Tecnológico de Monterrey

Exploratory Data Analysis and Visualization

CS5056 Data Analytics

Francisco J. Cantú, Héctor Ceballos
Tecnológico de Monterrey

February 17, 2021

Februrary-June, 2021

CS5056 Data Analytics

February 17, 2021

Class 2/16

Agenda

SUMMARY OF CLASS 1

DATA PREPARATION

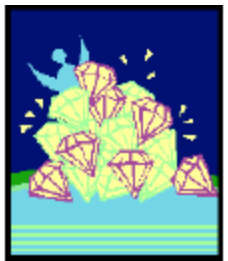
LABELING

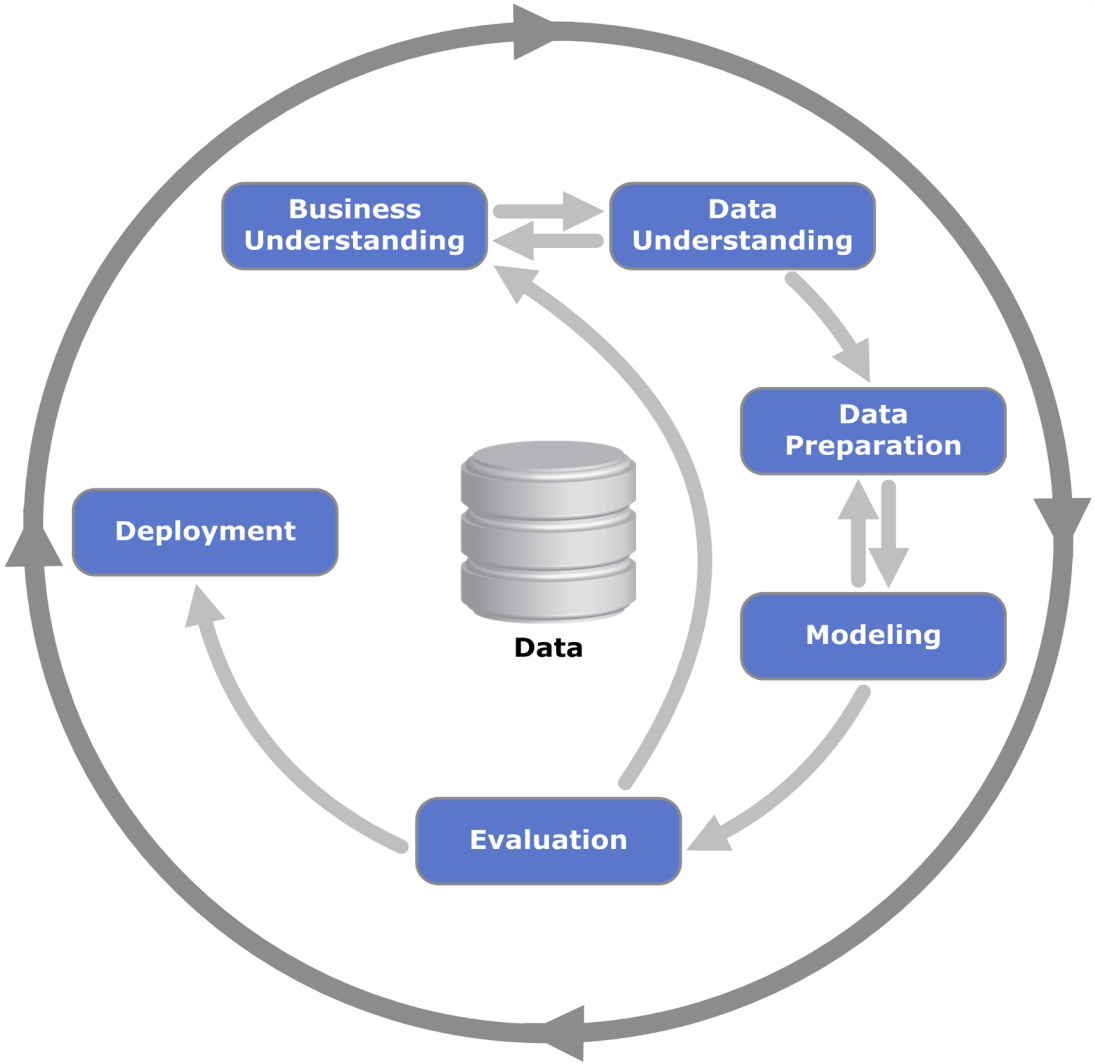
VISUALIZATION

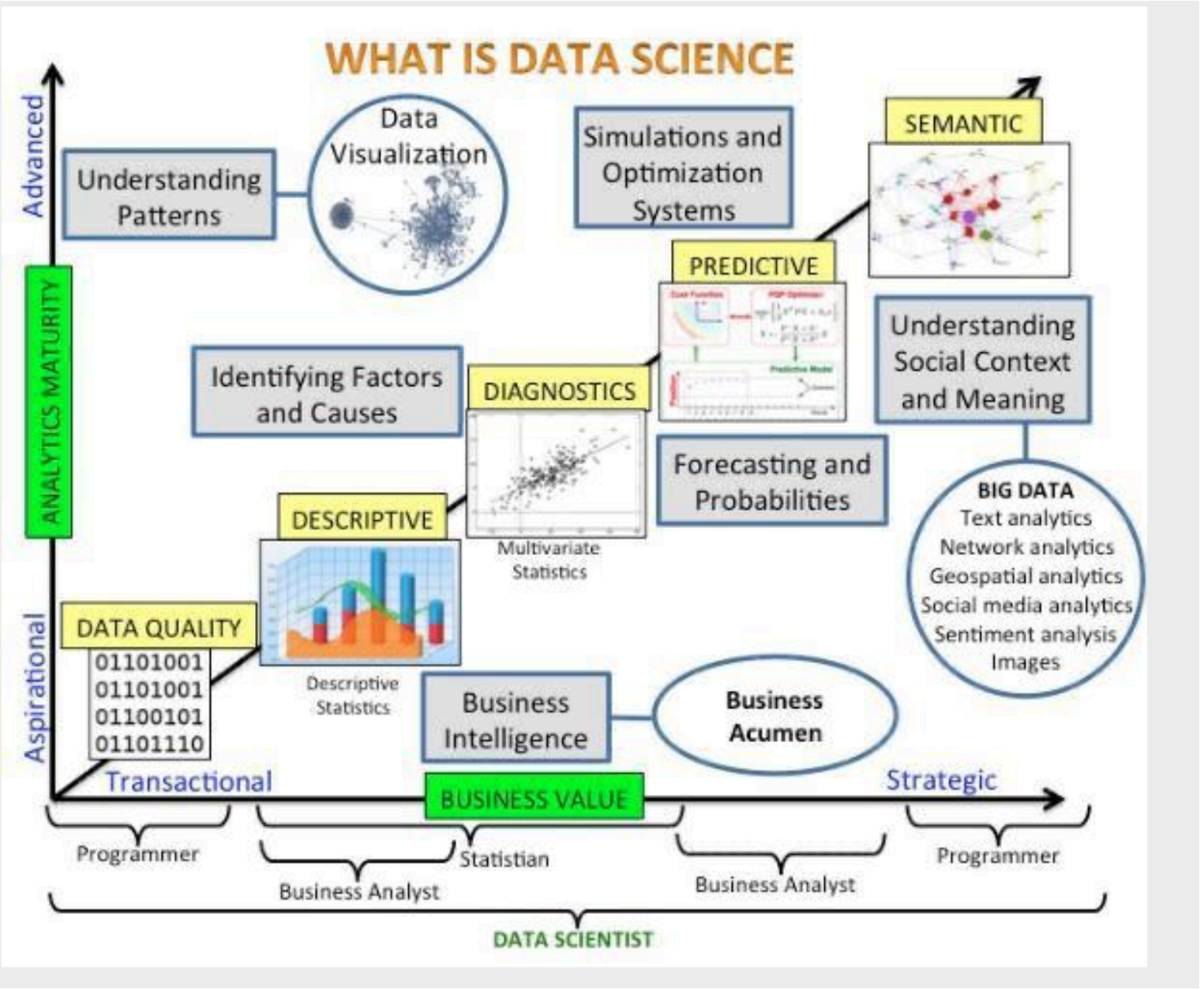
