Basic Data Understanding

Lecture 2 (Data Science)
17 April 2020
http://shala2020.github.io/

Outline

- Goals of data science
- Data types
- Measures of central tendency
- Order statistics
- Types of plots
- Distributions

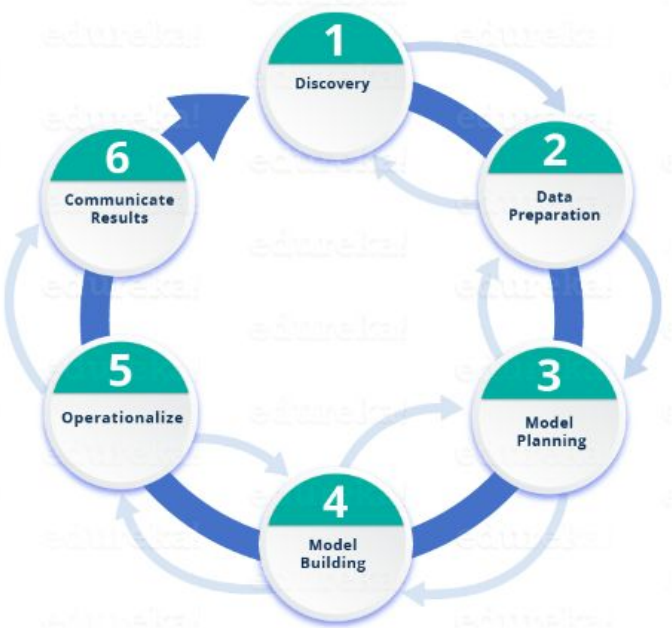
Goals of data science (DS)

- DS is used to make decisions and predictions
  - Predictive causal analytics,
  - Prescriptive analytics, and
  - Machine learning.
- Role of a data scientist is different from that of a data analyst

Lifecycle of Data Science

Reference: [Edureka blog]

- ➔ Discovery
- ➔ Data preparation
- ➔ Model planning
- ➔ Model building
- ➔ Operationalize
- ➔ Communicate results



Types of data

Reference: Statistics By Jim

- Quantitative data
  - Continuous - Histogram (single variable), Scatter plot (two variables), etc.
  - Discrete - bar chart
- Qualitative data

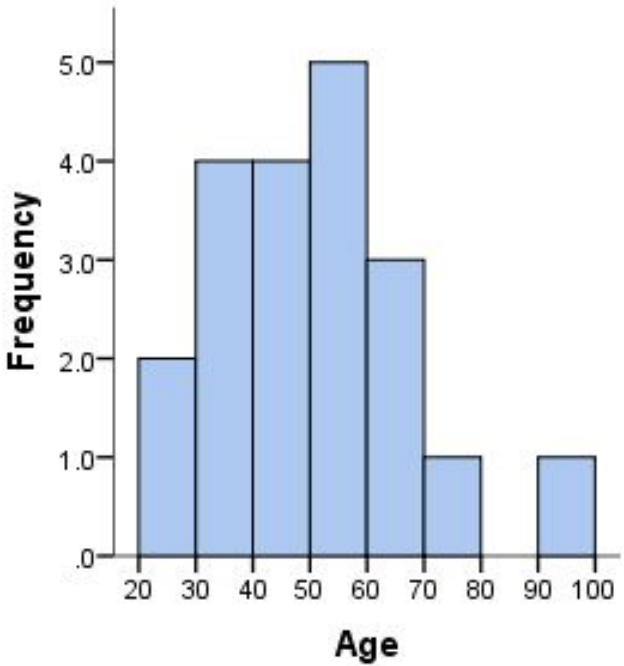
Types of data contd.

- Qualitative data
  - Categorical,
  - Binary, and
  - Ordinal
- Choosing statistical analyses based on data types

Basics of histogram

Reference: Laerd statistics

- Number of bins (intervals)
  - Neither too small nor too large
- No "gaps" between the bars



Measures of central tendency

Reference: Laerd statistics

- ➔ Single value that attempts to describe a set of data
  - ◆ Mean, median, and mode
- ➔ There is no best, but using only one is definitely worst!

When to use mean/median/mode

Reference: Laerd statistics

- ➔ Median is usually preferred over the mean (or mode), when our data is skewed.
- ➔ In case of normal distribution i.e. when the data is perfectly normal, the mean, median and mode are identical.

