greater than 0.05!

Correlation

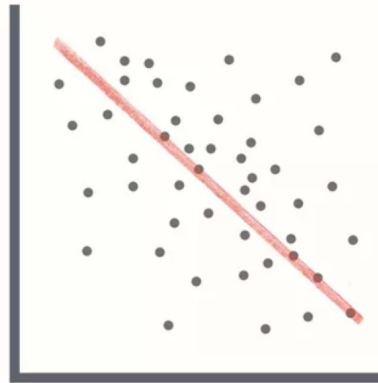
- One measure of the strength of the association between two numerical variables is correlation.
- Correlation describes the strength of the linear association between two variables.
- Correlation coefficient is between -1 and +1
-1 indicates a perfect negative linear association and +1 indicates a perfect positive linear association.
The correlation coefficient of 0, indicates that there is no linear relationship in the two variables. -0.1 and +0.1, indicates no linear relationship *or a very weak* linear relationship
- Correlation coefficient is sensitive to outliers.
- Correlation coefficient is unitless.

Reference(s): <https://www.investopedia.com/terms/c/correlationcoefficient.asp>
<https://www.investopedia.com/ask/answers/032515/what-does-it-mean-if-correlation-coefficient-positive-negative-or-zero.asp>

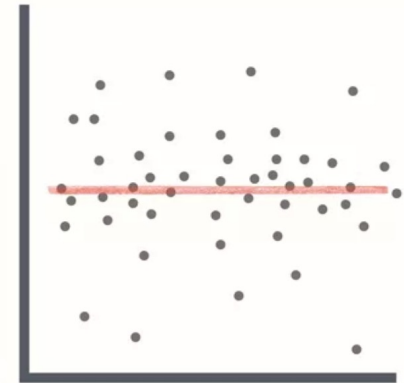
Correlation...



Positive Correlation



Negative Correlation



No Correlation

Image/Photo Credit: <https://www.investopedia.com/ask/answers/032515/what-does-it-mean-if-correlation-coefficient-positive-negative-or-zero.asp>

Input/Output

- Input: input go by different names,
input: *covariates, features, predictors, independent/explanatory variables*, sometimes just variables

$$X = (x_1, x_2, \dots, x_n)$$

- **Output:** The output variable called *response* or *dependent/predicted variable*, typically denoted by Y

- Suppose that we observe quantitative response Y with p different predictor variables,

X_1, X_2, \dots, X_p .

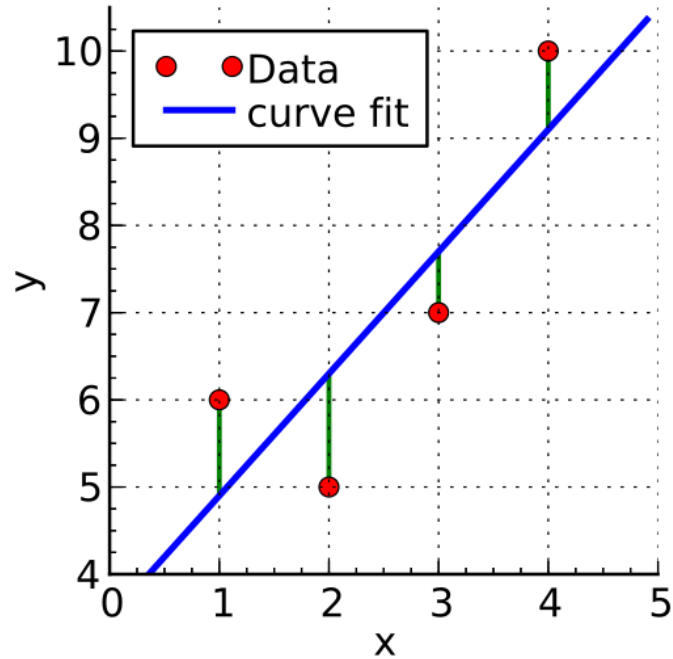
- We assume some relationship between Y and $X = (x_1, x_2, \dots, x_p)$, which can be written as:

$$Y = f(x) + \varepsilon$$

f is an unknown
function of x

random error term, which is
independent of x

Regression

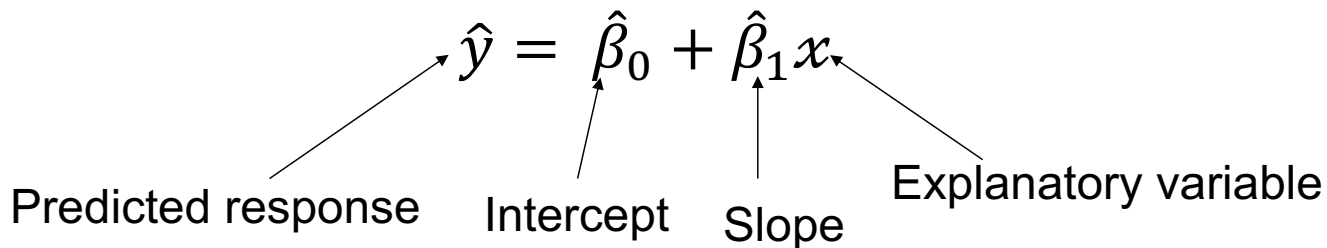


Simple Linear Regression

- The most commonly used approach is the *Least Squares*
- Least Squares approach chooses $\hat{\beta}_0$ and $\hat{\beta}_1$ to minimize the RSS

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$$

Predicted response Intercept Slope Explanatory variable



- \hat{y} = Predicted value of the response variable
- x = Explanatory variable (x)
- $\hat{\beta}_0$ = Intercept
- $\hat{\beta}_1$ = Slope

Residuals ...

