Visual Interfaces to Engage Analysts in Data Engineering

Yaji Sripada

Department of Computing Science

University of Aberdeen

Automating Data Science (De Bei et al, 2022)

CDA=Confirmatory Data Analysis

EDA=Exploratory Data Analysis

Data Engineering	Model Building (CDA)
Data Exploration (EDA)	Exploitation

Less dependent on domain context

More dependent on domain context

More open-ended

Less Open-ended

Forms of Automation in Data Science

- Mechanization
  - E.g. AutoML – the success story of Automated data science
- **Composition**
  - Workflow automation platforms
- **Assistance**
  - Tools that support data scientists

Forms of Automation in Data Engineering (DE)

- Mechanization
  - DE can’t be end-to-end automated like Model building
  - Because DE is a dynamic target – every batch of data may suffer from a different set of issues!
- **Composition**
  - Workflow automation platforms (e.g. Airflow, already covered in the course)
- **Assistance**
  - Tools that support data engineers

Two types of Assistance for Human-Computer Collaboration in DE

- Backend Assistance
  - Build libraries that can automatically (or semi-automatically) preprocess raw data
- Frontend Assistance
  - Data engineers themselves can work with data if they are effectively engaged in the DE process
- Human-Computer Collaboration is achieved by combining both these types of assistance

Backend Assistance

- Develop Algorithms that preprocess input data
- Examples
  - SimpleImputer class in Scikit-Learn replaces missing values in a column using the mean value of the data column
  - Drop_duplicates function in Pandas drops duplicate records in dataframes
- Ultimately, we want significant improvements here
  - Only then can significant levels of efficiency gains in DE be achieved

Frontend Assistance

- Data engineer input needed, even in the presence of robust backend assistance
  - Every organization and their datasets may have unique contexts that cannot be built into backend assistance
  - Data engineers need to make decisions about the DE pipelines and the algorithms they call
- Data engineers can only make these decisions if they are engaged in the DE process.
- Humans cannot understand data without data visualisations/information visualisations (InfoVis)

InfoVis

- InfoVis is the process of representing data visually
    - Visual presentation of abstractions or relationships underlying input data
    - To enable users to gain useful insights into the data
- Focus is on designing a data representation scheme
    - That makes the underlying ‘information’ visible & comprehensible to the user
- For rendering the representation scheme
    - Computer graphics technology is exploited
- Examples
    - Newsmap
    - Touchgraph

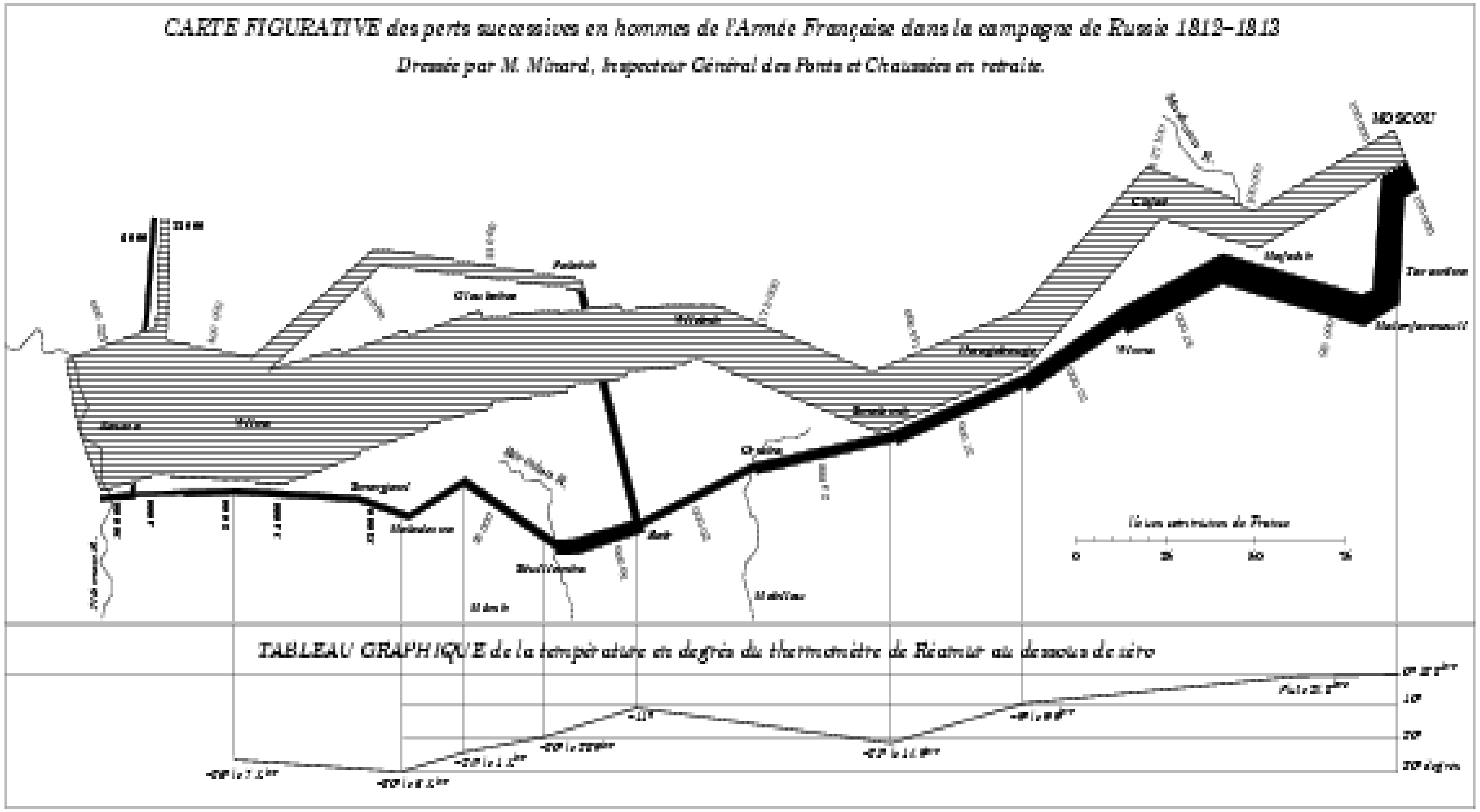
InfoVis (2)