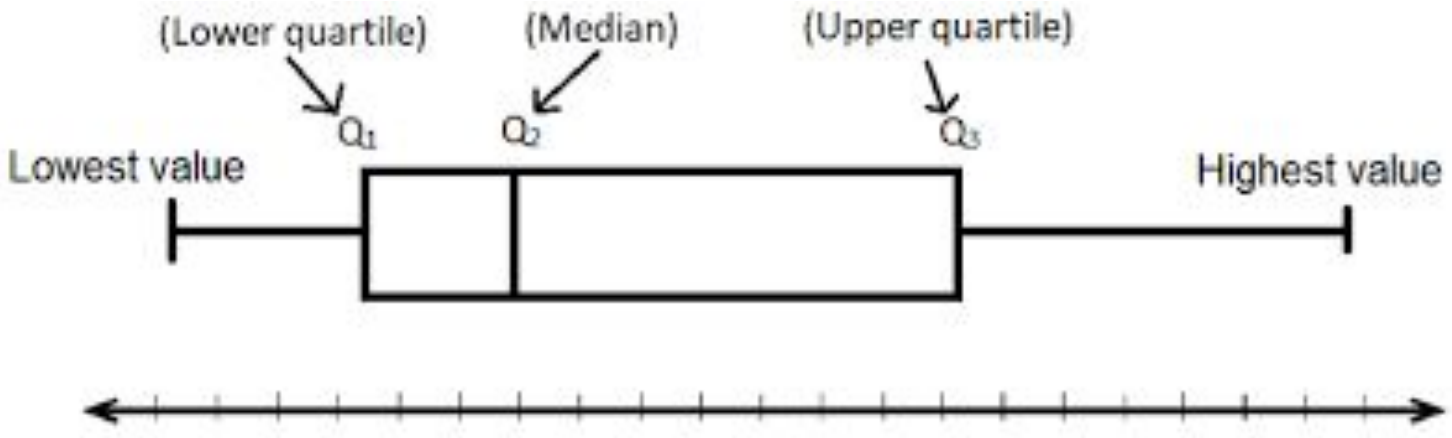
Order statistics

Reference: ICS UCI

- ➔ It refers to statistical methods that depend only on the ordering of the data and not on its numerical values
- ➔ The mean and mode of the data are not order statistic
- ➔ The most commonly used order statistic is the median

Basics of boxplot

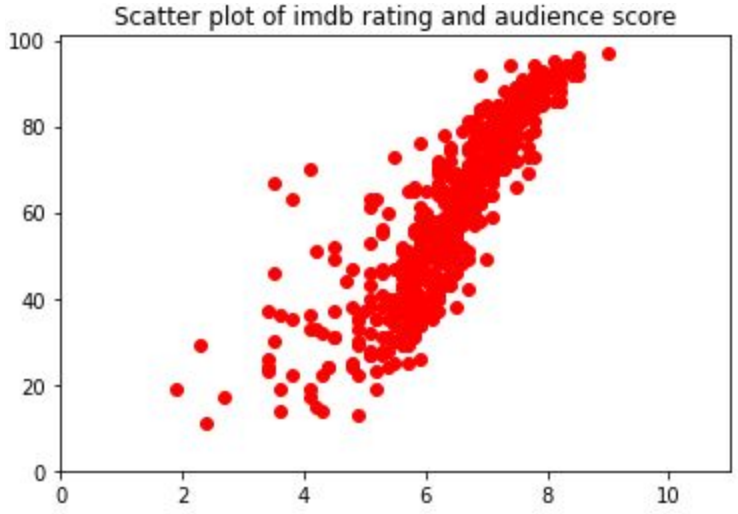
Reference: [Dimensionless]



Scatter Plot

Scatter plots are used to plot data points on a horizontal and a vertical axis to show how much one variable is affected by another, which is known as correlation between variables.

Note: - Auto correlation is different from Correlation. We will cover that in upcoming lectures.



Announcements

- Assignment - to be released just after the lecture (17 April).
  - Submission due on 19 April (9pm IST)
- Tutorial - 18 April (9pm IST)
- Next lecture (DS) - 20 April (9pm IST)