- The residual is defined as the difference between the observed value and the predicted value. (Difference between the observed value and the predicted value of the response variable for a given data point).

$$e_i = y_i - \hat{y}_i \quad \text{represents the } i\text{th residual,}$$

this is the difference between the i th observed response value and the i th response value that is predicted by the linear model.

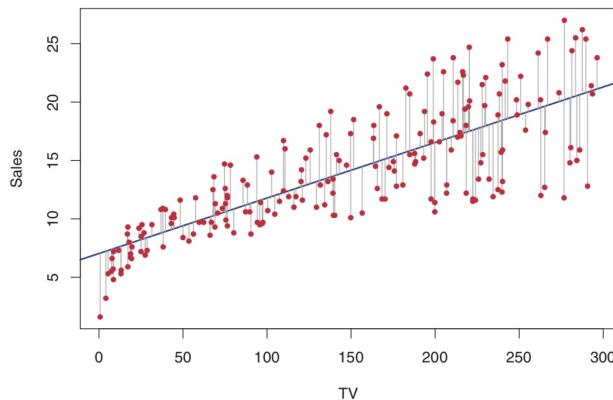
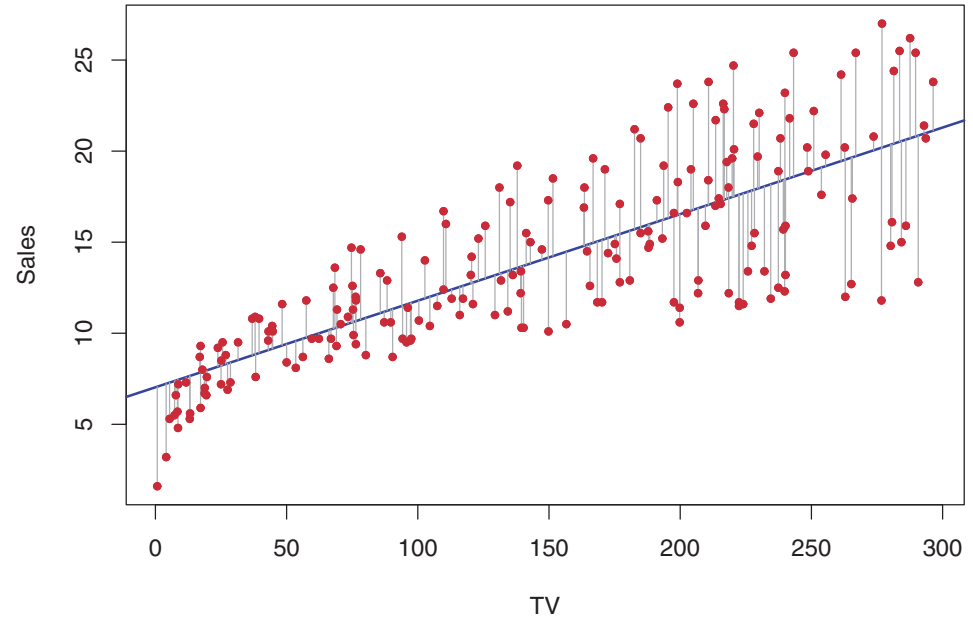


Image Credit: Introduction to Statistical Learning with Applications in R, 7th Edition, Chapter 3 – Linear Regression

Reference: Introduction to Statistical Learning with Applications in R, 7th Edition, Chapter 3 - Linear Regression

Linear Model

- Sales vs. TV ad spending
- Sales in 1000s of units
- TV ad spending in 1000s of \$



Evaluating Linear Models

Values of coefficients >> their Std. errors

High t-statistic

Very low p-value

	Coefficient	Std. error	t-statistic	p-value
Intercept	7.0325	0.4578	15.36	< 0.0001
TV	0.0475	0.0027	17.67	< 0.0001

Hypothesis (more TV ads → more sales)

H0 : There is no relationship between X and Y

Ha : There is some relationship between X and Y

$$t_{\hat{\beta}} = \frac{\hat{\beta} - \beta_0}{SE(\hat{\beta})}$$

Reject the null hypothesis!

Residual Standard Error

- Mean sales $\approx 14,000$ units

RSE = 3.26 = 3,260 units
good/bad?

- R^2
- measures the proportion of the variability in Y that can be explained using X
 - has a value between 0,1

Quantity	Value
Residual standard error	3.26
R^2	0.612
F-statistic	312.1

$$\text{RSE} = \sqrt{\frac{1}{n-2} \text{RSS}} = \sqrt{\frac{1}{n-2} \sum_{i=1}^n (y_i - \hat{y}_i)^2}$$

$$R^2 = \frac{\text{TSS} - \text{RSS}}{\text{TSS}} = 1 - \frac{\text{RSS}}{\text{TSS}}$$

$$\text{TSS} = \sum (y_i - \bar{y})^2$$

Data Analysis II

Errors & Uncertainty

Errors

- Personal errors are mistakes on the part of the experimenter. It is your responsibility to make sure that there are no errors in recording data or performing calculations
- Systematic errors tend to decrease or increase all measurements of a quantity, (for instance all the measurements are too large). E.g. calibration
- Random errors are also known as statistical uncertainties, and are a series of small, unknown, and uncontrollable events

