COMP20008
Elements of data processing

Semester 1 2020

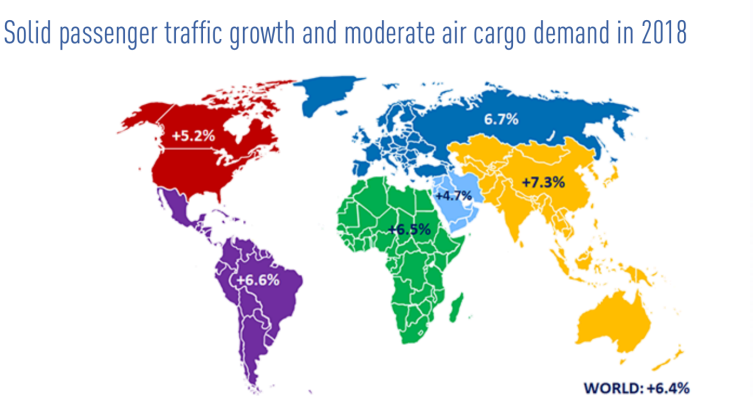
Lecture 2: Basic summary statistics and data visualisation

How do you describe data succinctly?

- Compute numbers that summarise the data
- Make pictures to look at; humans are very good at analysing information in a visual format

- Types of data
  - Continuous
    - Math operations can be performed
  - Discrete
    - Categorical (no ordering)
    - Ordinal (order is explicitly imposed)

Data summarisation



https://www.icao.int/Newsroom/Pages/Solid-passenger-traffic-growth-and-moderate-air-cargo-demand-in-2018.aspx

Raw data:

Passenger-id,
Name,
Time,
Airline,
Flight No,
Seat,
Class,
Meal choice,
Departure date yyyy-mm-dd,
Departure time,
Departure location,
Departure terminal,
Departure gate,
Arrival date yyyy-mm-dd,
Arrival time,
Arrival location,
Arrival terminal,
Arrival terminal,
Arrival gate

Data summarisation

https://gis.icao.int/eganp/jpgfc/FLOWCHARTFIRbackgroundproposewhite.jpg

Simple descriptive statistics

- Count (N)
- Sum
- Min
- Max
- Average
- Frequency
- Variance, Standard deviation

Central tendency

Data has a distribution

Typical value, centre, or location of the distribution

- Mean
- Median
- Mode

Dispersion (variability)

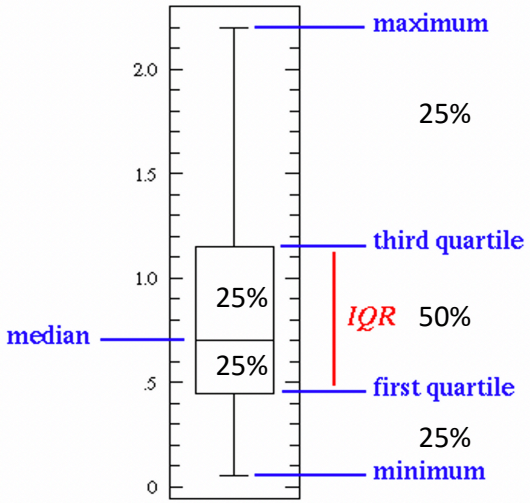
The extent to which a distribution is stretched

- Range
- Variance
- Standard deviation
- Quartiles and percentiles

Why visualisation?

- Converting data into a visual format
  - Reveals the structure of the data, information in the data
  - Characteristics of the data, relationships between objects or relationships between features
  - Simplifies the data
- See patterns
- Spot anomalies, outliers
- Visualisation can help show data quality
- Basic visualisations: boxplots, histograms, scatter plots, bar plots

Boxplot



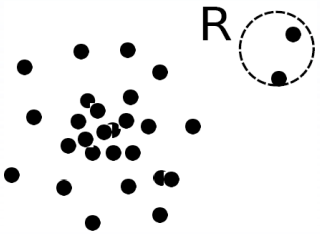
Shows distribution based on a 5-number summary of a set of data points (e.g. distance)

- Median: the middle number of the data
- First quartile(Q1): middle number of the data below the medium.
- Third quartile(Q3): middle number of the data above the medium
- Minimum: the smallest number of the data
- Maximum: the largest number of the data