- Good InfoVis techniques are based on
  - Good understanding of the information structures underlying the data
  - Good understanding of human perception and cognition
  - Good graphics libraries
- Limited screen sizes pose a serious challenge for using IV on very large data sets
- Therefore, the main task is to pack large information into a simple graphic
  - Highlighting all the required (important) information
- Creative art?

Textbook Example 1

- Napolean's 1812 campaign on Russia
- Input data
  - Size of army
    - at the start of the campaign = 442,000
    - at the end of the campaign = 10,000
  - Location of the army (2 dimensions)
  - Direction of the army's movement
  - Temperature and
  - Time

Minard’s Drawing



Created in 1861 by French engineer Charles Joseph Minard

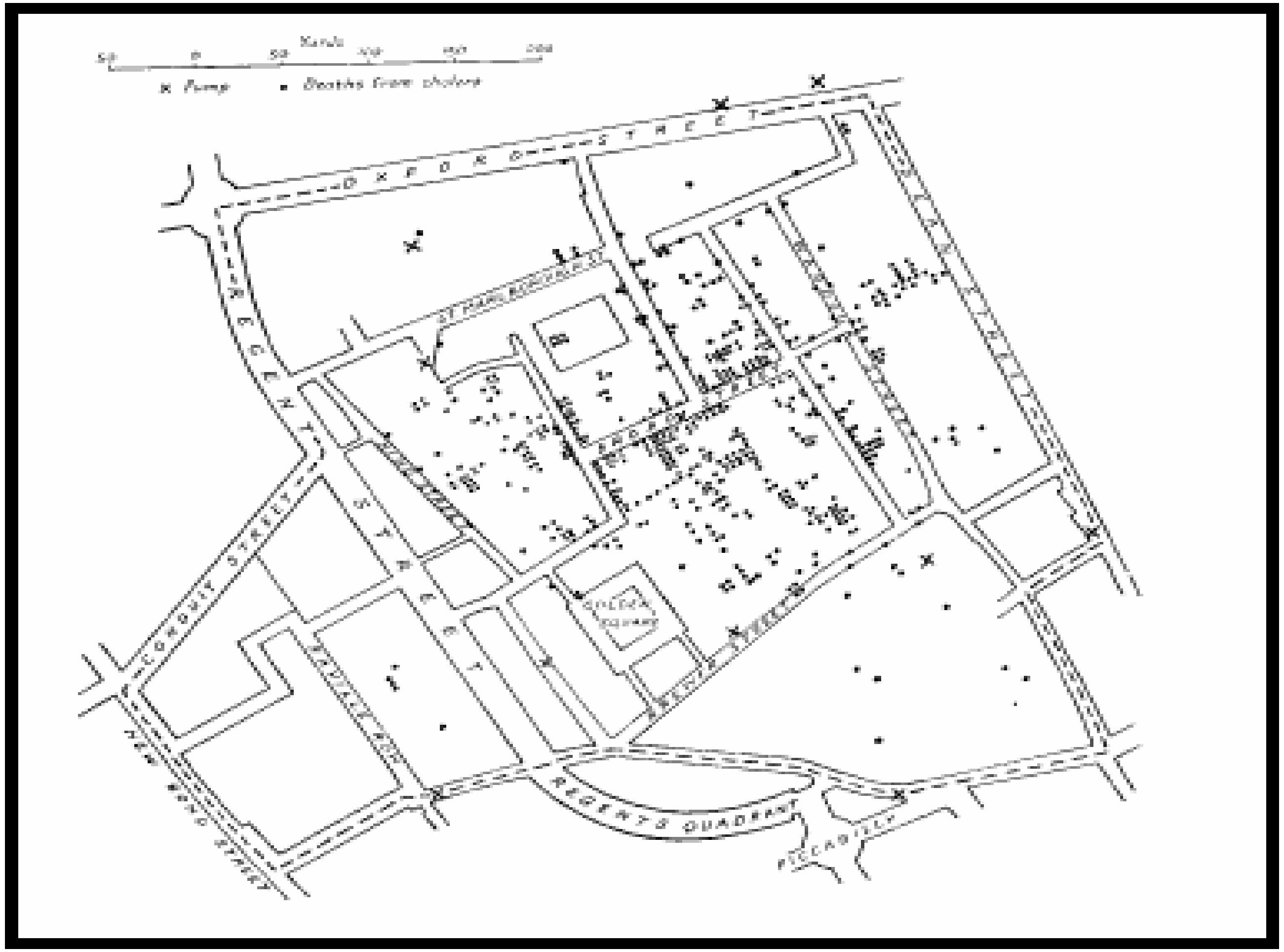
Minard’s Drawing (2)