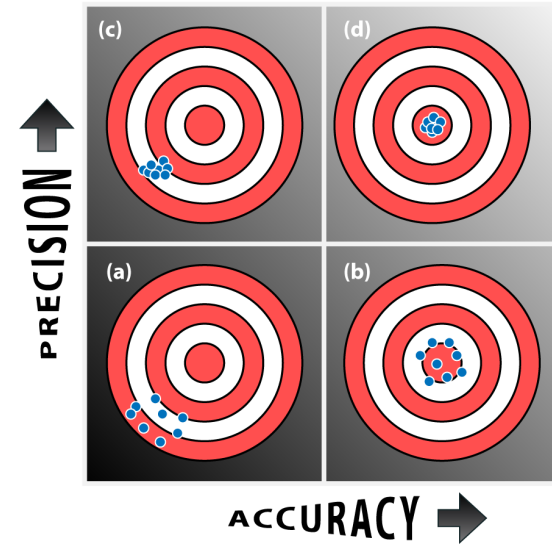
Errors

- Statistical uncertainties are much easier to assign, because there are rules for estimating the value

e.g. If you are reading a ruler, the statistical uncertainty is half of the smallest division on the ruler. Even if you are recording a digital readout, the uncertainty is half of the smallest place given. This type of error should always be recorded for any measurement

Standard measures of error

- Absolute deviation
 - is simply the difference between an experimentally determined value and the true value
- Relative deviation
 - is a more meaningful value than the absolute deviation because it accounts for the relative size of the error. The relative percentage deviation is given by the absolute deviation divided by the true value and multiplied by 100%
- Standard deviation



Some considerations

- Possibly more important than our answer is our confidence in the answer.
- Our confidence is quantified by uncertainties.
- Once we combine numbers, we need to be able to assess how the uncertainties change for the combination.
- This is called **propagation of errors** or more correctly the propagation of our understanding/ estimate of errors in the result we are looking at...

Resolution

Accuracy and Generalization



Actual soil interdigitation

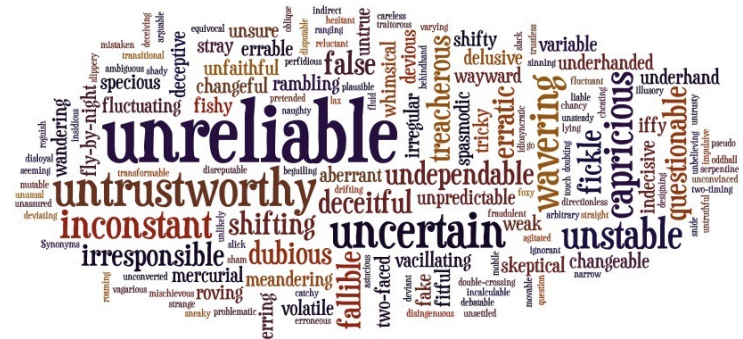


Generalization on map

Different soil type boundaries are generalized when mapping an area, but are actually vague and graduated. Differences in scale allow finer resolution, but only if the original data was collected at a finer resolution.

Reliability

- Changes in data over time
- Non-uniform coverage
- Map scales
- Observation density
- Sampling theorem (aliasing)
- Surrogate data and their relevance
- Round-off errors in computers



Propagating errors

- This is an unfortunate term – it means making sure that the result of the analysis carries with it a calculation (rather than an estimate) of the error.

e.g. if $x = y + z$ (your analysis), then $\Delta x = \Delta y + \Delta z$

e.g. if $x = y + z$ (your analysis), then $\Delta x = \Delta y + \Delta z$!

- It's not as simple for other calculations.
- When the function is not merely addition, subtraction, multiplication, or division, the error propagation must be defined by the total derivative of the function.

