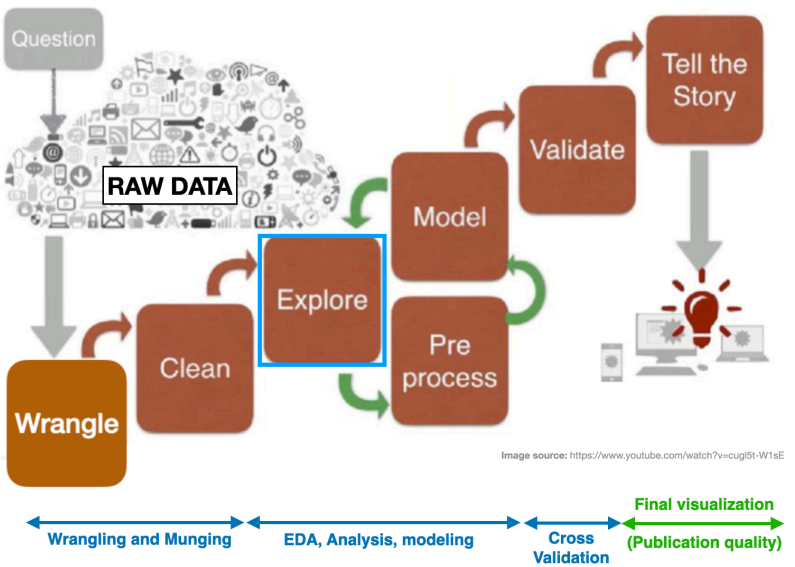
Exploratory data analysis

A brief introduction

Exploring data patterns, distributions, anomalies, and relationships. Informing insights and decisions through initial analysis



Motivation

- You have clean data … now what.

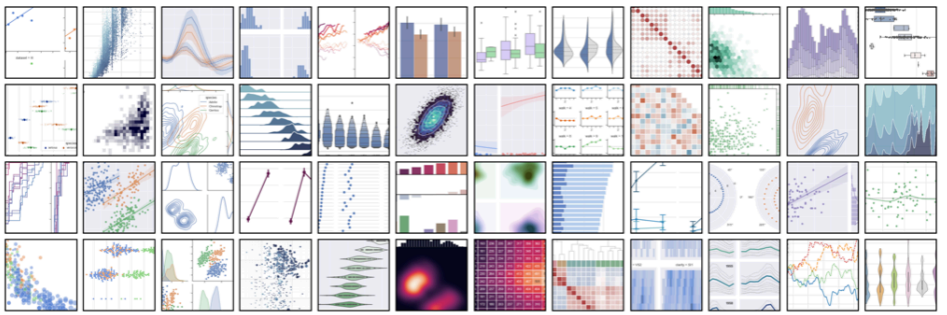
Clean data



ID	height (ft)	weight (lb)	age (years)	income ($)	adult/teen/child
1	h_1	w_1	a_1	i_1	C
2	h_2	w_2	a_2	i_2	A
3	h_3	w_3	a_3	i_3	C
⋮	⋮	⋮	⋮	⋮	⋮
N	h_N	w_N	a_N	i_N	T

EXPLORATORY DATA ANALYSIS (EDA)!!



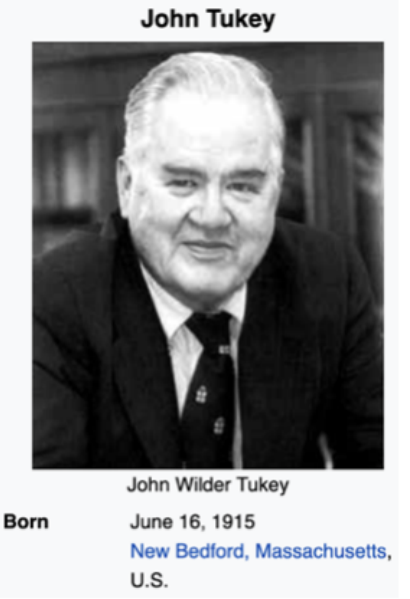


- EDA uncovers data trends, outliers, and relationships, guiding informed decisions
- Validates assumption which helps to ensure accurate modeling.
- Generally enhances data understanding and insights.”