- Considered the best graphic ever produced
  - Inspiration for modern InfoVis researchers
- Plots all the data corresponding to all six input variables
- Clearly shows the message underlying the input data
  - Gradual reduction in the size of the army
  - Linked to the gradual fall in temperatures
- Input data is complex
- Yet, the most important information is abstracted out and presented in a simple graphic

Textbook Example 2

- London cholera epidemic of 1854
- At that time, two hypotheses of causes of cholera:
  - Cholera is related to miasmas concentrated in the swampy areas of the city
  - Cholera is related to ingestion of contaminated water
- Input Data
  - Locations of deaths due to cholera
  - Locations of water pumps

Dr Snow's Cholera Map



Dots
locate
deaths
due to
cholera

Crosses
Locate water
pumps

Dr Snow's Cholera Map (2)

- Plotting the input data on the map helped Dr Snow
  - to detect the epicentre of the epidemic
  - Close to a pump on Broad Street
- Considered a classic case of visualization helping reasoning with data

Design & Technology

- There are two requirements for developing visualizations
  - Graphic Design
    - mapping information (raw or filtered) into a graphic
      - Mapping data/information to display variables
        - Position, orientation, size, motion, colour etc.
  - Technology
    - achieving the design programmatically
      - Graphics programming, flash programming etc

Graphic Design

- Mapping
  - Data to some graphical element
    - Such as a cross.
  - data attributes to the attributes of the graphical element
    - Such as colour, size, shape etc.
- Order of priority for representing quantitative data
  - Position
  - Length
  - orientation
  - Size
  - colour

Inputs to the design process

- Data - size and data type
- User Task
- User characteristics
- System resources - PC vs Graphics workstation
- Standards/guidelines

Designing Information Visualizations

- Gospel like guidelines
  - If the underlying data is simple, keep the graphic simple
  - If the underlying data is complex, make the graphic look simple (e.g., Minard's Graphic)
  - Always tell the truth - Do not distort the data
  - Maximize the data-ink ratio (Edward Tufte, www.edwardtufte.com)
    - Data-ink ratio= data-ink/total ink used on the graphic

Visual Information Seeking Mantra

- Modern visualizations are highly interactive
  - Users wish to seek information visually and interactively
- Visual Information Seeking Mantra recommends designing interfaces using the following guideline

  “Overview first, zoom and filter, then details on demand”
- Details of the mantra are given in the Task by Type Taxonomy (TTT) proposed by Prof. Shneiderman, HCI Lab, University of Maryland (UMD)
- TTT is a framework for organizing visualizations. Involves
  - 7 tasks and
  - 7 data types

7 Tasks

- The 7 interactive tasks users wish to perform:
  - **Overview**: Gain an overview of the entire collection.
  - **Zoom** : Zoom in on items of interest
  - **Filter**: filter out uninteresting items.
  - **Details-on-demand**: Select an item or group and get details when needed.
  - **Relate**: View relationships among items.
  - **History**: Keep a history of actions to support undo, replay, and progressive refinement.
  - **Extract**: Allow extraction of sub-collections and of the query parameters.