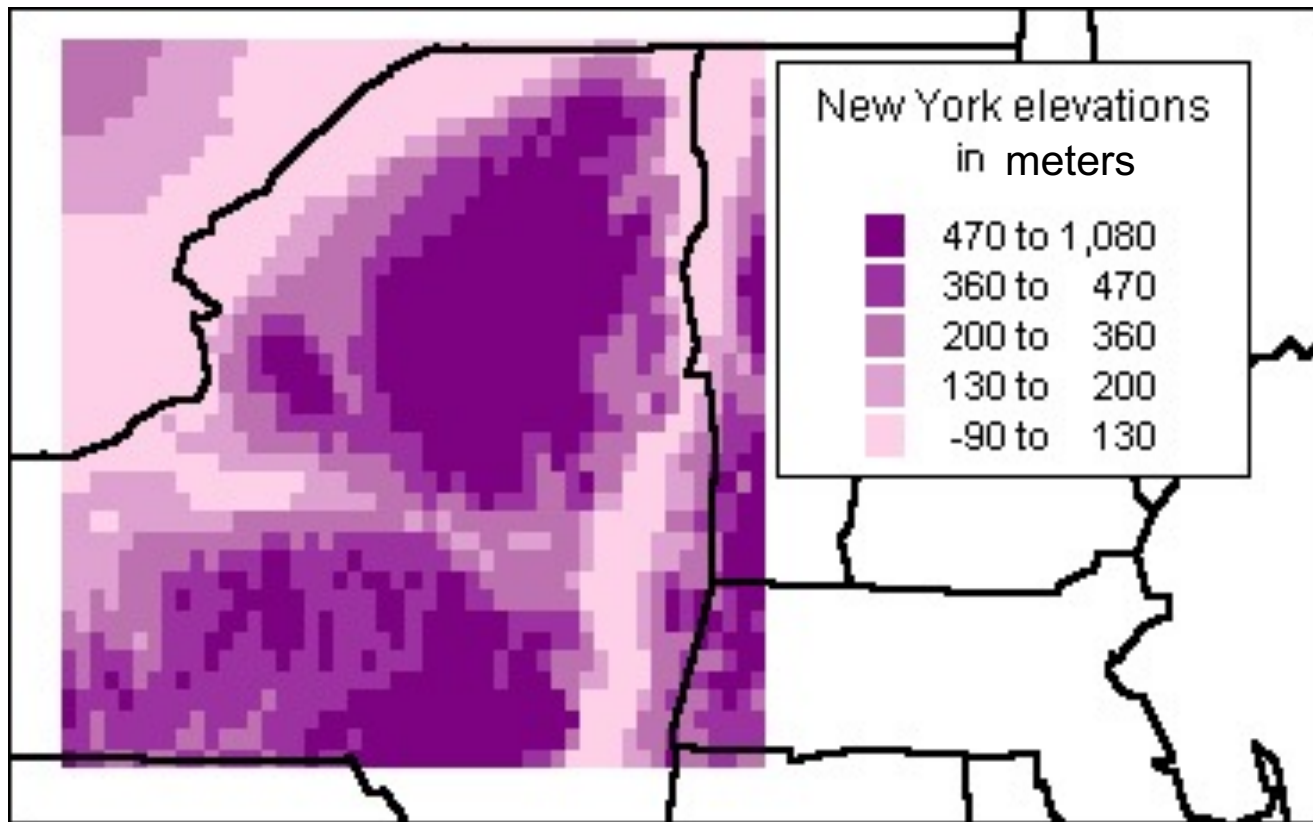
Error propagation

- Errors arise from data quality, model quality and data/model interaction.
- We need to know the sources of the errors and how they propagate through our model.
- The simplest representation of errors is to treat observations/attributes as statistical data – use mean and standard deviation.

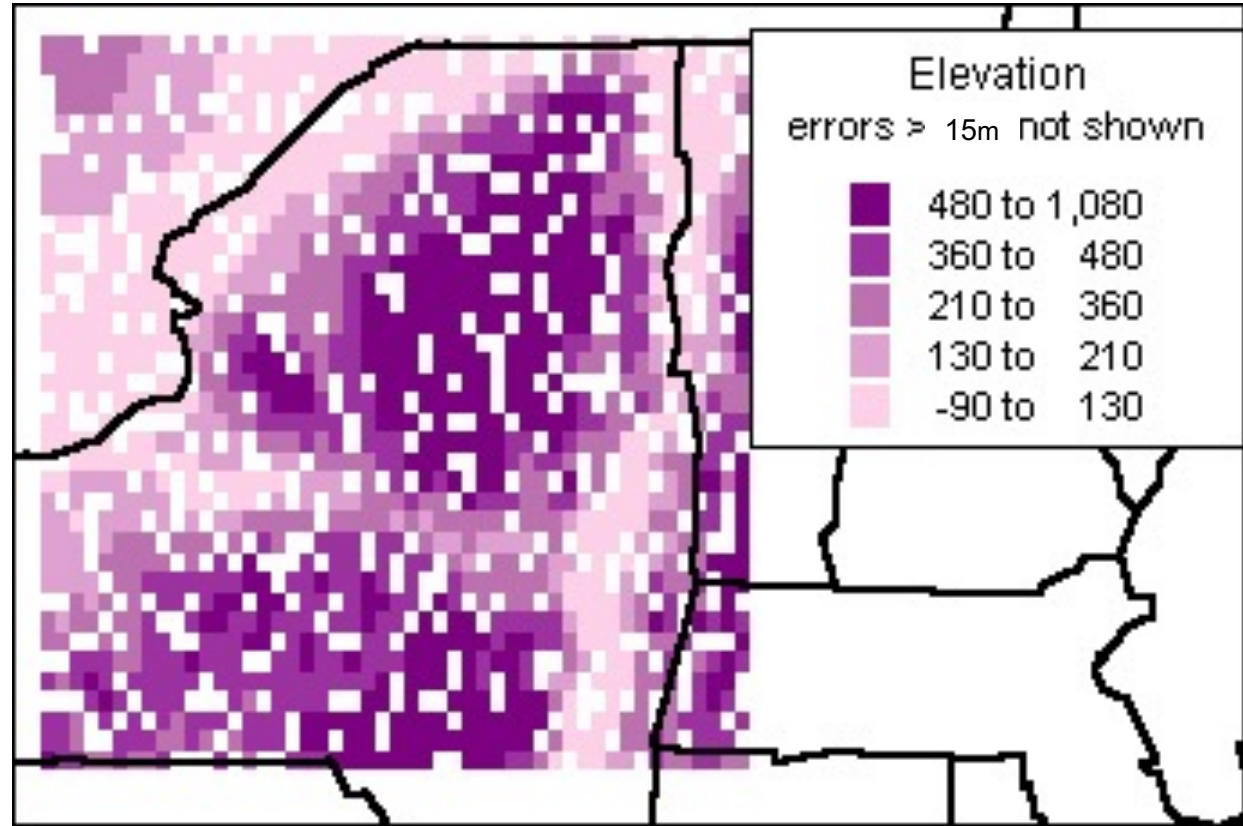
Dealing with errors

- In analyses:
 - Report on the statistical properties.
 - Does it pass tests at some confidence level?
- In visualizations:
 - Exclude data that are not reliable (plot only a subset of data).
 - Annotate the figure with some measure of confidence.