- IQR = interquartile range = $Q3-Q1$

Outlier analysis

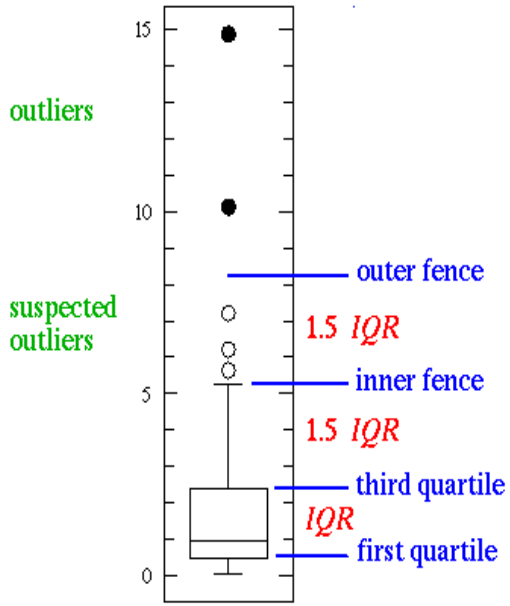
- **Outlier**: A data object that **deviates significantly** from the normal objects as if it were **generated by a different mechanism** (Hawkins, 1980)
  - Ex.: Unusual credit card purchase, sports: Michael Jordan, Lance Franklin, …
- From a statistics perspective
  - Normal (non-outlier) objects are generated using some statistical process
  - The outlier objects deviate from this generating process



Why do we care?

- Compute the average age of people in this room
  - Skewed results
- Compute the average salary of people in this room
  - What if Donald Trump is in the audience?
- Compute the average bank account withdrawal
  - What if the account has been hacked?

Outliers and Tukey Boxplots



- Median: the middle number of the data
- First quartile(Q1): middle number of the data below the medium.
- Third quartile(Q3): middle number of the data above the medium
- IQR = interquartile range = $Q3\text{-}Q1$

Whiskers (inner fence)

- 'Minimum': Lowest number within $1.5 \times IQR$ of Q1
- 'Maximum': Highest number within $1.5 \times IQR$ of Q3

Suspected outliers (open black)

- $> 1.5 \times IQR$ above Q3 or below Q1

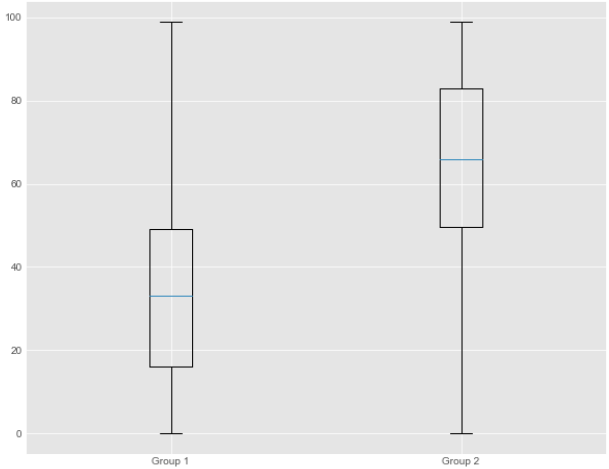
Outliers (filled black)

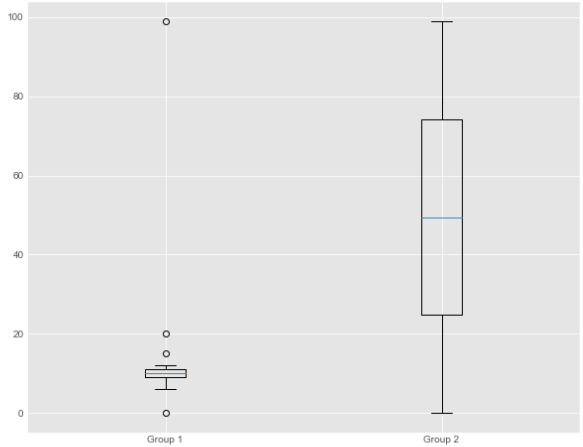
- $> 3 \times IQR$ above Q3 or below Q1

outlier $< Q_1 - 1.5\,IQR$ or $> Q_3 + 1.5\,IQR$

Boxplots – patterns

- Symmetric or skewed? Tightly or loosely grouped?





