Exploratory Data Analysis

Data Exploration and Visualisation

Dr Zak Varty

Outline

1. What is Exploratory Data Analysis?
2. What is **not** Exploratory Data Analysis?
3. Issues around Exploratory Data Analysis.

What is an Exploratory Data Analysis?

EDA as a way to know your data

> Exploratory Data Analysis: quick and simple exerpts, summaries and plots to better understand a data set.

- Iterative, not put into production
- EDA notebooks can be helpful
- Document and share what is often an ad-hoc process
- Balance between reproducibility and time cose

EDA as a conversation starter

- An effective EDA sets a precedent for open communication with the stakeholder and project manager.
  - Establish rapport and trust and buy-in early in the project;
  - **Stakeholder:** subject-specific knowledge and data collection expertise;
  - **Manager:** prioritise projects for best **business** outcome.

EDA as project scoping

> EDA is an initial assessment of whether the available data measure the correct values, in sufficient quality and quantity, to answer a particular question.

This requires:

- A well defined question or line of investigation
- A record of data collection methods and the interpretation of each variable (data card)
- Documentation on the structure, precision, completeness and quantity of data available.

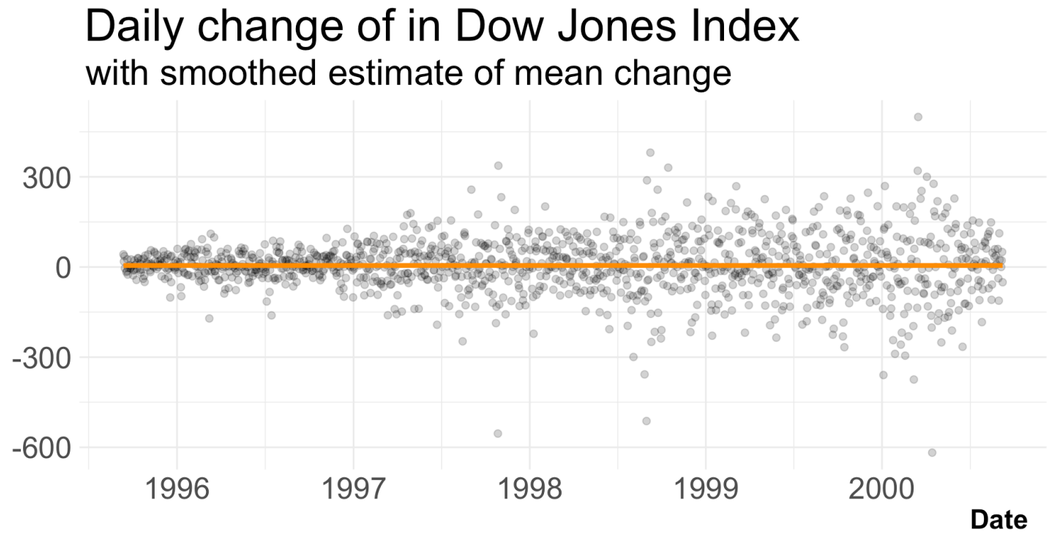
EDA as an investigation

- measurement noise or misclassification
- values and dependence between measured variables
- missing values and their structure
- signal strength and data size: simplest best and worst case

What is EDA not?

What is not Exploratory Data Analysis?

- EDA is not modelling.
- EDA is not IDA.
- EDA is not assumption free.
- EDA is not prescriptive.