7 Data Types

- 1 D Linear
- 2D Map
- 3D World
- Multi-dimensional
- Temporal
- Tree
- Network

Graphics Technology

- Computer Graphics is a major field of Computing Science.
- Two Approaches:
  - Raster Graphics
  - Vector Graphics (e.g., OpenGL and SVG)
- More in the Lab Class

SVG

- Scalable Vector Graphics
  - An XML-based Web Language to textually specify vector graphics
  - E.g. SVG specification of a circle

```
<svg width="300" height="200" xmlns="http://www.w3.org/2000/svg">
  <circle cx="125" cy="110" r="20" fill="red" />
</svg>
```

- W3C Recommendation
- Browser support for SVG content
  - Firefox provides built in support
  - IE needs an Adobe Plug-in
- SVG content can be created
  - Using text editors (static)
  - Programmatically (dynamic)

SVG in an HTML page

- Three methods
- Using the <embed> tag

  <embed src=“circle.svg" width="300" height="100" type="image/svg+xml" pluginspage="http://www.adobe.com/svg/viewer/install/" />

- Using the <object> tag
  - <object data="rect.svg" width="300" height="100" type="image/svg+xml" codebase="http://www.adobe.com/svg/viewer/install/" />
- Using the <iframe> tag
  - <iframe src="rect.svg" width="300" height="100"> </iframe>

Exploratory Data Analysis (EDA)

- Data Scientists need to gain useful insights into the input data first
  - We saw this in the previous lecture
- Exploratory Data Analysis (EDA) helps to achieve this
- EDA offers several techniques to comprehend data
- But EDA is more than a library of data analysis techniques
- EDA is an approach to data analysis
- EDA involves inspecting data without any assumptions
  - Mostly using information graphics
  - Modern InfoVis tools use many of the EDA techniques which we study here
- Insights gained from EDA help select the appropriate data analysis (InfoVis) techniques
- **We simply repurpose EDA methods to help with building frontend assistance for DE**

Descriptive Statistics

- Descriptive statistical methods quantitatively describe the main features of data
  - We learn the key ideas using visuals
- Main data features
  - measures of central tendency – represent a ‘center’ around which measurements are distributed
    - e.g. mean and median
  - measures of variability – represent the ‘spread’ of the data from the ‘center’
    - e.g. standard deviation
  - measures of relative standing – represent the ‘relative position’ of specific measurements in the data
    - e.g quantiles

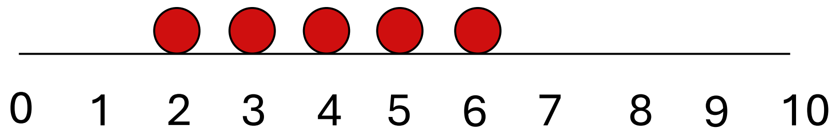
Mean

- Sum all the numbers and divide by their count

  $x = (x_1+x_2+ \ldots +x_n)/n$

- For the example data
  - Mean = (2+3+4+5+6)/5

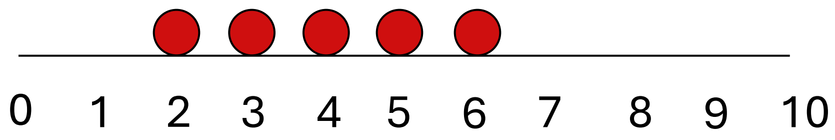
    = 4
  - 4 is the ‘center’
- The information graphic used here is called a dot diagram



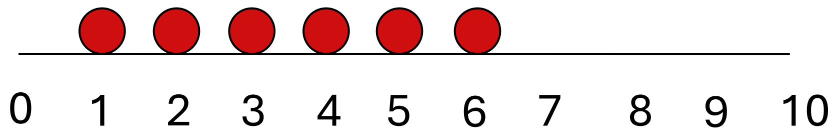
Median

- The exact middle value
- When count is odd just find the middle value of the sorted data
- When count is even find the mean of the middle two values
- For example data 1
  - Median is 4
  - 4 is the ‘center’
- For example data 2
  - Median is (3+4)/2 = 3.5
  - 3.5 is the ‘center’

Data 1



Data 2

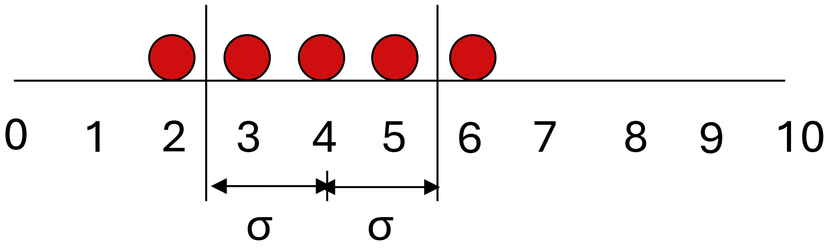


Standard Deviation

- Computation steps
  - Compute mean
  - Compute each measurement's deviations from the mean
  - Square the deviations
  - Sum the squared deviations
  - Divide by (count-1)
  - Compute the square root

