Exploratory Data Analysis

Damian Thomas

2017-02-03

Topics

1. Review
   - R Data Structures
   - Brackets
   - Subsetting
   - Iteration
   - Functions
2. Importing Data
3. Exploratory Data Analysis
   - What is it?
4. Techniques
   - Five number summary
   - Five number summary
   - Box plot

R Data Structures

Table: R Data Structures by Content Type and Number of Dimensions

	Homogeneous	Heterogeneous
1d	Atomic vector	List
2d	Matrix	Data Frame
*n*d	Array	

Anscombe's Quartet

```
> #
> setwd("~/projects/hu/Ecog314_Spring2017/lecture3/")
> anscombe <- read.csv("data/anscombe.csv")
> str(anscombe)

'data.frame':        11 obs. of  8 variables:
 $ x1: int  10 8 13 9 11 14 6 4 12 7 ...
 $ x2: int  10 8 13 9 11 14 6 4 12 7 ...
 $ x3: int  10 8 13 9 11 14 6 4 12 7 ...
 $ x4: int  8 8 8 8 8 8 8 19 8 8 ...
 $ y1: num  8.04 6.95 7.58 8.81 8.33 ...
 $ y2: num  9.14 8.14 8.74 8.77 9.26 8.1 6.13 3.1 9.13 7.26
 $ y3: num  7.46 6.77 12.74 7.11 7.81 ...
 $ y4: num  6.58 5.76 7.71 8.84 8.47 7.04 5.25 12.5 5.56 7.
```

What is Exploratory Data Analysis?

In statistics, exploratory data analysis (EDA) is an approach to analyzing data sets to summarize their main characteristics, often with visual methods.[1]

[1] Source: https://en.Wikipedia.org/wiki/Exploratory_data_analysis

EDA is:

- Data focused
- Informal. No model is specified
- Gain insight into the data generating process.
- Learn about the data, underlying structure
- Summarize the data without losing information.
- Gather key information required to build a model.
- Generate questions
- help decide what sort of model fits

EDA is *not:*

- Model focused
- Dependent on assumptions (randomness, normality, etc.)
- A rigorous formal approach
- Model Specification (regressions, ANOVA)
- Parameter estimation
- Hypothesis testing /statistical inference

Techniques

- Summary statistics
- Visualizations

Tukey's five number summary

- minimum: smallest value
- lower quartile: 25th percentile
- median: middle value
- upper quartile: 75th percentile
- maximum: largest value

Tukey's five number summary in R

Key summary statistics

- Extremes
- Location
- Spread

-
-
-