Boxplots – patterns

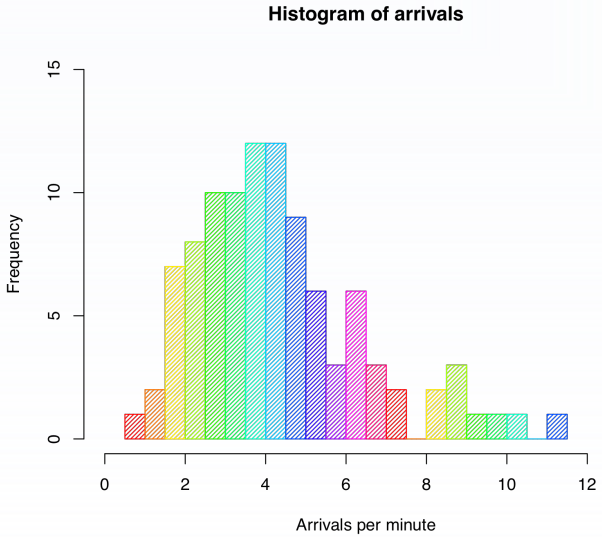
Do we know the count of the data?

No

Can 2 sets of data with different distribution result in the same boxplot?

Yes

Histograms



By DanielPenfield - Own work, CC BY-SA 3.0, https://commons.wikimedia.org/w/index.php?curid=9401898

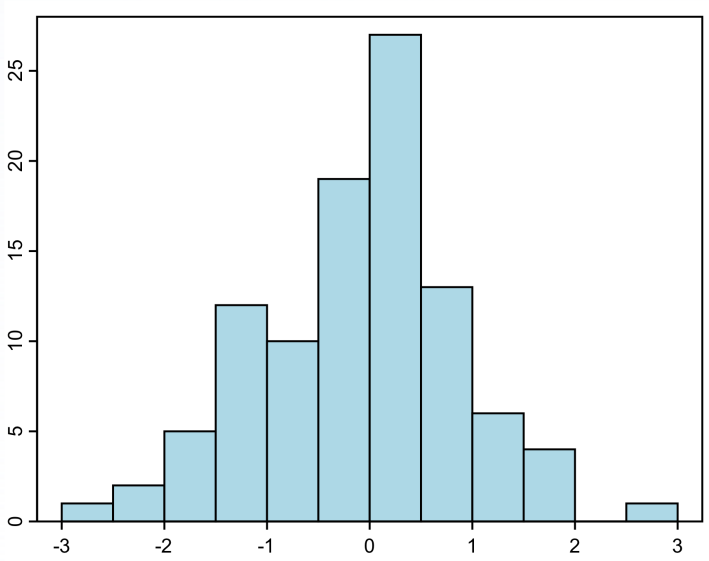
Also shows the distribution of the data points

- **Frequency** distribution of a set of continuous data points.
- Inspect the underlying distribution (shape), is it normal? skewed? outliers?

Histograms with equal-width bins

- Commonly used histograms
- X-axis: Divide the range of values into consecutive, non-overlapping, and equal width intervals.
- Y-axis: **height proportional to the frequency of the bin**