$\sigma = \sqrt{}(\sum(x_i\text{-}x)^2)/(n\text{-}1)$

Data 1



Mean = 4

Deviations: -2, -1, 0, 1, 2

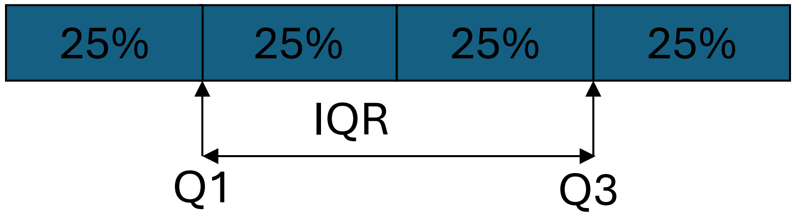
Squared deviations: 4, 1, 0, 1, 4

Sum = 10

Standard deviation = √(10/4) = 1.58

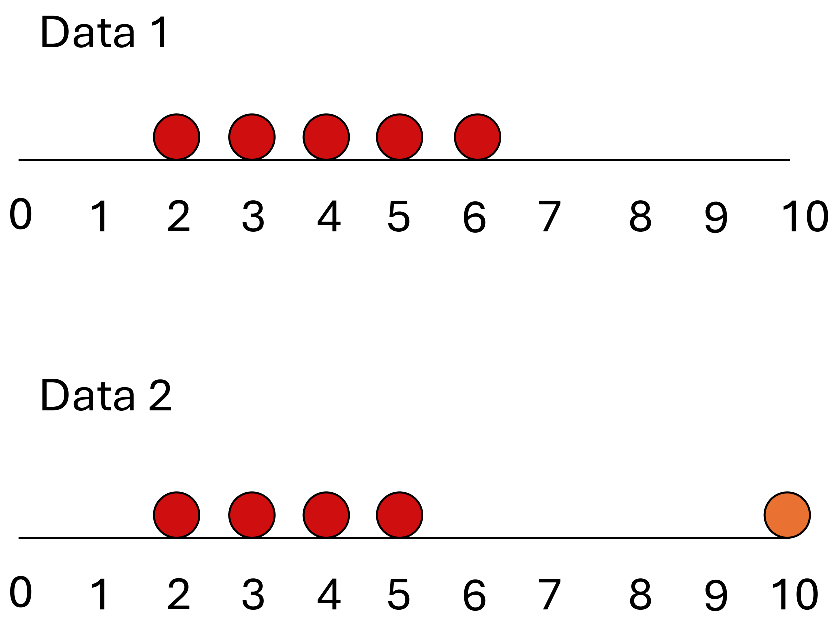
Quartiles

- Median is the 2nd quartile
- 1st quartile is the measurement with 25% measurements smaller and 75% larger – lower quartile (Q1)
- 3rd quartile is the measurement with 75% measurements smaller and 25% larger – upper quartile (Q3)
- Inter quartile range (IQR) is the difference between Q3 and Q1
    - Q3-Q1



Median VS Mean

- When data has outliers median is more robust
  - The blue data point is the outlier in data 2
- When data distribution is skewed median is more meaningful
- For example data 1
  - Mean=4 and median=4
- For example data 2
  - Mean=24/5 and median=4



Stem and Leaf Plot

- This plot organizes data for easy visual inspection
  - Min and max values
  - Data distribution
- Unlike descriptive statistics, this plot shows all the data
  - No information loss
  - Individual values can be inspected
- Structure of the plot
  - Stem – the digits in the largest place (e.g. tens place)
  - Leaves – the digits in the smallest place (e.g. ones place)
  - Leaves are listed to the left of stem separated by '|'
- Possible to place leaves from another data set to the right of the stem for comparing two data distributions

