Feature Analysis and Visualization

MIE223

Winter 2025

1 Feature Analysis and Visualization

1.1 Exploratory Data Analysis (EDA)

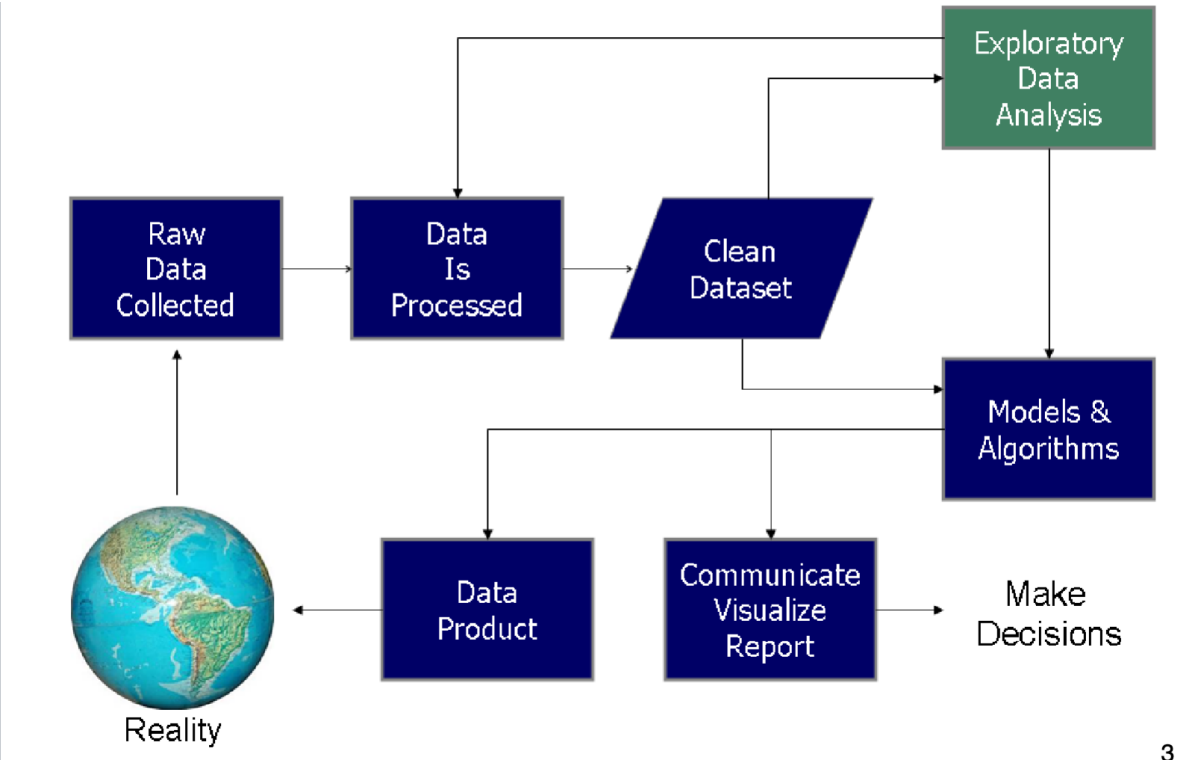
John Tukey (Bell Labs, 1970's) (Inventor of the box-plot)

- The first time corporations had a lot of digital data
- Data on semiconductor processes, networks

Observed that statistics was preoccupied with hypothesis testing

- But there were no established methodologies for **generating (data-driven) hypotheses**
- Espoused a visual methodology and a new language S (R became the free version)

