Exploratory Data Analysis

Richard Vale, 23 May 2019

Exploratory Data Analysis (EDA)

- Exploratory (looking at the data) versus Confirmatory (modelling; hypothesis tests)
- Very important step
- Goals:

  cleaning; visualization; outlier detection; preparation of data for modelling; formulation of questions; missing values; etc.
- Influences predictive performance; benefits outweigh disadvantages

Procedure

- Use a REPL (read-evaluate-print-loop)
- Summary statistics and plots
- Start with univariate analysis (*anecdote)
- Then do bi/multivariate

- Do not be afraid to use models as part of the exploratory analysis! Even a boxplot is a model

Pandas

- In this tutorial, we will use Pandas to explore various different data types (numerical; text; categorical; dates)

- Data set of insurance complaints in Connecticut

References

- https://rpubs.com/jasdumas/eda-ct-insurance
- Tukey, Exploratory Data Analysis, 1977
- Kaggle writeups
- Singaporean rogue train incident; a great example of using EDA to solve a problem (with Python code):

  https://blog.data.gov.sg/how-we-caught-the-circle-line-rogue-train-with-data-79405c86ab6a

Note: in writeups, people may not describe **all** their EDA steps. I strongly recommend doing at least a plot or summary of every individual variable.