Data

29, 44, 12, 53, 21, 34, 39, 25, 48, 23, 17, 24, 27, 32, 34, 15, 42, 21, 28, 37

Stem and Leaf Plot

1 | 2 7 5

2 | 9 1 5 3 4 7 1 8

3 | 4 9 2 4 7

4 | 4 8 2

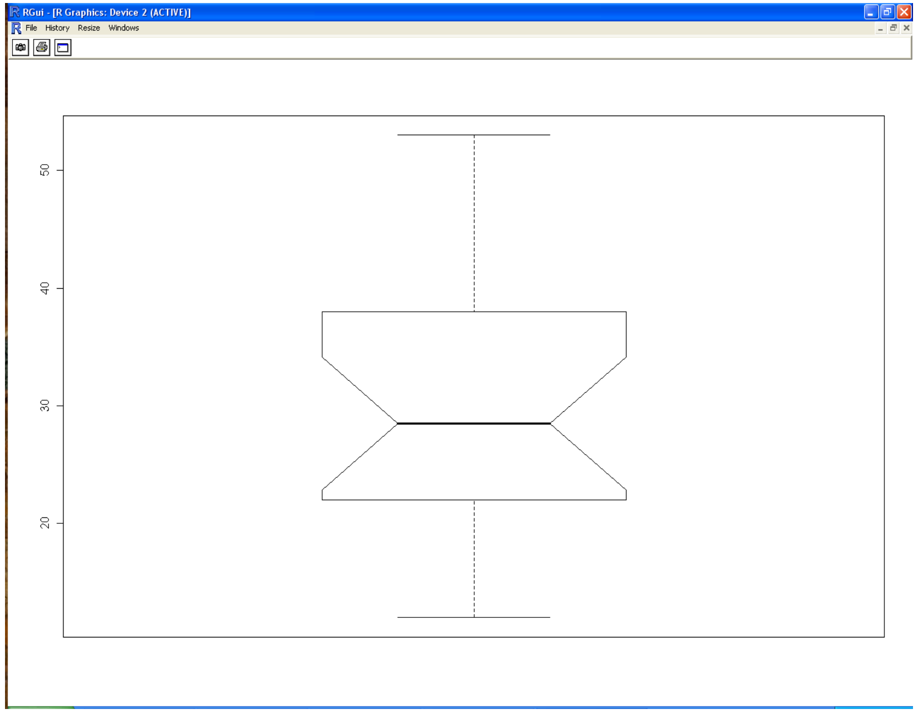
5 | 3

Box Plot

- A five value summary plot of data
  - Minimum, maximum
  - Median
  - 1st and 3rd quartiles
- Often used in conjunction with a histogram in EDA
- Structure of the plot
  - Box represents the IQR (the middle 50% values)
  - The horizontal line in the box shows the median
  - Vertical lines extend above and below the box
  - Ends of vertical lines called whiskers indicate the max and min values
    - If max and min fall within 1.5*IQR
  - Shows outliers above/below the whiskers

Data

29, 44, 12, 53, 21, 34, 39, 25, 48, 23, 17, 24, 27, 32, 34, 15, 42, 21, 28, 37



Standardization

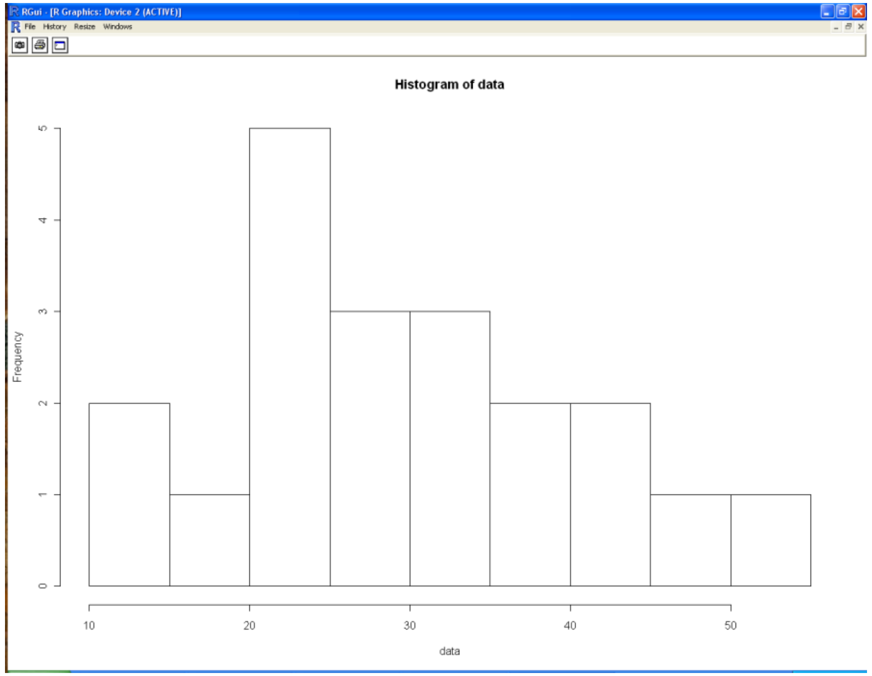
- Data sets originate from several sources, and there are bound to be differences in measurements
    - Comparing data from different distributions is hard
- Standard deviation of a data set is used as a yardstick for adjusting for such distribution-specific differences
- Individual measurements are converted into what are called standard measurements, called z scores
- An individual measurement is expressed in terms of the number of standard deviations, σ it is away from the mean, μ
- Z score of $x = (x- \mu)/ \sigma$
    - Formula for standardizing attribute values
- Z scores are more meaningful for comparison
- When different attributes use different ranges of values, we use standardization

Histogram/Bar Chart

- Graphical display of <u>frequency distribution</u>
  - Counts of data falling in various ranges (bins)
  - Histogram for numeric data
  - Bar chart for nominal data
- Bin size selection is important
  - Too small – may show spurious patterns
  - Too large – may hide important patterns
- Several Variations possible
  - Plot relative frequencies instead of raw frequencies
  - Make the height of the histogram equal to the ‘relative frequency/width’
    - Area under the histogram is 1
- When observations come from continuous scale histograms can be approximated by continuous curves