HANDS-ON EXERCISES WITH PYTHON

HANDS-ON EXERCISES WITH R

Data Mining Cycle: CRISP-DM

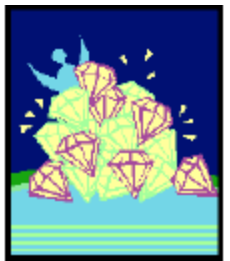




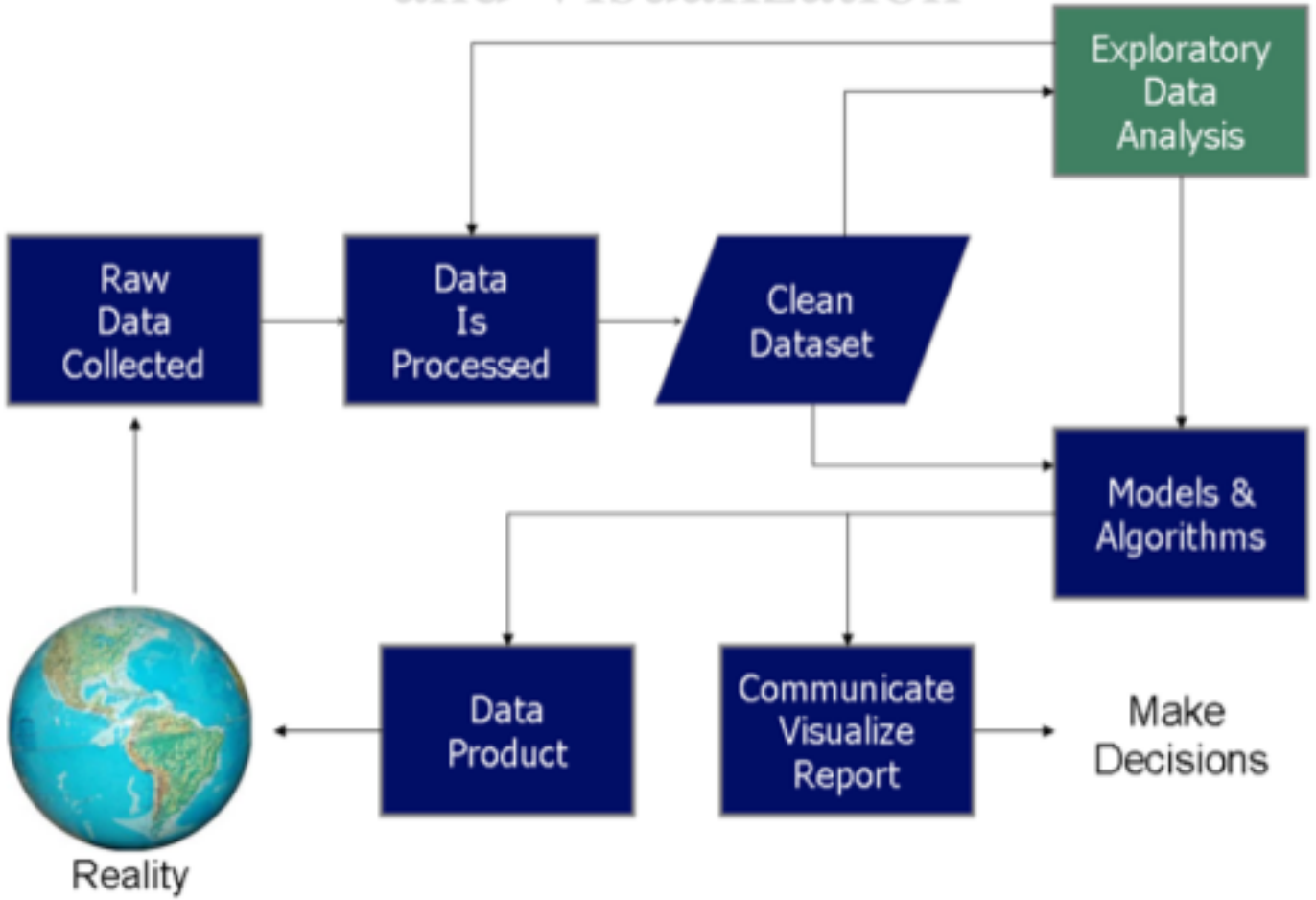

WHAT IS DATA SCIENCE
Advanced
ANALYTICS MATURITY
Aspirational
Understanding Patterns
Data Visualization
Simulations and Optimization Systems
SEMANTIC
PREDICTIVE
Understanding Social Context and Meaning
Identifying Factors and Causes
DIAGNOSTICS
Multivariate Statistics
Forecasting and Probabilities
DESCRIPTIVE
Descriptive Statistics
DATA QUALITY
01101001
01101001
01100101
01101110
BIG DATA
Text analytics
Network analytics
Geospatial analytics
Social media analytics
Sentiment analysis
Images
Business Intelligence
Business Acumen
Transactional
BUSINESS VALUE
Strategic
Programmer
Business Analyst
Statistian
Business Analyst
Programmer
DATA SCIENTIST