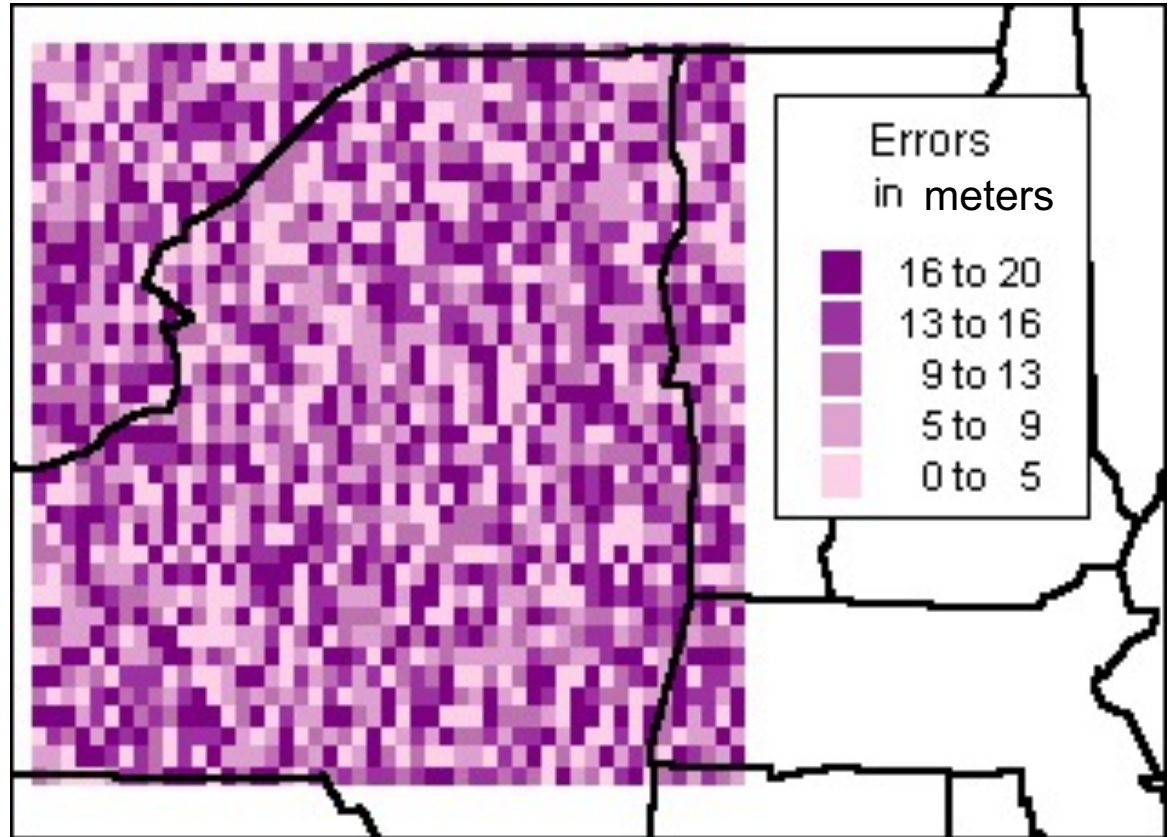
Elevation map



Larger errors 'whited out'



Elevation errors



Reporting results/ uncertainty

- Consider the number of significant digits in the result which is indicative of the certainty of the result.
- The number of significant digits depends on the measuring equipment you use and the precision of the measuring process - do not report digits beyond what was recorded.
- The number of significant digits in a value defines the precision of that value

Reporting results...

- In calculations, it is important to keep enough digits to avoid round off error.
- In general, keep at least one more digit than is significant in calculations to avoid round off error.
- It is not necessary to round every intermediate result in a series of calculations, but it is very important to round your final result to the correct number of significant digits.

Uncertainty

- Results are usually reported as result \pm uncertainty (or error).
- The uncertainty is given to one significant digit, and the result is rounded to that place.
- For example, a result might be reported as $12.7 \pm 0.1 \text{ m/s}^2$. A more precise result would be reported as $12.745 \pm 0.004 \text{ m/s}^2$. A result should not be reported as $12.70361 \pm 0.2 \text{ m/s}^2$.
- Units are very important to any result!

Secondary analysis

- Depending on where you are in the data analysis.
- Having a clear enough awareness of what has been done to the data (either by you or others) prior to the next analysis step is very important – it is very similar to sampling bias.
- Read the metadata (or create it) and documentation.

Visualizations

Considerations for visualizations as analysis

- What is the improvement in the understanding of the data as compared to the situation without visualization?
- Which visualization techniques are suitable for one's data?
 - e.g. Are direct volume rendering techniques to be preferred over surface rendering techniques?

Why visualization?

- Reducing amount of data, quantization
- Patterns
- Features
- Events
- Trends
- Irregularities
- Leading to presentation of data, i.e. information products
- *Exit points for analysis*