Overview

> Exploratory data analysis can never be the whole story, but nothing else can serve as the foundation stone – as the first step
>
> ~ John Tukey, 1977

> An approach/philosophy for data analysis that employs a variety of techniques (mostly graphical) to maximize insight into a data set, uncover underlying structure, extract important variables, detect outliers and anomalies, test underlying assumptions, develop parsimonious models and determine optimal factor settings.
>
> ~ National Institute of Standards and Technology (NIST)



https://en.wikipedia.org/wiki/John_Tukey

EDA goals

`IBM`

- Get a look at data before making any assumptions.
- Screen data and Identify obvious errors
- Better understand patterns within the data
- Detect outliers or anomalous events
- Ask questions and check/validate your assumptions
- Find interesting relations among the variables.

Source: https://www.ibm.com/cloud/learn/exploratory-data-analysis

`NIST`

The primary goal of EDA is to **maximize the analyst's insight** into a data set and into the **underlying structure of a data set**, while providing all of the **specific items** that an analyst would want to **extract from a data set,**

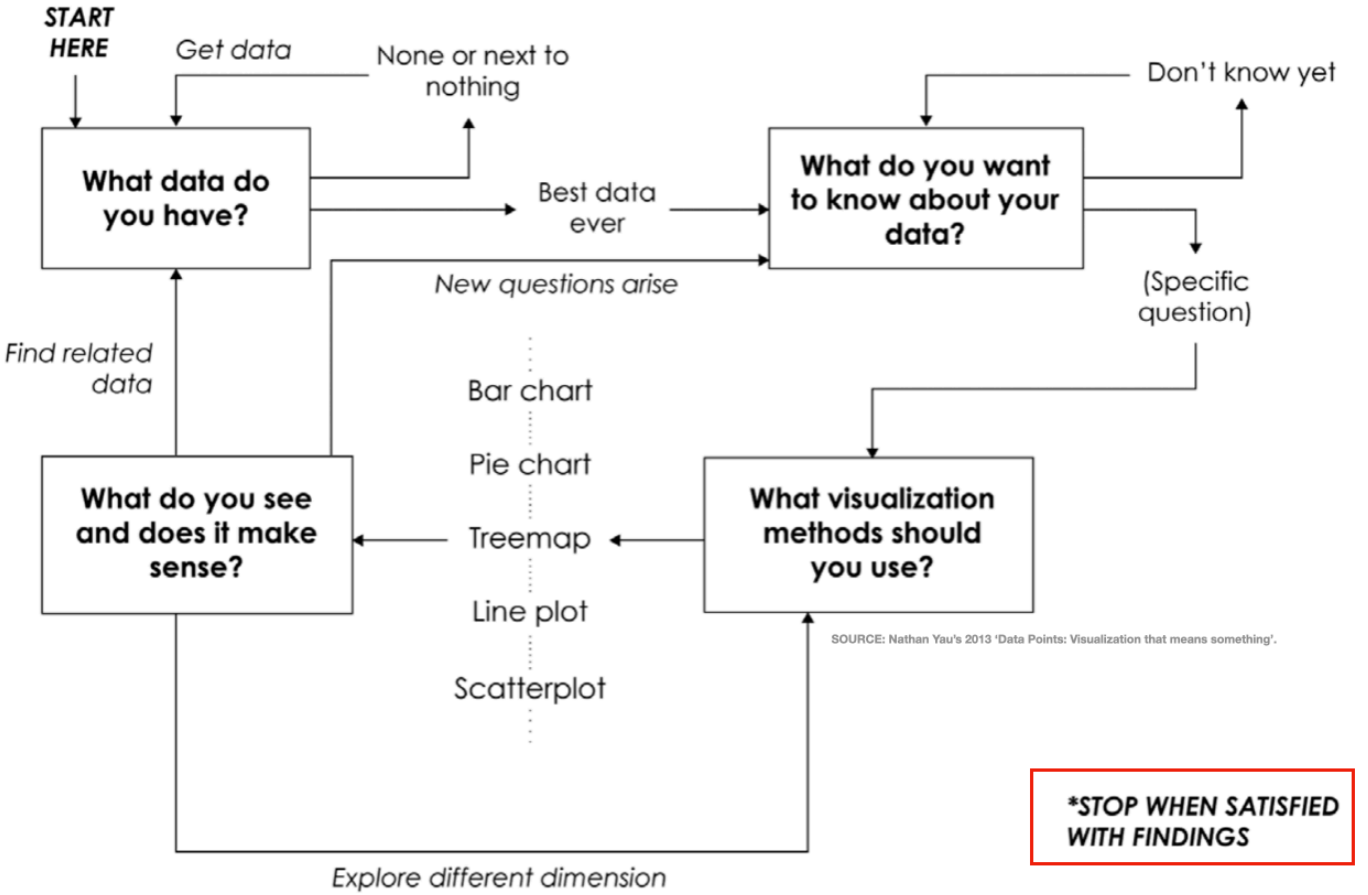
1. a good-fitting, parsimonious model
2. a list of outliers
3. a sense of robustness of conclusions
4. estimates for parameters
5. uncertainties for those estimates
6. a ranked list of important factors
7. conclusions as to whether individual factors are statistically significant
8. optimal settings