Histograms with equal-width bins

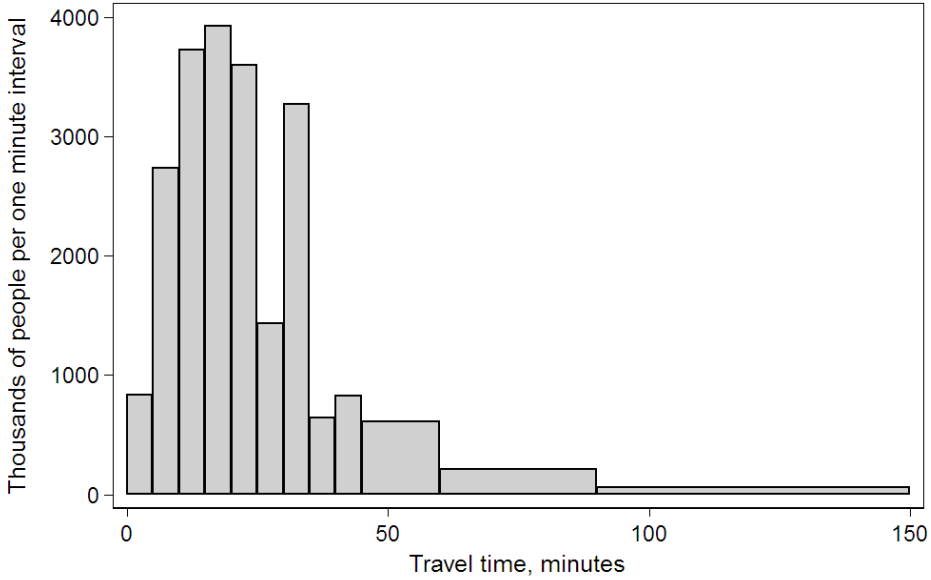


By Jkv - Own work, Public Domain, https://commons.wikimedia.org/w/index.php?curid=1886663

Histogram with variable-width bins

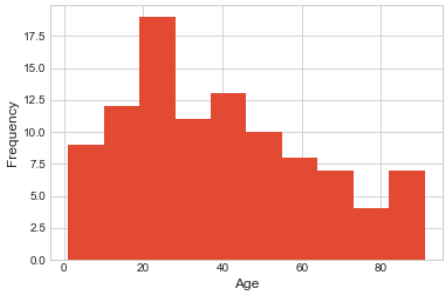
- X-axis: Divide the range of values into consecutive, non-overlapping, and variable width intervals.
- Y-axis: height proportional to frequency density—the number of cases per unit of the variable. The rectangle has its **area proportional to the frequency**

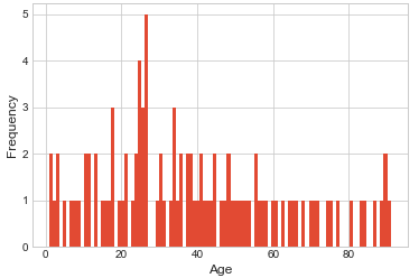
Histogram with variable width bins

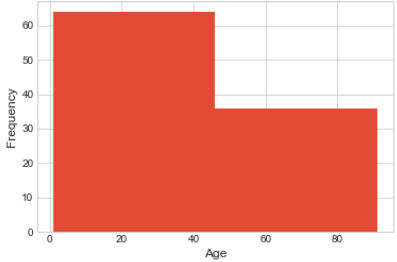


Histograms – cont.

- Histograms of the same dataset may look different with different bins selected.
- Problem: Hard to choose an appropriate bin size for histogram
  - Too small → normal objects in empty/rare bins, false positive
  - Too big → outliers in some frequent bins, false negative

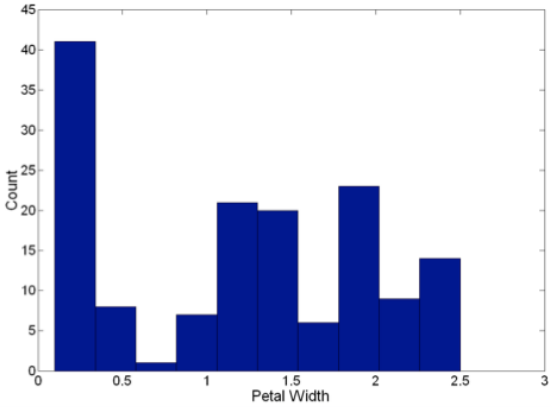


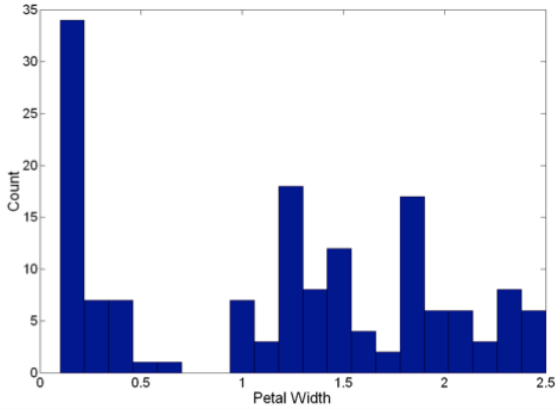




Histogram – number of bins

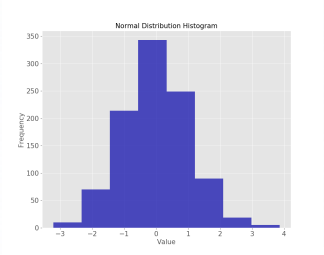
Petal Width (10 and 20 bins, respectively)