after_june_98	mean	sd
FALSE	5.916798	65.19093
TRUE	3.972929	119.56067

Exploratory Data Analysis Issues

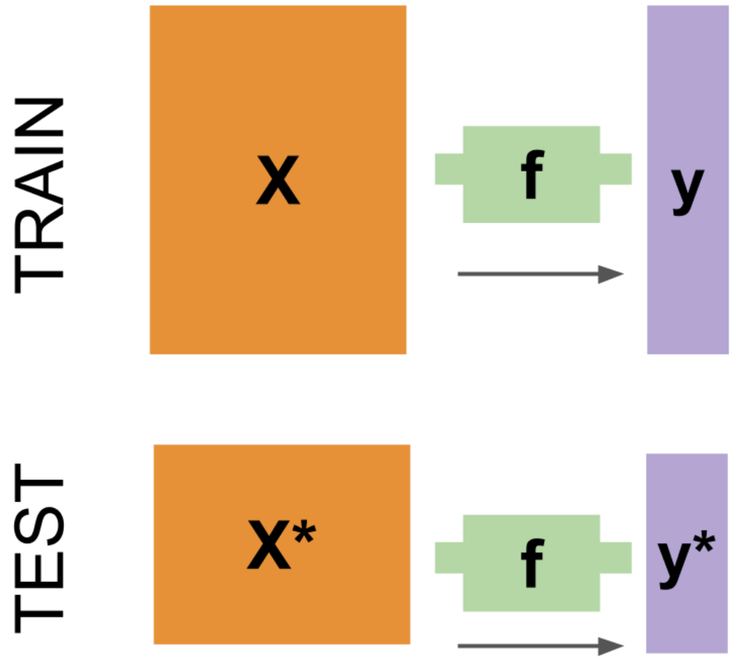
Too many choices: forking paths

- Data Science projects present a sequence of decisions.
- Too many options, difficult to decide **a priori**.
- EDA should help with this.

Example: selecting null model.

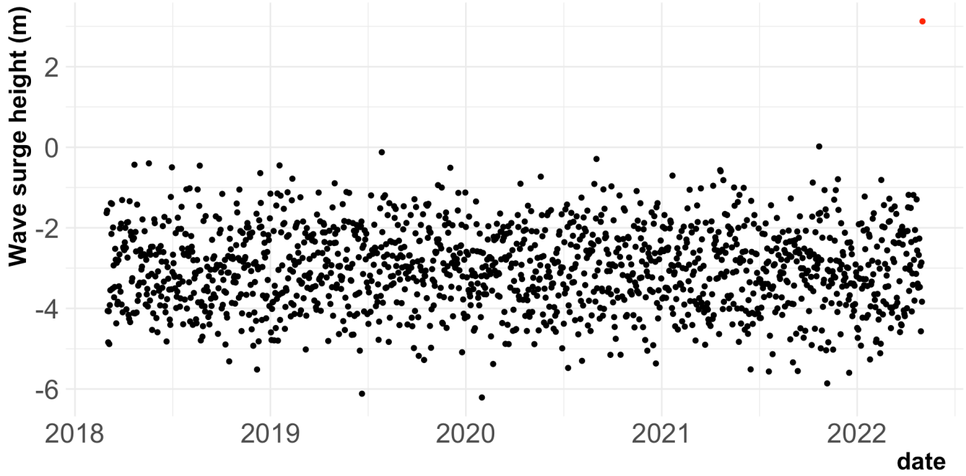
Data Leakage

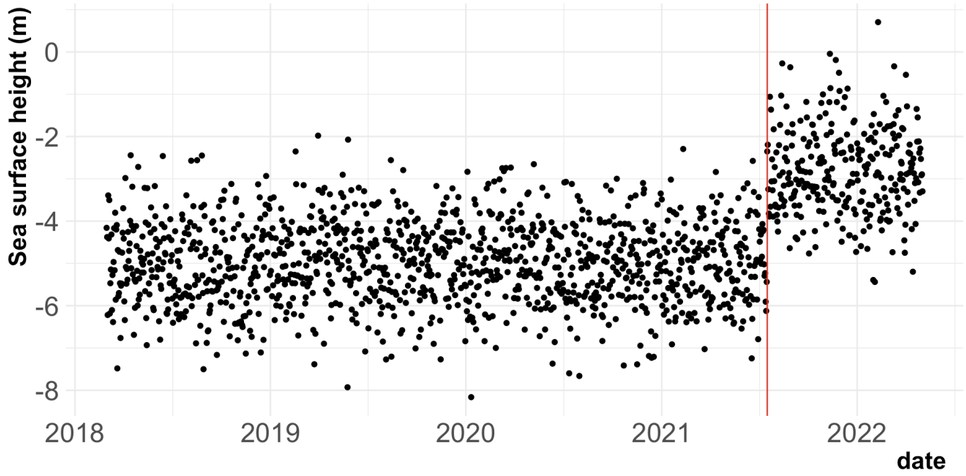
- Using information you wouldn't have access to fit a model or construct a prior.
- This "peeking" is often subtle or indirect making it hard to specify.
- Train / test split **or** using EDA to select question / model of interest.



Solutions

- Corrections to testing estimation procedures:
  - **Medical Stats** - multiple testing;
  - **Extremes** - flood defences;
  - **Changepoints** - time of dislocation.
- Avoided by preregistration.
- Humility and follow-up required in data science.





Learning More

Recap

- EDA is an important step in the life-cycle of a data science project.
- An EDA can guide our project but risks data leakage issues.

Learning more

- EDA not often available publicly or written about in detail.
- Learn from your own experience and explore lots of what other people do
- Some starting points:
  - EDA check list by Roger Peng
  - Exploratory Data Analysis for Complex Models by Andrew Gelman