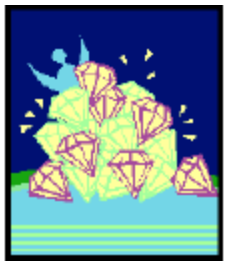
Hands-on Exercises Class 1

- Fortune500
- Marketing
- Churning
  - CRISP-DM
  - EDA



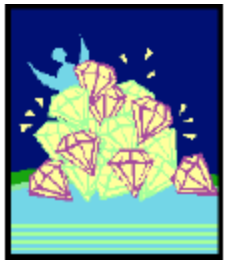
Exploratory Data Analysis and Visualization





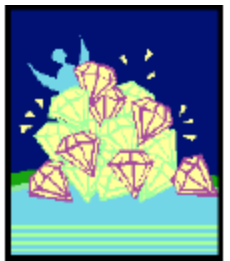
Your Dataset

- What is the discipline?
- What is it about?
- What problem would you like to address?
- Does it contain text, images, speech, combinations?
- Relational or Unstructured?
- How many rows and columns?
- Are there missing data?
- Are data clean?
- Is it in a cloud site?



Data Preparation

- Labeling Datasets
- Supervised learning



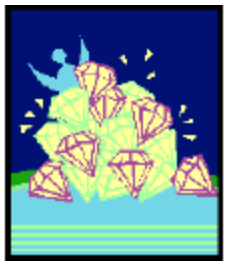
Completely Automated Public Turing Test to tell Computers and Humans Apart

Label on the correct Image(s)

Data Preparation

What is Data Labeling?

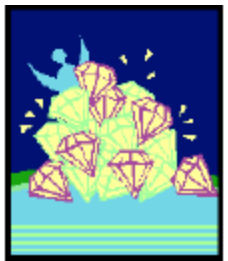
- https://aws.amazon.com/sagemaker/groundtruth/what-is-data-labeling/



Data Preparation

Types and Importance of Data Labeling?

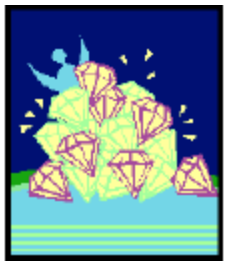
- https://whatis.techtarget.com/definition/data-labeling



Data Preparation

ImageNet

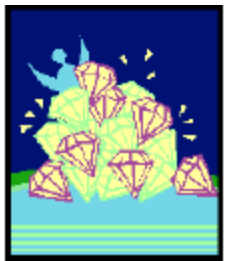
- https://en.wikipedia.org/wiki/ImageNet



Exploratory Data Analysis

- Exploratory Data Analysis (EDA) is an approach to analyzing data sets to summarize their main characteristics, often with visual methods

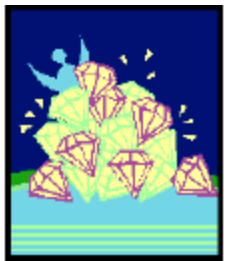
- A statistical model can be used or not, but primarily EDA is for seeing what the data can tell us beyond the formal modeling or hypothesis testing task.



Exploratory Data Analysis

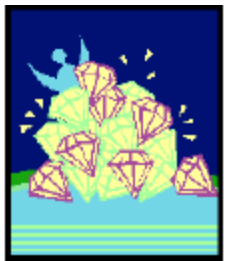
- EDA was promoted by John Tukey to encourage statisticians to explore the data and formulate hypotheses that could lead to new data collection and experiments

- EDA is different from initial data analysis (IDA) which focuses more narrowly on checking assumptions required for model fitting and hypothesis testing, and handling missing values and making transformations of variables as needed



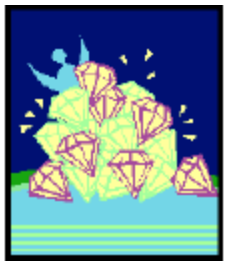
Approaches to EDA

- Descriptive Statistics
- Inferential Statistics
- Machine Learning
- Pattern Recognition



Approaches to EDA

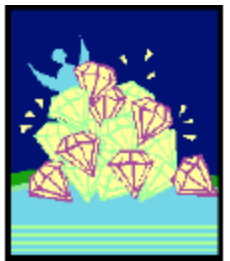
- Descriptive Statistics
  - Min, max, quantiles, median, mode, correlation
  - Visual methods: histograms, boxplots, scatterplots, violin plots, word clods, etc.
- Inferential Statistics
  - Probability axioms, mass and density distribution, mean, variance, covariance, expected value, hypothesis testing, the Z, t, F, Xi^2 statistics, analysis of variance, regression, panel data, time series analysis, etc.
- Machine Learning
  - Decision trees, clustering, heuristic search, greedy search, K-NN, neural nets, genetic algorithms, Bayesian networks, Boosting, etc.
- Pattern Recognition
  - Classification, discriminant analysis, K-Means, SVM, Markov chains, PCA, Filtering, Hidden Markov Models, etc.