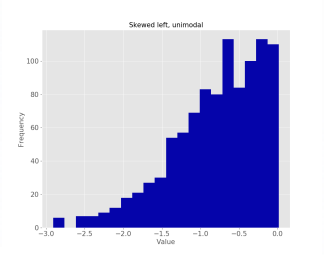


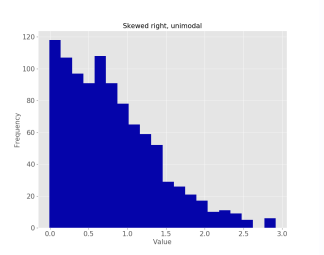


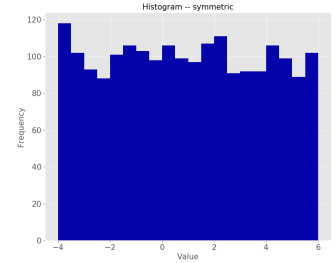
Histograms – patterns

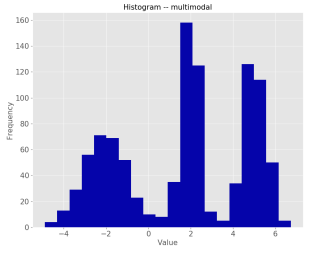
- Symmetric? Left/right skewed, unimodal, bimodal, multimodal?

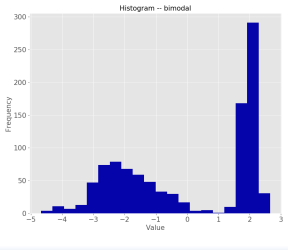












Outliers and histograms

- Paternity case: "The study of outliers", V. Barnett, Journal of the Royal Statistical Society, 27(3), 1978

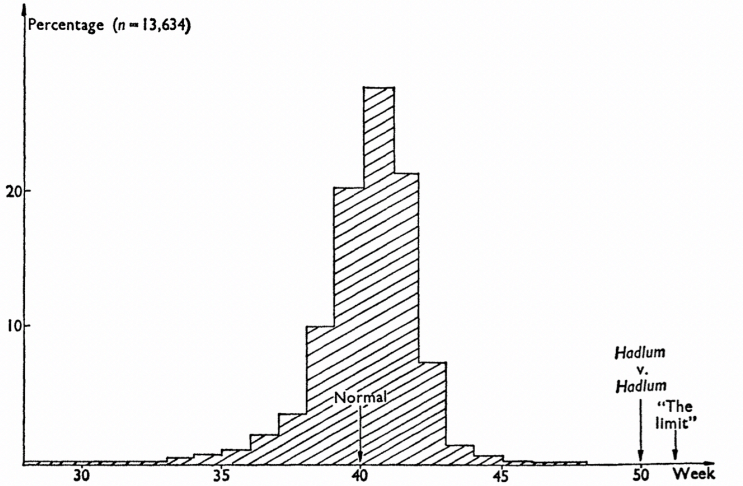
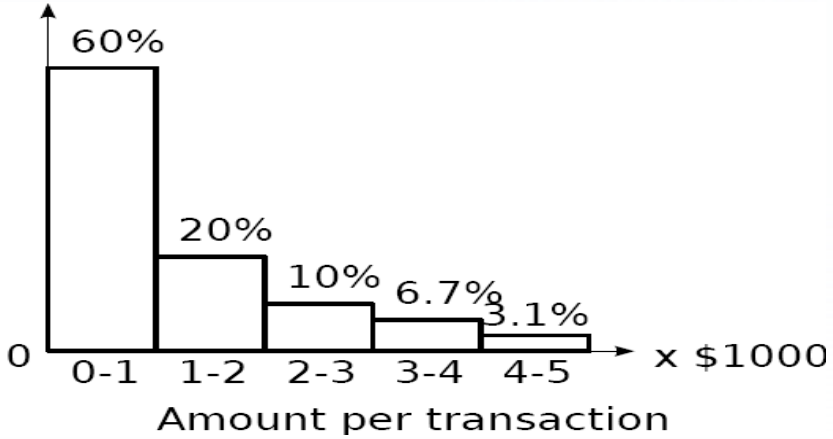