Types of visualization

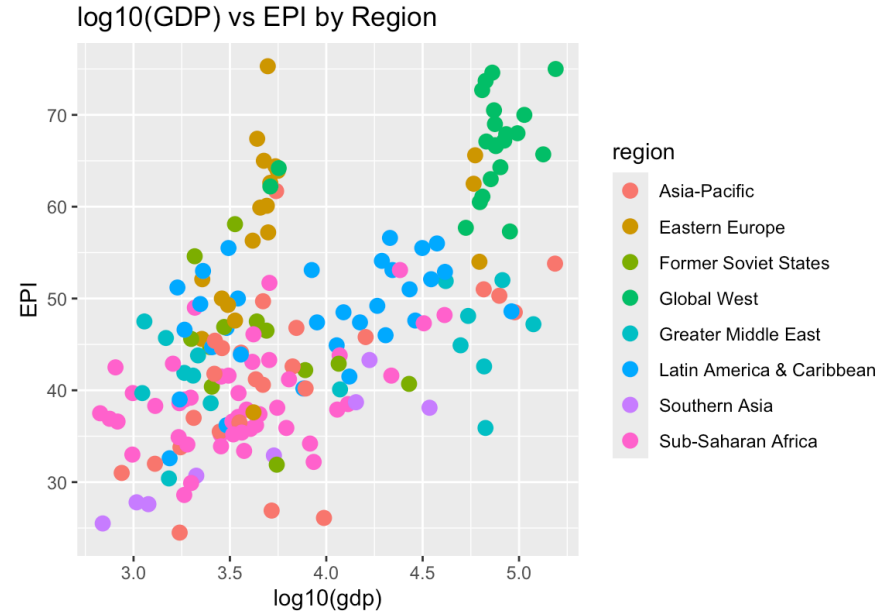
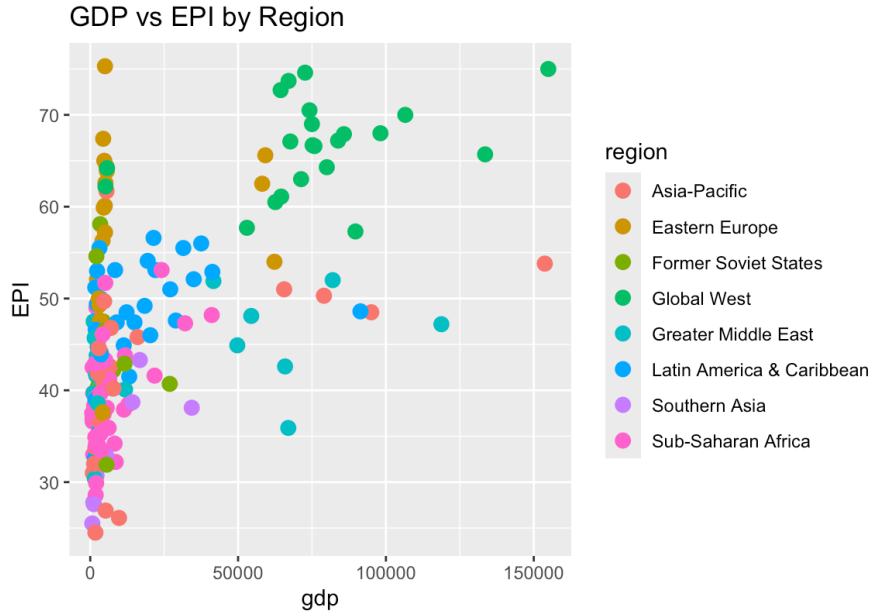
- Color coding (including false color)
- Classification of techniques is based on
 - Dimensionality
 - Information being sought, i.e. purpose
- Line/scatter/bar plots
- Networks
- Contours
- Volume rendering techniques
- Animation techniques
- Non-realistic, including 'cartoon/ artist' style

Visualizations

- Scatter Plot – Paired data (x,y)
- Describe the relationship between numerical variables.
- Make a note on the direction of the data points
 - Positive direction
 - Negative Direction
- Check for unusual observations
- See the relationship - Linear or Non-linear

Visualizations

- Scatter Plot – Paired data (x,y)

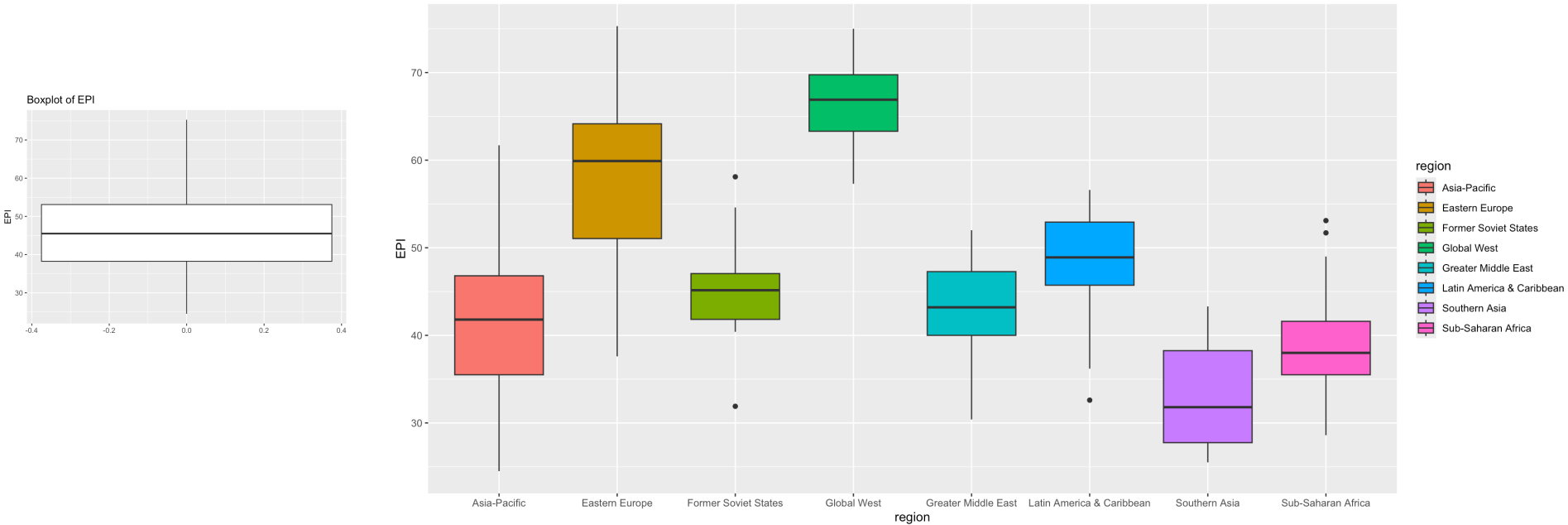


Visualizations

- Boxplot - box-and-whisker of a variable x
- Gives an overview of the range of the variable and where most of the observations are exist along that range
- Make a note of the minimum, maximum, quartiles
- Check for unusual observations (outliers)

Visualizations

- Boxplot - box-and-whisker of a variable x



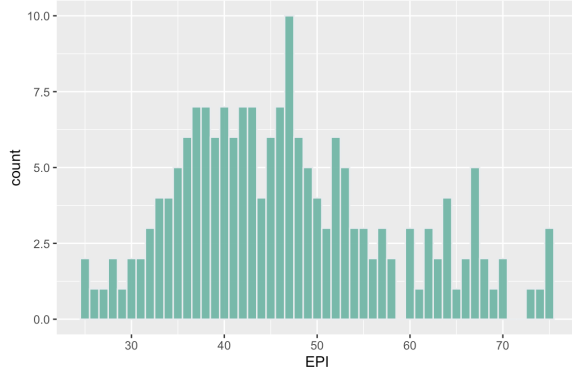
Visualizations

- Histogram – Distribution of variable X
 - Describe frequency of occurrence of values or ranges of values (bins) of x
 - Tune bin size parameter
 - Observe overall shape
 - Check for mixed distributions