1.2 Data Science Process



3

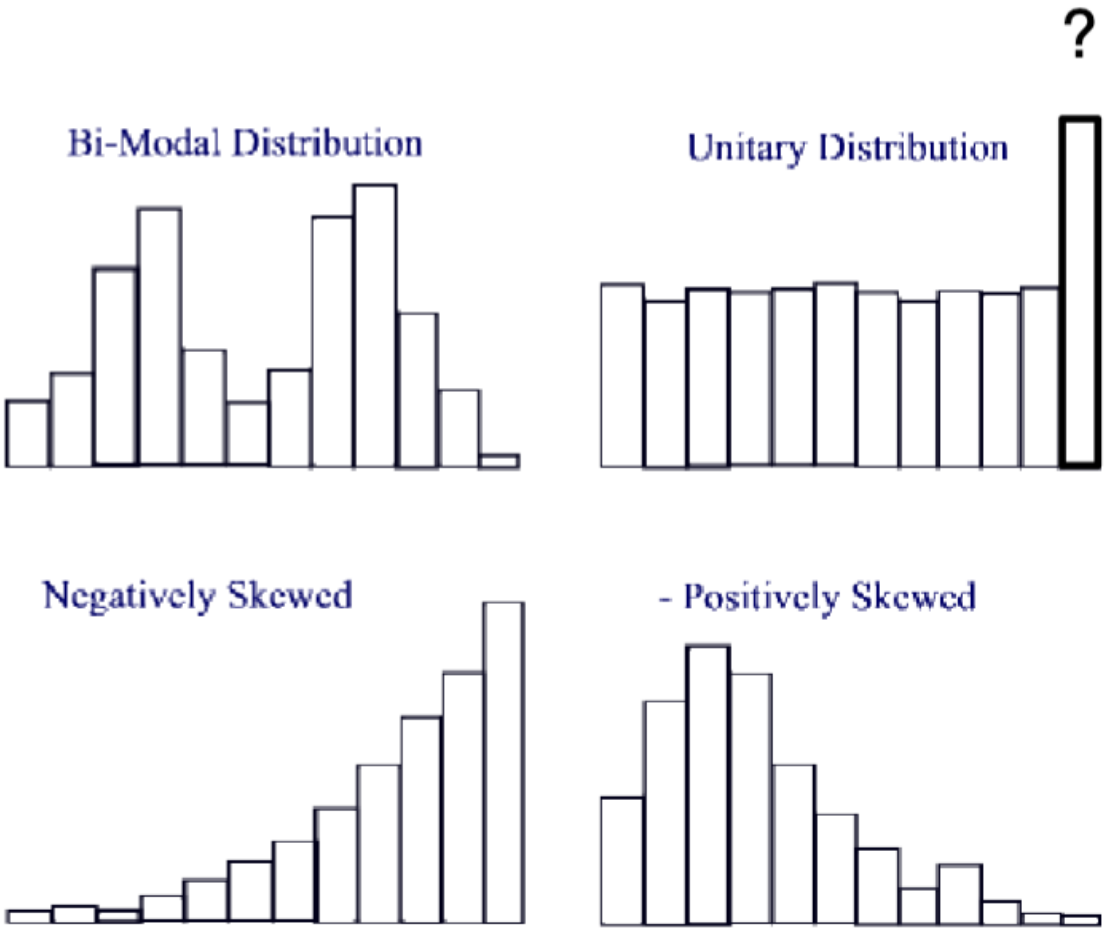
1.3 Data Cleaning: Brief Recap

80% of your Data Science time

Some pointers:

- Missing values – do not replace with 0 or -999

- Treat as missing
- Look for frequencies of values or outliers (-999)
  - **Histograms** invaluable
  - Examine **outliers**



2 Correlation

We'll be look at feature distributions independently (per column), but also relationships between features (columns). Correlation is one of the first tools you think of for looking at the relationship between two features (but has limitations)

2.1 Correlated variables

Two variables are correlated if changes in one variable, correspond to changes in the other
Correlation in general refers to the statistical association between two variables

- This statistical association allows us to make estimates for the one variable based on the value of the other

- While we typically think of linear relationships between variables, nonlinear might exist too

2.2 Linear correlation

- Two variables x and y are said to be linearly correlated if their relationship can be described by the following equation:

$$y = \alpha + \beta x, \beta \neq 0 \tag{1}$$

- This relationship will not be exact
- There will be an error associated with it, and hence, in reality:
- ϵ is the error term for the relationship

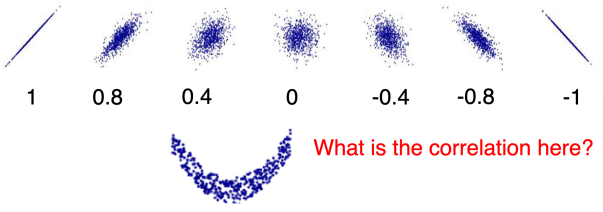
$$y = \alpha + \beta x + \epsilon, \beta \neq 0 \tag{2}$$

2.3 Pearson correlation coefficient

- In order to test the linear association between two variables x and y we can use the Pearson correlation coefficient r_{xy}

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

- Pearson correlation is in the range [-1,1]
  - 1: perfect/strong and positive linear correlation
  - -1: perfect/strong and negative linear correlation
  - 0: no linear correlation
- It is very crucial to understand that this correlation coefficient can only examine linear associations between two variables



Note 1.
- Pearson correlation coefficient is sensitive to outliers
- It is not robust to outliers
- It is not a good measure of correlation for non-linear relationships

3 Feature Analysis: Probability Recap

This is mathematical so we first we need to make sure everyone has the same understanding and intuition for notation

3.1 Random Variables

- Random variable (RV) denoted by uppercase letter (e.g. X)
- RVs take value assignments X = x where X is a
  - Discrete RV if x is in a countable set (binary: $x \in 0,1$; or dice)
  - Continuous RV if x is in an uncountable set (real: $x \in$ Reals)
- Write $x \in X$ for possible value assignments of X
- For all x, $P(X = x) \in [0,1]$
- P(X = x) is a proper distribution
  - Discrete RV:

    $$\sum_{x \in X} P(X = x) = 1$$

  -
  - Continuous RV:

    $$\int_{x \in X} P(X = x)\, dx = 1$$

  -
- Write P(x) for P(X=x), write P(X) for full distribution
- Representing probability distributions
  - Discrete (finite) RV: tabular
  - Continuous (infinite) RV: function

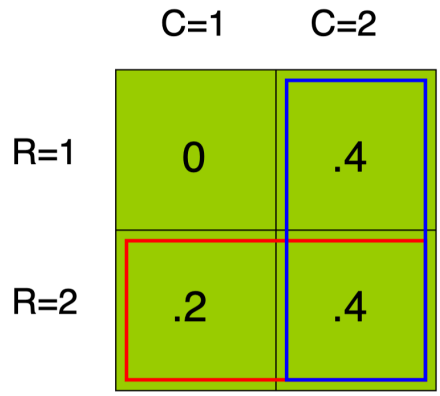
3.2 Joint Distributions on RVs

Aliens in your backyard

- Aliens in your backyard

P(R,C) =

R	C	Pr
1	1	0
1	2	.4
2	1	.2
2	2	.4



- P(R=2,C=2) = .4

Marginalize over C

- P(R=2) = $\sum_{c \in \{1,2\}}$ P(R=2,C=c) = .6
- P(R=1|C=2) = P(R=1,C=2)/P(C=2) = .5

Condition on C=2

Example from Andrew Moore @ CMU/Google

17