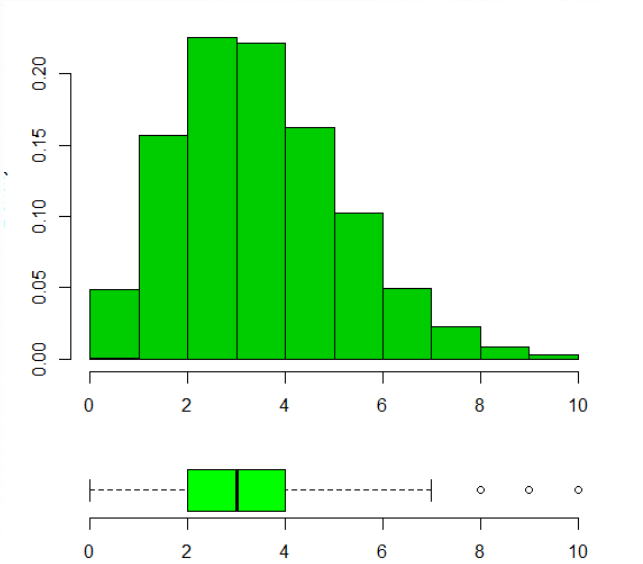
FIG. 1. Distribution of human gestation periods.

Outlier detection using histogram

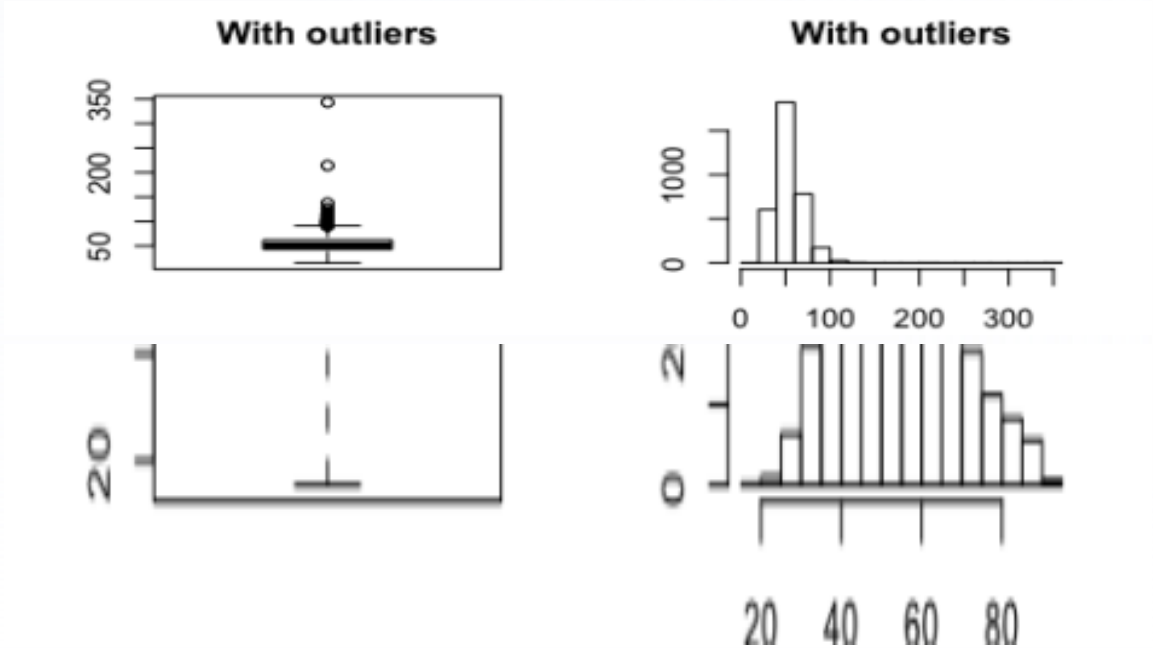
- Figure shows the histogram of purchase amounts in transactions
- A transaction in the amount of $7,500 is an outlier, since only 0.2% transactions have an amount higher than $5,000