Descriptive Statistics

- **Descriptive Statistics** is the branch of Statistics for analyzing those statistics

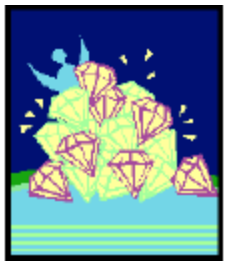
- A **descriptive statistic** is a summary statistic (a number) that quantitatively describes or summarizes features of a dataset



Descriptive Statistics

- It is distinguished from Inferential Statistics by its aim to summarize a sample, rather than use the data to learn about the population that the sample of data is thought to represent.

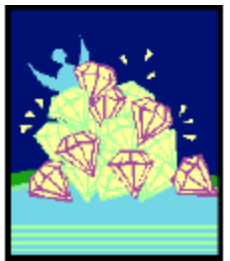
- Unlike inferential statistics, it is not developed on the basis of probability theory and are frequently nonparametric statistics. Thus, it is a type of Nonparametric Statistics



Nonparametric Statistics

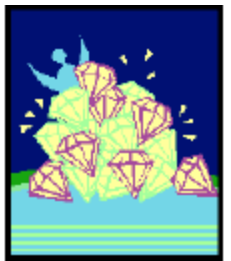
- It is the branch of Statistics that is not based only on parametrized families of probability distributions (common examples of parameters are the mean and variance).

- It is characterized on either being distribution-free or having a specified distribution but with the distribution's parameters unspecified. Nonparametric Statistics includes Descriptive Statistics



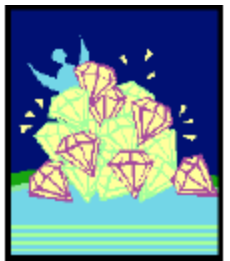
Inferential Statistics

- A set of values $x_1, x_2, \ldots, x_n$ (a dataset)
- Underlying distribution is unknown
- True parameters of the underlying distribution are unknown: mean, standard deviation, variance
- But, although unknown, true parameters can be estimated from the set of values
- For instance, the mean can be estimated by the average of the set of values
- Same for the standard deviation, the variance and other parameters



Inferential Statistics

- Random experiment
- Sample space
- Random Variable X
- Distribution function f(X)
- Probability values $p_1, p_2, \ldots, p_n$
- Probability **mass** function for discrete variables: p(X=x)
- Probability **density** function for continuous variables: p(X=x)
- Expected values E(X)
- Variance V(X)
- Covariance Cov(X,Y)



Parametric Statistics

- Arithmetic Mean

$$\bar{x} = \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right) = \frac{x_1 + x_2 + \cdots + x_n}{n}$$

For example, the arithmetic mean of five values: 4, 36, 45, 50, 75 is:

$$\frac{4 + 36 + 45 + 50 + 75}{5} = \frac{210}{5} = 42.$$

- Geometric Mean

$$\bar{x} = \left(\prod_{i=1}^{n} x_i\right)^{\frac{1}{n}} = (x_1 x_2 \cdots x_n)^{\frac{1}{n}}$$

For example, the geometric mean of five values: 4, 36, 45, 50, 75 is:

$$(4 \times 36 \times 45 \times 50 \times 75)^{\frac{1}{5}} = \sqrt[5]{24\,300\,000} = 30.$$

Harmonic mean (HM) [edit]

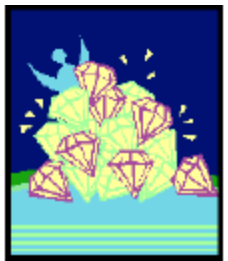
- Harmonic Mean

$$\bar{x} = n\left(\sum_{i=1}^{n} \frac{1}{x_i}\right)^{-1}$$

For example, the harmonic mean of the five values: 4, 36, 45, 50, 75 is

$$\frac{5}{\frac{1}{4} + \frac{1}{36} + \frac{1}{45} + \frac{1}{50} + \frac{1}{75}} = \frac{5}{\frac{1}{3}} = 15.$$

$$\mathrm{AM} \geq \mathrm{GM} \geq \mathrm{HM}$$



Parametric Statistics

- Weighted Mean:

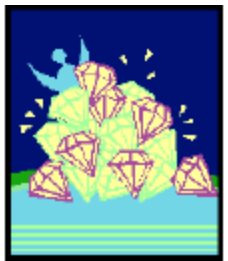
$$\bar{x} = \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i}.$$

- Mean of a continuous
- function

$$y_{\text{ave}}(a, b) = \frac{1}{b-a} \int_a^b f(x)\, dx$$

- Median: Value at the middle or central point

- Mode: Most repeated value



Visualization

Chart Suggestions—A Thought-Starter

What would you like to show?

Comparison

- Among Items
  - Two Variables per Item
    - Variable Width Column Chart
  - One Variable per Item
    - Many Categories
      - Table or Table with Embedded Charts
    - Few Categories
      - Many Items
        - Bar Chart
      - Few Items
        - Column Chart
- Over Time
  - Many Periods
    - Cyclical Data
      - Circular Area Chart
    - Non-Cyclical Data
      - Line Chart
  - Few Periods
    - Single or Few Categories
      - Column Chart
    - Many Categories
      - Line Chart

Relationship

- Two Variables
  - Scatter Chart
- Three Variables
  - Bubble Chart

Distribution

- Single Variable
  - Few Data Points
    - Column Histogram
  - Many Data Points
    - Line Histogram
- Two Variables
  - Scatter Chart
- Three Variables
  - 3D Area Chart

Composition

- Changing Over Time
  - Few Periods
    - Only Relative Differences Matter
      - Stacked 100% Column Chart
    - Relative and Absolute Differences Matter
      - Stacked Column Chart
  - Many Periods
    - Only Relative Differences Matter
      - Stacked 100% Area Chart
    - Relative and Absolute Differences Matter
      - Stacked Area Chart
- Static
  - Simple Share of Total
    - Pie Chart
  - Accumulation or Subtraction to Total
    - Waterfall Chart
  - Components of Components
    - Stacked 100% Column Chart with Subcomponents