Source: https://www.itl.nist.gov/div898/handbook/eda/section1/eda14.htm

NOTE: `Parsimonious` (frugal) models are **simple models with great explanatory predictive power**. They explain data with a minimum number of parameters, or predictor variables.

EDA is an iterative process

- EDA and data cleaning is a time-consuming `non-linear` iterative process!
  - Sometimes you need to loop back, gather more data, re-clean, or start over.
- Often, the majority of your time as a data scientist is dedicated to this



EDA Tools

Graphical techniques:

- Bivariate visualizations
- Box plot
- Histogram
- Pareto chart
- Scatter plot
- Stem-and-leaf plot
- Parallel coordinates plots
- Targeted projection pursuit
- Interactive versions of these plots

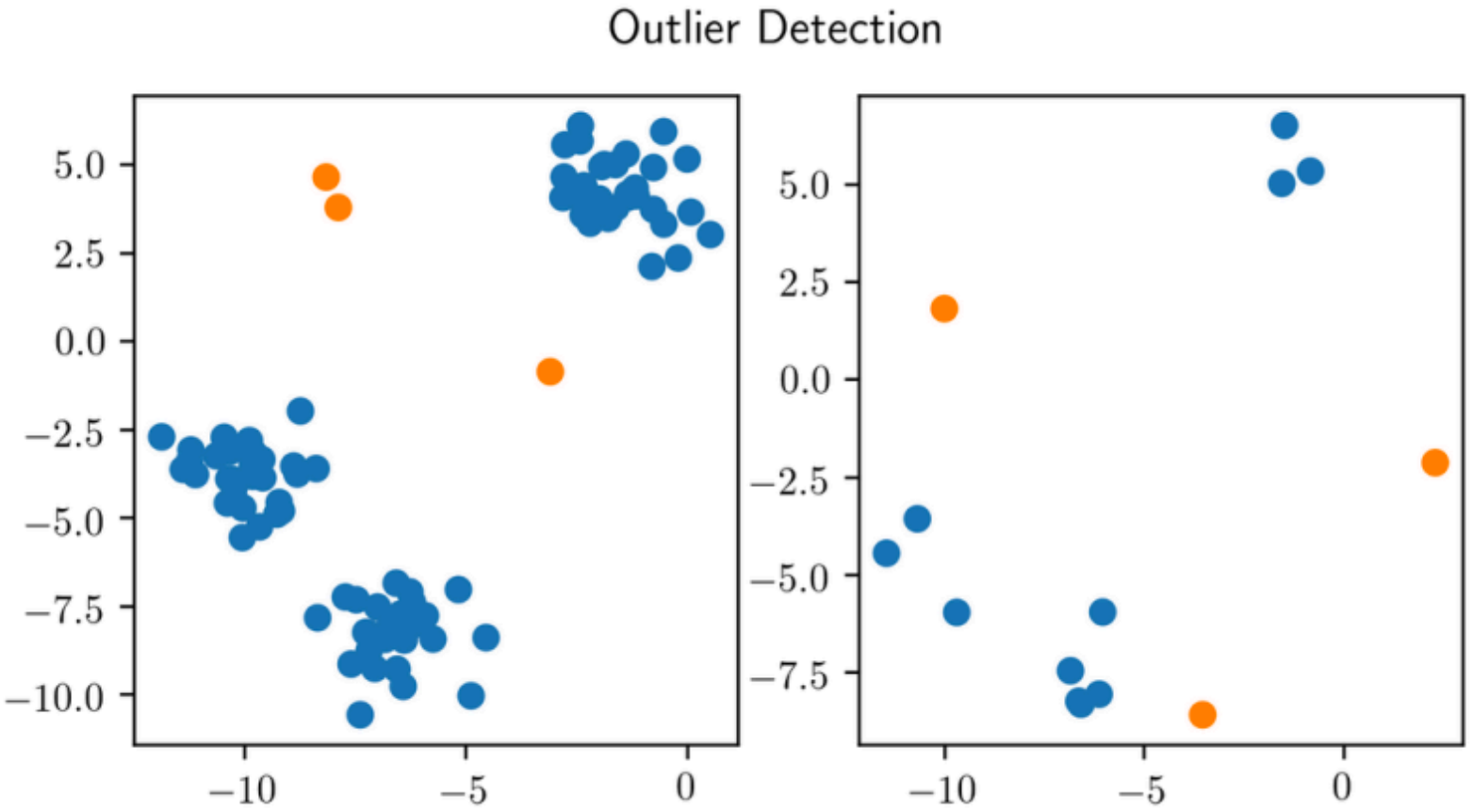
Quantitative techniques

- Clustering and dimension reduction techniques
  - e.g. K-means Clustering and PCA
- Summary statistics
- Multilinear PCA
- Nonlinear dimensionality reduction (NLDR)
- Iconography of correlations

Source: https://en.wikipedia.org/wiki/Exploratory_data_analysis

Outlier & anomaly detection