Align histogram with boxplot

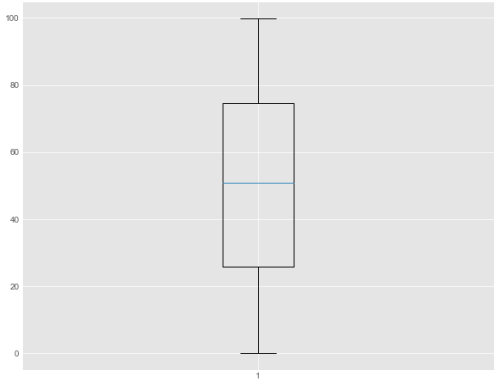


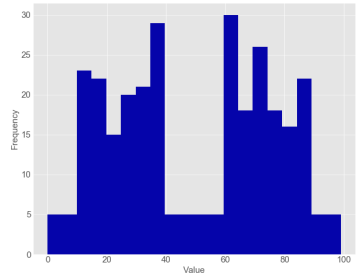
Outlier Check



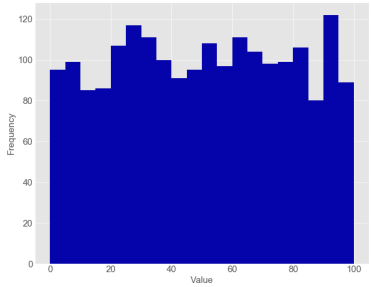
Histograms vs boxplots

Is this the boxplot of A or B?





A

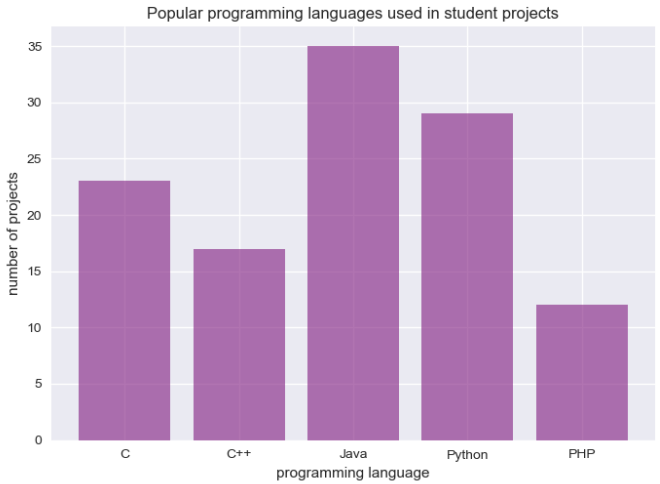


B

Bar plots

- Summarise data points over a categorical variable.

  X-axis: categorical variable
  Y-axis: numeric quantity



Bar plots vs histograms

- Histograms:
  - X-axis is intervals of a numeric variable
  - Y-axis is the frequency or frequency-density
  - Only sensible to be ordered in one way
- Bar plots:
  - X-axis is a categorical variable
  - Y-axis is a numeric quantity
  - Can be in any order

They look similar but they have different grammars.

Scatter plots

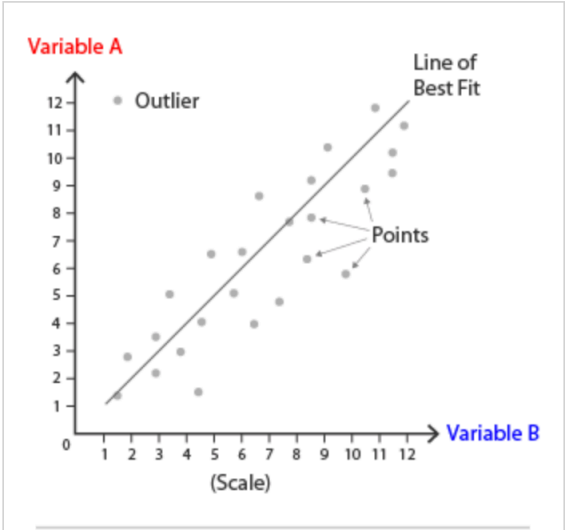
https://www.data-to-viz.com/graph/scatter.html