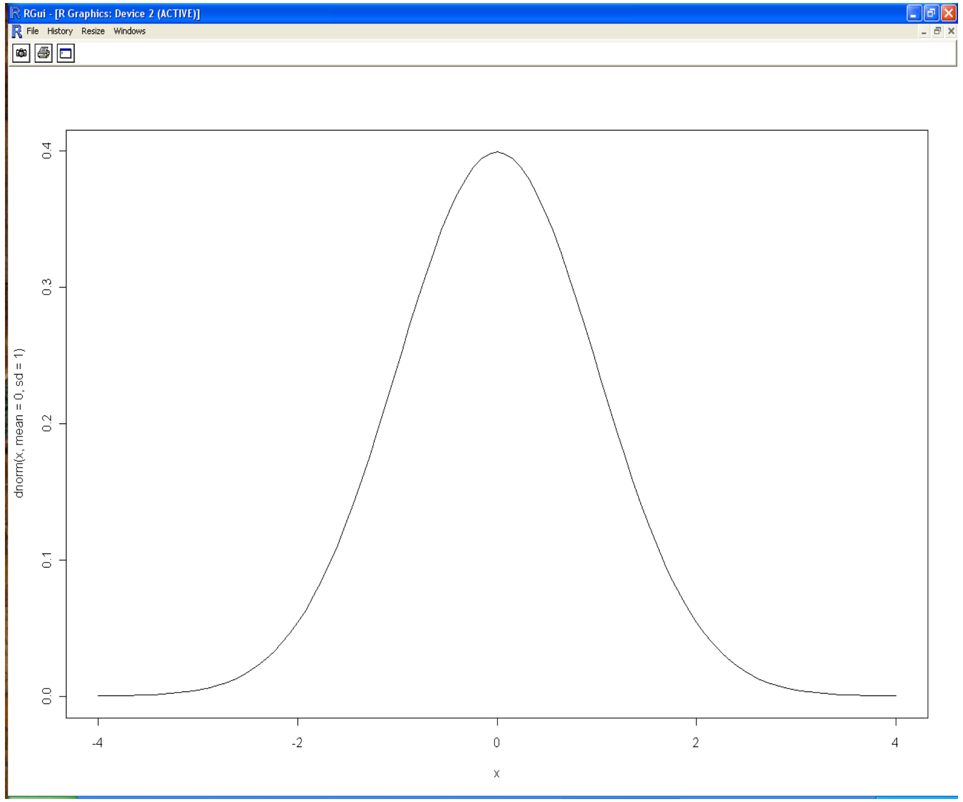
Data

29, 44, 12, 53, 21, 34, 39, 25, 48, 23, 17, 24, 27, 32, 34, 15, 42, 21, 28, 37



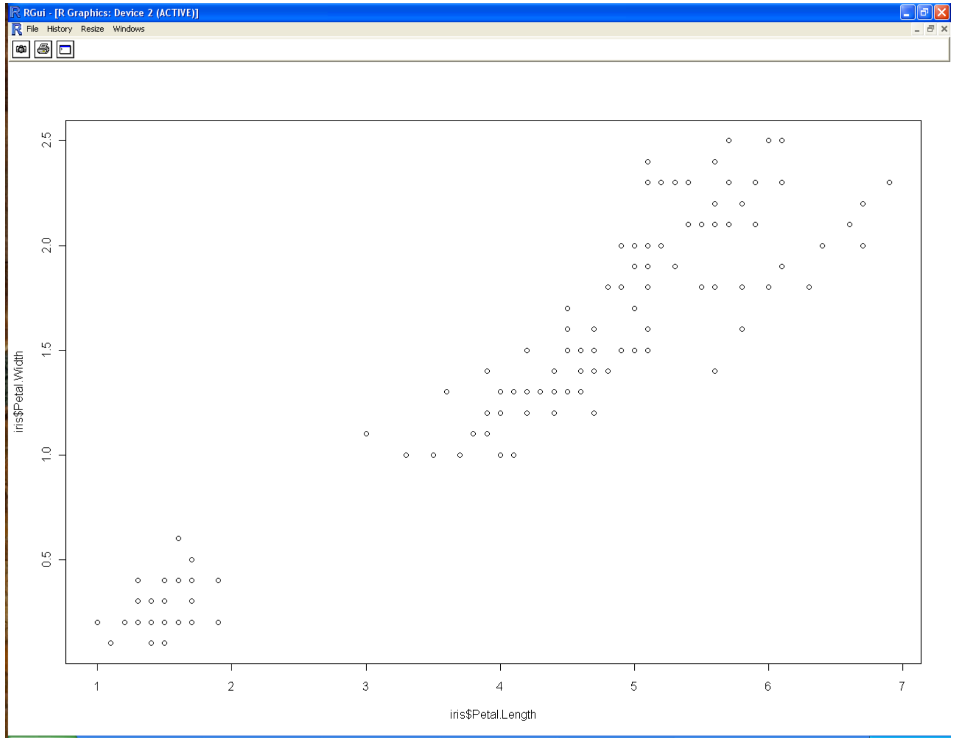
Normal Distribution

- Distributions of several data sets are bell shaped
  - Symmetric distribution
  - With peak of the bell at the mean, μ of the data
  - With spread (extent) of the bell defined by the standard deviation, σ of the data
- For example, height, weight and IQ scores are normally distributed
- **The 68-95-99.7% Rule**
  - 68% of measurements fall within $\mu - \sigma$ and $\mu + \sigma$
  - 95% of measurements fall within $\mu - 2\sigma$ and $\mu + 2\sigma$
  - 99.7% of observations fall within $\mu - 3\sigma$ and $\mu + 3\sigma$



