[redacted]

Chapter 1: Introduction

Exploratory data analysis (EDA)

- techniques for • summarising, visualising & reviewing data,

building an inhition for how the underlying process that generated data works.

- Basic approach
  - generate questions about you data.
  - search for answers by visualising, transforming and modelling your data.
  - use what you learn to refine your questions &/or generate questions.
- Data-driven (model-free) - What can the data tell us? as
- Is fundamentally a creative process and a critical first step in analysing the data.
- Is often described as a philosophy, and there are no hard-and-fast rules for how you approach it.
- methods can be numerical (i.e., descriptive statistics), graphical, or tabular.
  - graphical methods: make it very easy to discover trends and patterns in a data set.

- In tabular data
  - expect each record or observation to represent a set of measurements of a single object or event.
  - Each type of measurement is called a variable or an attribute of the data.
  - The number of attributes is called the dimension of the data.

Visualisation

Why visualisation

- aid model construction and check plausibility of model assumptions.
  - pattern discovery: clusters, outliers, trends
  - contextual knowledge: expectations for dataset, explanations for patterns
  - action: humans learn and take action.
- It is intuitive
- It is fast
- It is flexible
- It is insightful

- conversion of numbers → images
  - convert numerical analysis into visual analysis

In general, we create visualisation to help us
- answer questions
- make decisions
- present argument or tell a story.

- what do you want your visualisation to show about your data?

1) distribution
- how a variable or variables in the dataset distribute over a range of possible values.

2) relationship
- how the values of multiple variables in the dataset relate.

3) composition
- how the dataset breaks down into subgroups.

4) comparison
- how trends in multiple variable or dataset compare.

• Visualisation is the graphic representation of information and data by using visual element like charts, graphs and maps.

• Three types of goals for visualisation
1) to explore
- nothing is known, visualisation used for data exploratio
2) to analyse
- there are hypotheses, visualisation used for verification or falsification.
3) to present
- "everything" known about the data, visualisation used for communication of results.

• Visualisation types
1) Infographics (information graph in short)
- are visual representation of facts, events or numbers.
• The visual patterns & trends
- are used in such a way that human cognition is enhanced.

2) scientific visualisation
- visualising results of simulations, experiments or observations
• frequently data is multi-dimensional.

3) Data visualisation
- are used for searching for interesting phenomena

- Data visualisation
  - tools provide accessible way to see and understand trends, outliers and the patterns in data.
  - is the translation of the data set into a meaningful and the easy to understand visual media.
  - presents objective, quantifiable information.
  - can be static or interactive.

Good & bad visualization examples

- Junkcharts

Good data visualisation

- The purpose of computing is insight, not numbers.
- The purpose of visualisation is insight, not pictures.
- Good data visualisation
  - makes data accessible
  - combines strengths of human & computers
  - enables insight
  - communicates

Graphical summaries of data.

- A Good picture is worth a 1,000 words
- There's more than 1 right way to go about visualising data,
  - but there are many, MANY wrong ways to do it.

How to make a bad graph?

- The aim of good data graphics:
  - Display data accurately & clearly
- Some rules for displaying data badly:
  - Display as little information as possible
  - Obscure what you do show (with chart junk)
  - use proud-3d & color gratuitously
  - make a pie chart (preferably in color & 3d)
  - Use a poorly chosen scale

Chart Junk

- a chart that does not represent data (or is a scale or label) as not just unnecessary, but harmful.
- Extraneous visual elements that distract from the message
  e.g. - heavy or dark grid lines,
  - ornamented chart axes & display frames,
  - pictures or icons within data graphs
- A related term is the data-to-ink ratio
  - which is a term to describe how many visual items
  - (how much "ink") representing data there are in a chart
  - in relation to how much there is overall.

Visualisation: What is it?

Let's try to define it:

- Attempt 1: Presenting complex data in a way that is easy to understand. But in what way?
- Attempt 2: By looking at a good visualisation you should be able to:
  - easily answer a lot of questions about the data
  - see features/ regularities / facts that make you ask more questions
  - 'read' what the presented wants to 'write'

Summary of the basic principles

- A plot is a map between data & visual elements.
- It consists of layers that share some common properties
- Each layer has
  - data - aesthetic mapping, - scales, - geometry
  - a statistical transformation & a coordinate system
  - together may define how the plot will look like.
- Layers & some additional properties
  - can be added on top of any graph
  - as long as there are no contradictions between the definitions

Statistical & Infographics

- Statistical graphics
  - are intended to show structure of data, and
  - to support statistical inferences
  - may need much concentration to understand, & many users may be confused...
- Infographics present the data as a story, or as an aesthetic object
  - may be informationally useless, but users love them.
- The best infographics display information well & tell a story.

Is the data any good?

- Dimensionality of data sets
  - univariate: measurement made on 1 variable per subject.
  - bivariate: measurement made on 2 variables per subject.
  - multivariate: measurement made on many variables per subject.

- Common issues with data:
  - missing values: how do we fill in?
  - wrong values: how can we detect & correct?
  - messy format
  - not usable: the data cannot answer the question posed.

Plan of course

- Bivariate plots (X Vs Y)
  - plot design; rescaling, rescaling, rescaling; many types of plotting; comparing distributions.
- Investigating a data-set
  - generating questions; answering questions; what to plot & how?
- Visual perception
- Multivariate data visualisation
  - principal components analysis
  - clustering
  - non-linear dimensionality reduction
    - such as stochastic neighbour embedding & t-SNE
- statistical tests, significance, & snooping
  - standard & home-made tests;
  - statistical significance & how to interpret it;
  - data-snooping.

Each graph tells a message

- A graph is like a sentence or paragraph.
  - It should have a single, clear purpose.
- Several graphs may be needed for one dataset.

- Scaling
  - linear
  - linear-logarithmic

Conclusions

- Real data is hard to understand.
- EDA
  - provides ways of presenting data that make the data easier to understand.
- Visualisations should clarify, not confuse.
- After making a visualisation
  - ask yourself if you understand the data better than
  - you did before looking at it.
- If the answer is no, then you might consider a different visualisation.
- Avoid chart junk & overly "flashy" graphics.
- sticking to the basic plot types
  - that can be made in R will help significantly
  - when you are starting out.