- Two numeric variables
- E.g. Sale price vs Ground living area

Scatter plots

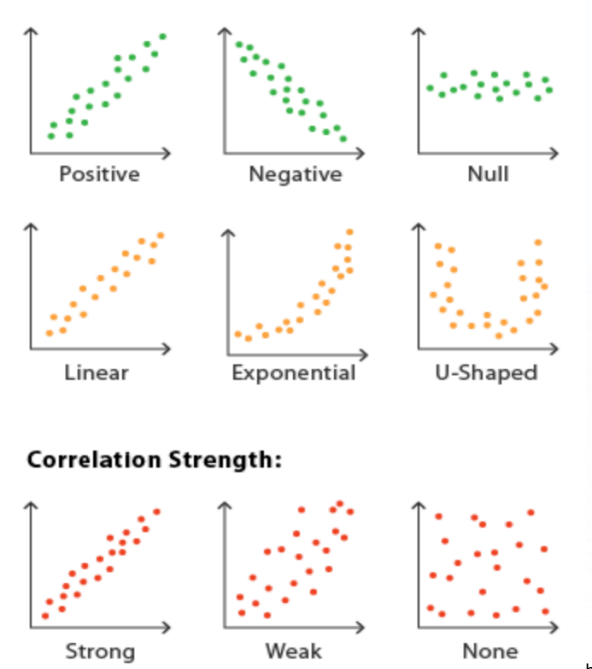
- X-axis: one numeric variable
- Y-axis: the other numeric variable
- A dot is a data point with 2-values as the x, y coordinates.

Anatomy



https://datavizcatalogue.com/methods/scatterplot.html

Scatter plots – patterns



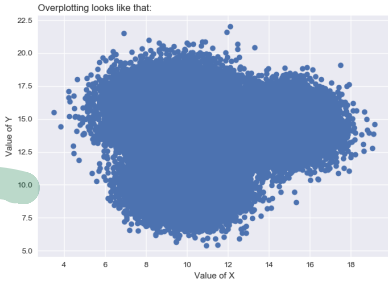
Relationship between two variables

https://datavizcatalogue.com/methods/scatterplot.html

‘Overplotting’ in scatter plots

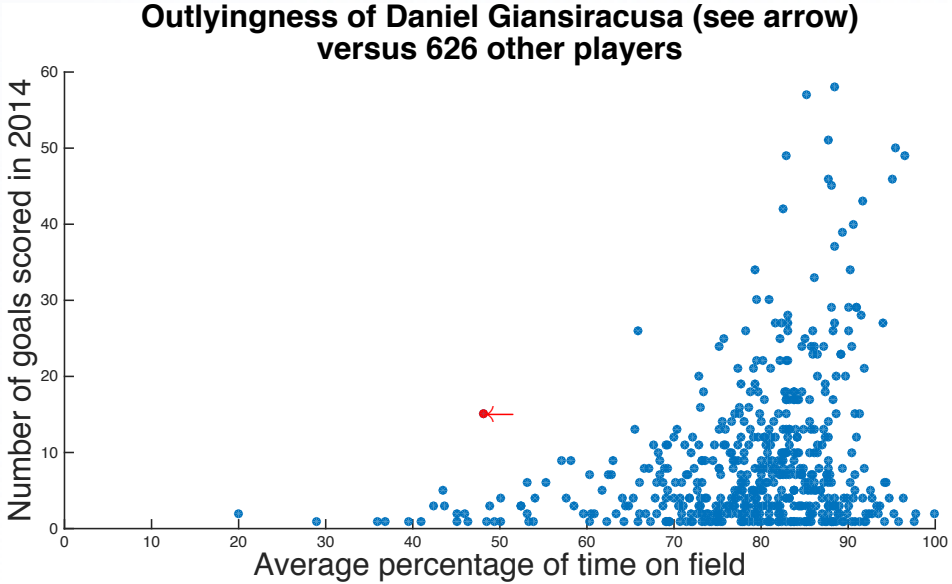
When there are many data points, dots tend to overlap

- Reduce dot size
- Sampling
- Jitter (for moderate overplotting)
- Use other plots



https://python-graph-gallery.com/134-how-to-avoid-overplotting-with-python/

Outliers and scatter plots



Elements of a good visualisation

- Meaningful title.
- Appropriate scales, annotation
- Suitability to the dataset and the context of the data question
- Can be interpreted on its own.
  - Caption can be used to explain the context, the dataset, and a brief interpretation of plot, where appropriate.
- Has no redundant, information unimportant to the plot.
- Customisation and improvisation may be required.

Summary

- Get to know your data
- Think
- Try different visualisations
- Will be bread and butter for you later on