- Anomaly detection identifies deviations from expected patterns in data.
- It is Essential for spotting outliers, fraud, errors, or unusual behavior.



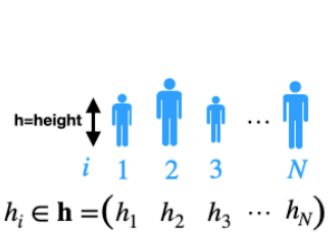
Univariate methods

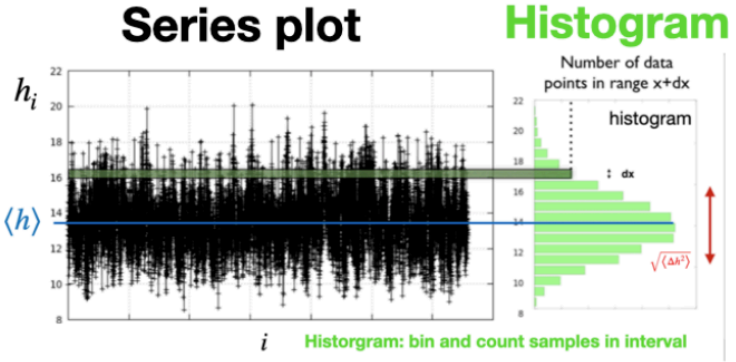
- `Univariate non-graphical.` This is the simplest form of data analysis, where the data being analyzed consists of just one variable. Since it's a single variable, it doesn't deal with relationships. The main purpose of univariate analysis is to describe the data and find patterns that exist within it.
- `Univariate graphical`. Non-graphical methods don't provide a full picture of the data. Graphical methods are therefore required. Common types of univariate graphics include:
  - Box plots, which graphically depict the five-number summary of minimum, first quartile, median, third quartile, and maximum.
  - Histograms, a bar plot in which each bar represents the frequency (count) or proportion (count/total count) of cases for a range of values.
  - Stem-and-leaf plots, which show all data values and the shape of the distribution.