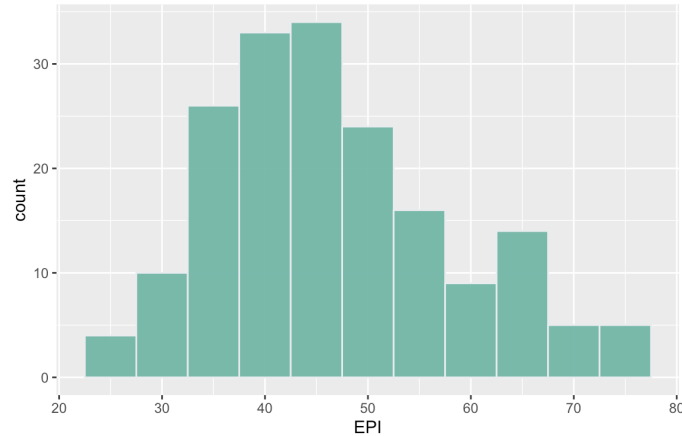
Visualizations

- Histogram – Distribution of variable X

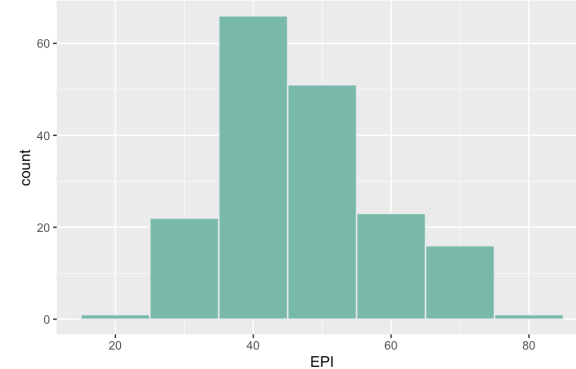
Histogram of EPI - Bin size = 1



Histogram of EPI - Bin size = 5



Histogram of EPI - Bin size = 10



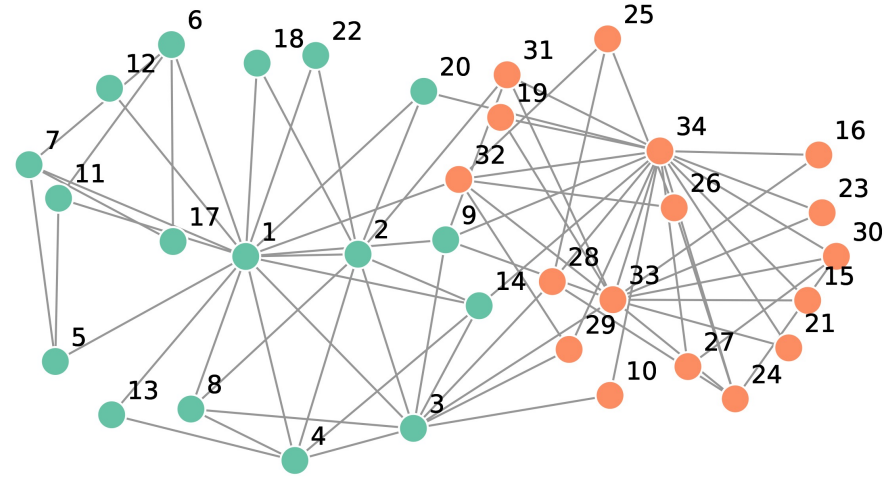
Visualizations

- Network Diagram (Graph) – Relationships between entities
- Represents entities as vertices and the connections between them as edges.
- High-dimensional representation.
- Layout must be computed (e.g. force-directed layout)
- Annotations add information (node sizes/colors, etc.)

Visualizations

- Network Diagram (Graph) – Relationships between entities

- Network of karate club members
- Colors represent a factional split between the instructor and administrator














































































credit: [Paresnah](#) - license [CC BY-SA 4.0](#) (no changes)

Visualizations

- More network diagrams (that I worked on)!

<https://dtdi.carnegiescience.edu/>

A PERIODIC TABLE OF VISUALIZATION METHODS

 C continuum	 Data Visualization Visual representations of quantitative data in schematic form (either with or without axes)										 G graphic facilitation						
 Tb table	 Ca cartesian coordinates	 Information Visualization The use of interactive visual representations of data to amplify cognition. This means that the data is transformed into an image, it is mapped to screen space. The image can be changed by users as they proceed working with it										 Me meeting trace	 Mm metro map	 Tm temple	 St story template	 Tr tree	 Ct cartoon
 Pi pie chart	 L line chart	 Concept Visualization Methods to elaborate (mostly) qualitative concepts, ideas, plans, and analyses.										 Co communication diagram	 Fp flight plan	 Cs concept skeleton	 Br bridge	 Fu funnel	 Ri rich picture
 B bar chart	 Ac area chart	 R radar chart	 Pa parallel coordinates	 Hy hyperbolic tree	 Cy cycle diagram	 T timeline	 Ve vean diagram	 Mi mindmap	 Sq square of oppositions	 Cc concentric circles	 Ar argument slide	 Sw swim lane diagram	 Gc gantt chart	 Pm perspectives diagram	 D dilemma diagram	 Pr parameter ruler	 Kn knowledge map
 Hi histogram	 Sc scatterplot	 Sa sankey diagram	 In information lense	 E entity relationship diagram	 Pt petri net	 Fl flow chart	 Cl clustering	 Lc layer chart	 Py pyramid technique	 Ce cause-effect chains	 Tl tollman map	 Dt decision tree	 Cp cpm critical path method	 Cf concept fan	 Co concept map	 Ic iceberg	 Lm learning map
 Tk tukey box plot	 Sp spectrogram	 Da data map	 Tp treemap	 Cn cone tree	 Sy system dyn./simulation	 Df data flow diagram	 Se semantic network	 So soft system modeling	 Sn synergy map	 Fo force field diagram	 Ib ibis argumentation map	 Pr process event chains	 Pe pert chart	 Ev evocative knowledge map	 V vee diagram	 Hh heaven 'a' hell chart	 I informal