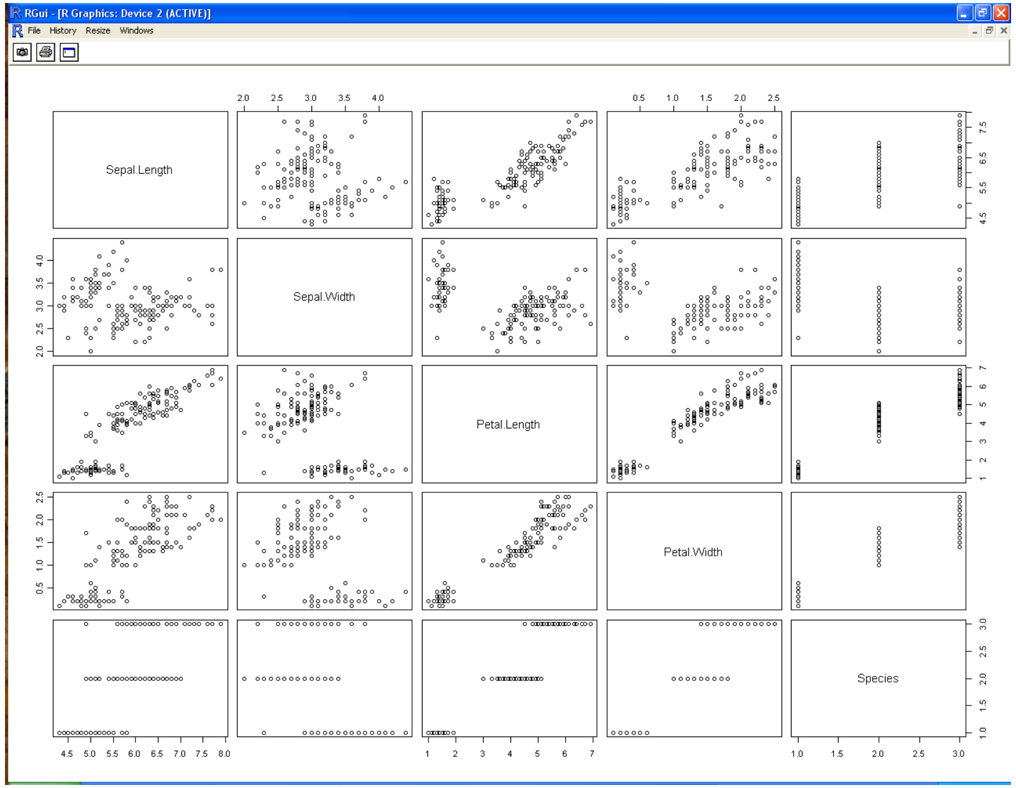
Scatter Plot

- Scatter plots are two dimensional graphs with
    - explanatory attribute plotted on the x-axis
    - Response attribute plotted on the y-axis
- Useful for understanding the relationship between two attributes
- Features of the relationship
    - strength
    - shape (linear or curve)
    - Direction
    - Outliers
- Scatter plot of iris$Petal.Width against iris$Petal.Length (refer to practical 1 about IRIS data) is shown here



Scatter Plot Matrix

- When multiple attributes need to be visualized all at once
  - Scatter plots are drawn for every pair of attributes and arranged into a 2D matrix.
- Useful for spotting relationships among attributes
  - Similar to a scatter plot
- Scatter plot matrix of IRIS data is shown here
  - Attributes are shown on the diagonal
- Later in the course we learn to use parallel coordinates for plotting multi-attribute data



EDA Answers Questions

- All the techniques presented so far are the tools useful for EDA
- But without an understanding built from the EDA, effective use of tools is not possible
  - A detective investigating a crime scene needs tools for obtaining fingerprints.
  - Also needs an understanding (common sense) to know where to look for fingerprints
    - Are doorknobs better places than door hinges?
- EDA helps to answer a lot of questions
  - What is a typical value?
  - What is the uncertainty of a typical value?
  - What is a good distributional fit for the data?
  - What are the relationships between two attributes?
  - etc

References

1. Tijl De Bie, Luc De Raedt, José Hernández-Orallo, Holger H. Hoos, Padhraic Smyth, Christopher K. I. Williams, Automating Data Science, Communications of the ACM, March 2022, Vol. 65 No. 3, Pages 76-87.(2022)