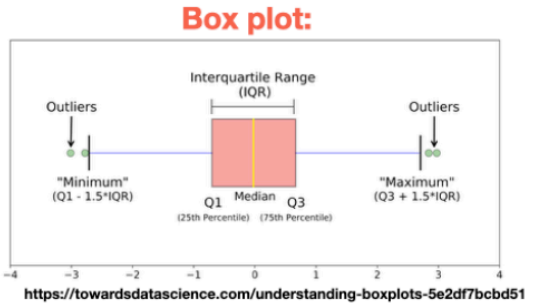
Source: https://www.ibm.com/cloud/learn/exploratory-data-analysis

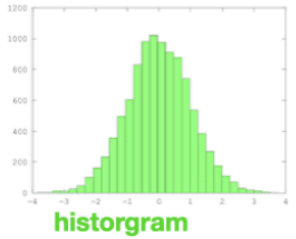
Uni-variate visual EDA

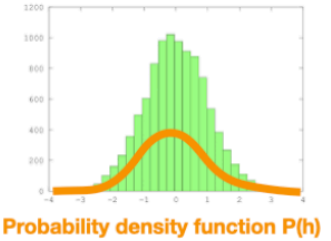
- Single variable EDA largely boils down to visualizing the distribution points for a single dimensions of your data
- There are many ways to visualize uni-variate distributions, we will discuss them in more detail in DSAN-5100



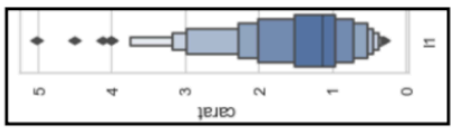


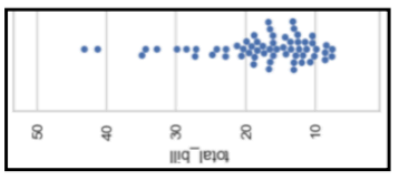


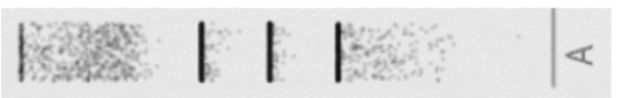


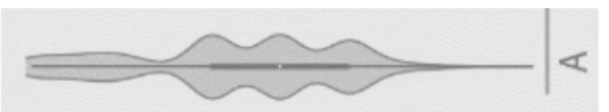














Five number summaries

- Statistical summary comprising minimum, first quartile, median, third quartile, and maximum values, providing insights into data distribution and spread.
  - Five-Number Summary: $\min, Q_1, Q_2, Q_3, \text{Max}$
  - Interquartile range (IQR): $IQR = Q_3 - Q_1$
  - Limits of the dataset:
    - Lower limit = $Q_1 - 1.5 \times 1QR$
    - Upper limit = $Q_3 + 1.5$ x IQR
  - `Outliers`: The objects below the lower limit and above the upper limit are potential outliers.
  - `Percentile`: divides the data into hundredths (100 equal parts) $P_1, P_2, \ldots, P_{99}$

Multi-variable EDA

- Multivariate data arises from more than one statistical variable.
- `Multivariate nongraphical`:
  - Generally show the relationship between two or more variables of the data through cross-correlation statistics.
- `Multivariate graphical`: Multivariate data uses graphics to display relationships between two or more sets of data.
  - The most used graphic is a grouped bar plot or bar chart with each group representing one level of one of the variables and each bar within a group representing the levels of the other variable.
  - Multivariate chart, which is a graphical representation of the relationships between factors and a response.
  - Heat map, which is a graphical representation of data where values are depicted by color.
  - Scatter plot, which is used to plot data points on a horizontal and a vertical axis to show how much one variable is affected by another.
  - Bubble chart, which is a data visualization that displays multiple circles (bubbles) in a two-dimensional plot.

Source: https://www.ibm.com/cloud/learn/exploratory-data-analysis

Multi-variate visual EDA

- Multi-variable EDA largely boils down to the following, we will discuss this in more detail in DSAN-5100
  - 1. Visualizing multi-variable distribution points for multiple dimensions of your data
  - 2. Visualizing data correlations and relationships in your data (pair-plots, correlation heatmaps, etc)
  - 3. Quantifying data correlations (Pearson and Spearman correlation coefficients)