Cy **Process Visualization**





























Note: Depending on your location and connection speed it can take some time to load a pop-up picture.

version 1.5

© Ralph Lengler & Martin J. Eppler, www.visual-literacy.org

Hy **Structure Visualization**

-  **Overview**
-  **Detail**
-  **Detail AND Overview**
-  **Divergent thinking**
-  **Convergent thinking**

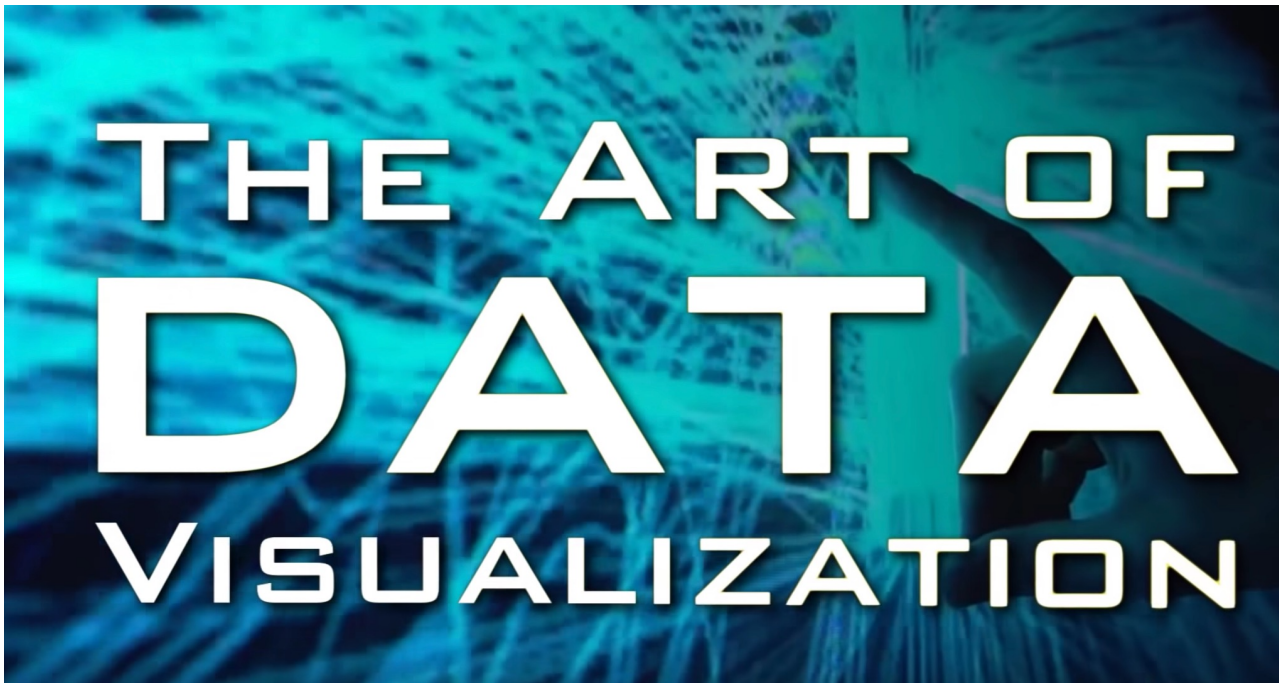
 Su supply demand curve	 Pc performance charting	 St strategy map	 Oc organisation chart	 Ho house of quality	 Fd feedback diagram	 Ft failure tree	 Mq magic quadrant	 Ld life-cycle diagram	 Po porter's five forces	 S s-cycle	 Sm stakeholder map	 Is ishikawa diagram	 Tc technology roadmap
 Ed edgeworth box	 Pf portfolio diagram	 Sg strategic game board	 Mz mintzberg's organigram	 Z zwick's morphological box	 Ad affinity diagram	 De decision discovery diagram	 Bm bcg matrix	 Stc strategy canvas	 Vc value chain	 Hy hype-cycle	 Sr stakeholder rating map	 Ta taps	 Sd spray diagram

https://www.visual-literacy.org/periodic_table/periodic_table.html

Managing visualization products

- The importance of a 'self-describing' product
- Visualization products are not *just* consumed by people
- How many images, graphics files do you have on your computer for which the origin, purpose, use is still known?
- How are these logically organized?

Motivation: Art of Data Visualization



<https://www.youtube.com/watch?v=AdSZJzb-aX8>

Use, citation, attribution

- Think about and implement a way for others (including you) to easily use, cite, attribute any analysis or visualization you develop
- This *must* include suitable connections to the underlying (aka backbone) data – and note this may not just be the full data set!
- Naming, logical organization, etc. are key
- Make them a resource, e.g. URI / URL

See <http://commons.esipfed.org/node/308>

Reproducibility

- The documentation around procedures used in the analysis and visualization are very often neglected – DO NOT make this mistake
- Treat this *just* like a data collection (or generation) exercise
- Follow your management plan
- Despite the lack or minimal metadata/ metainformation standards, capture and record it
- Get someone else to verify that it works

Thanks!