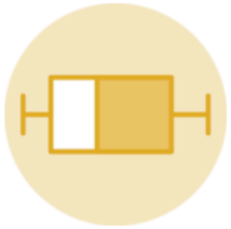
Visualization

Distribution


Violin


Density


Histogram


Boxplot


Ridgeline

Correlation


Scatter


Heatmap


Correlogram


Bubble


Connected scatter


Density 2d

Visualization

Ranking


Barplot


Spider / Radar

Wordcloud


Parallel


Lollipop


Circular Barplot

Part of a whole


Grouped and Stacked barplot


Treemap


Doughnut


Pie chart


Dendrogram


Circular packing

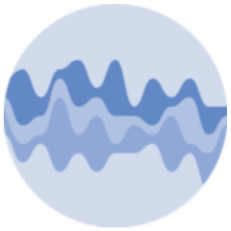
Visualization

Evolution

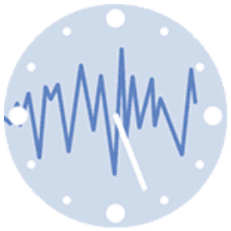
Line plot

Area

Stacked area



Streamchart



Time Series

Map

Map

Choropleth

Hexbin map

Cartogram

Connection

Bubble map

Visualization

Flow


Chord diagram


Network


Sankey


Arc diagram


Edge bundling

General knowledge

Ggplot2


Animation

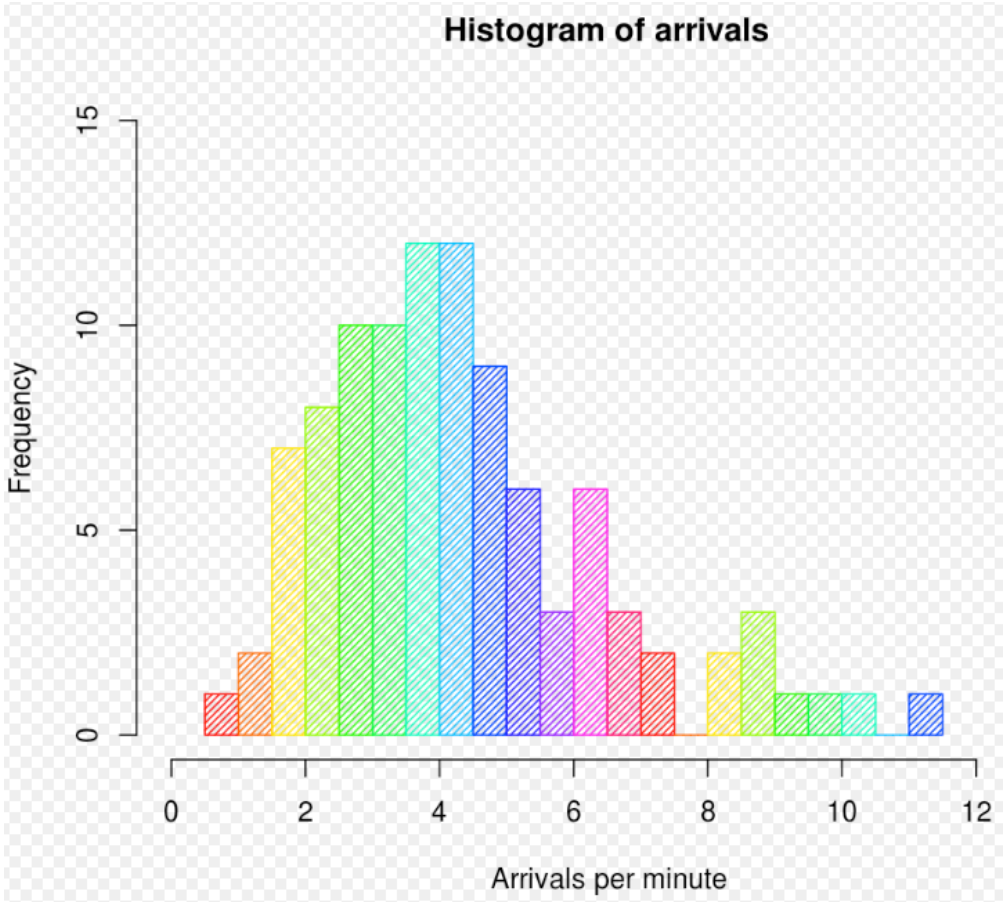
Interactivity


3D


Caveats

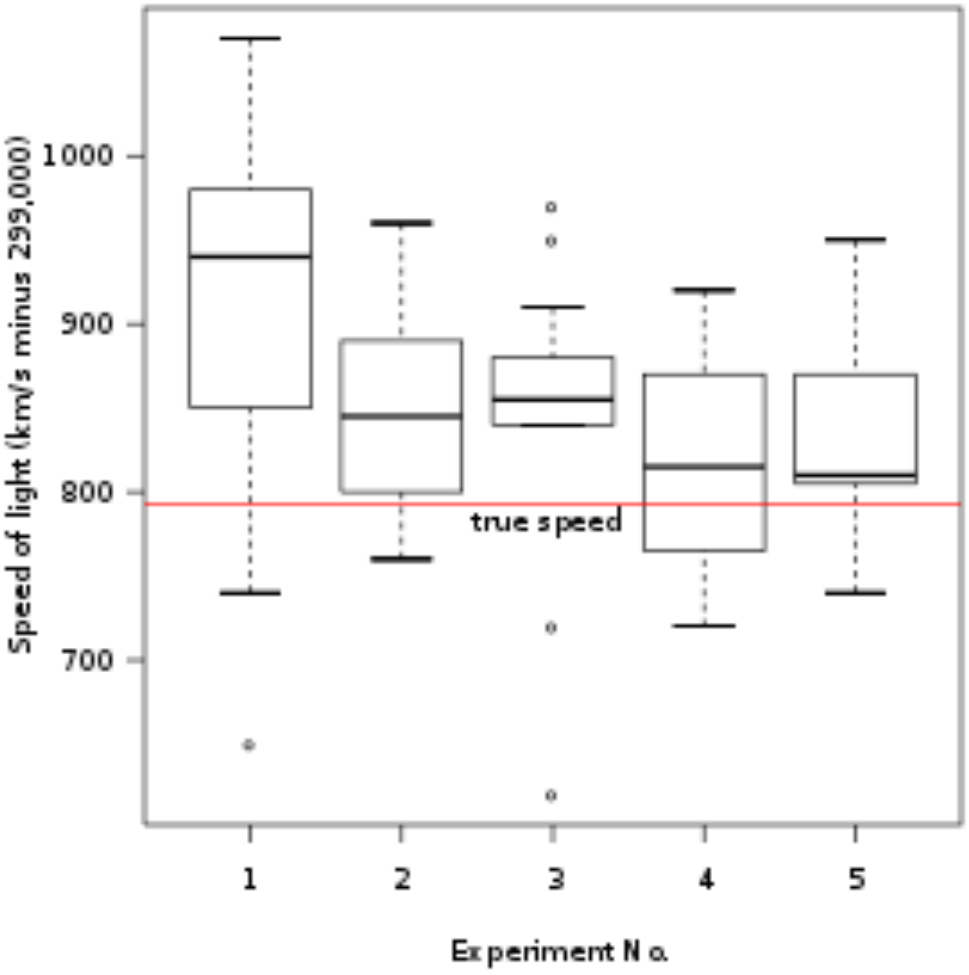

Data art

Visualization



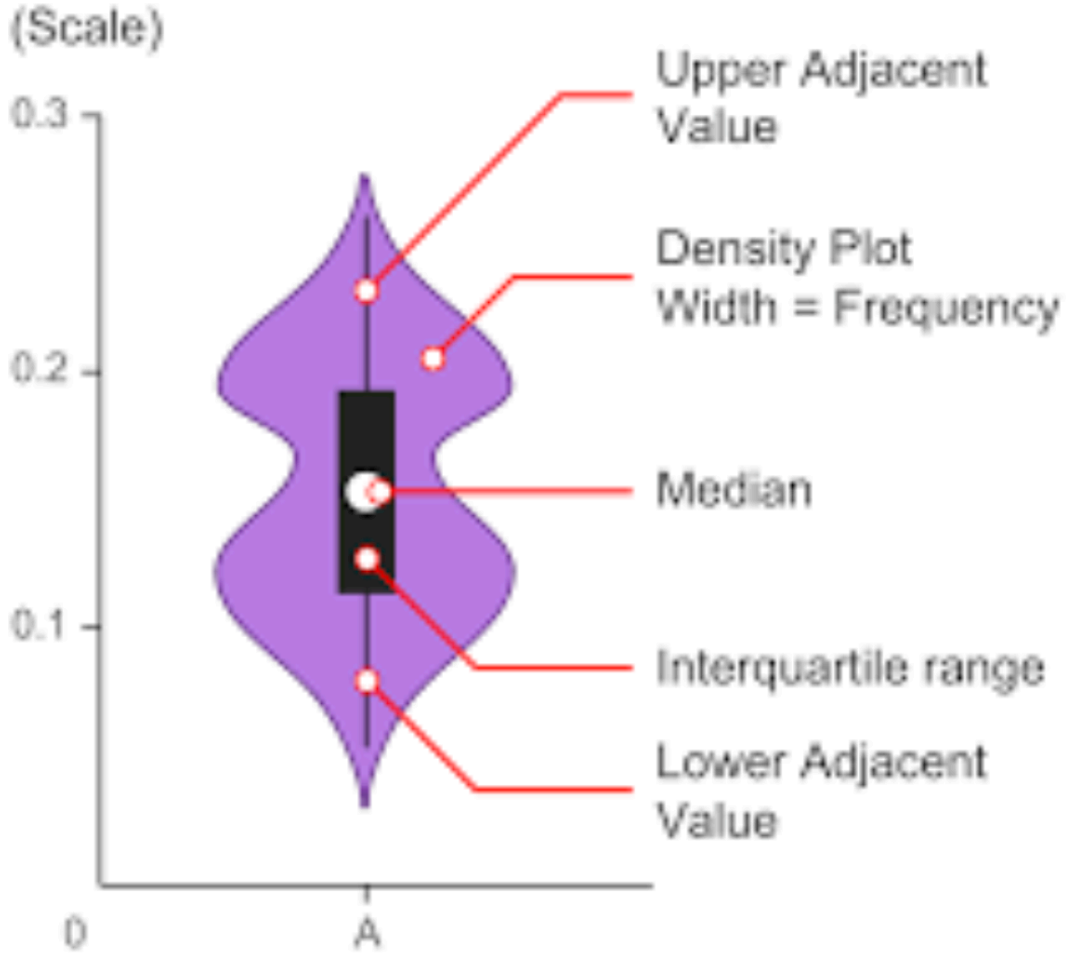
Histogram

Visualization



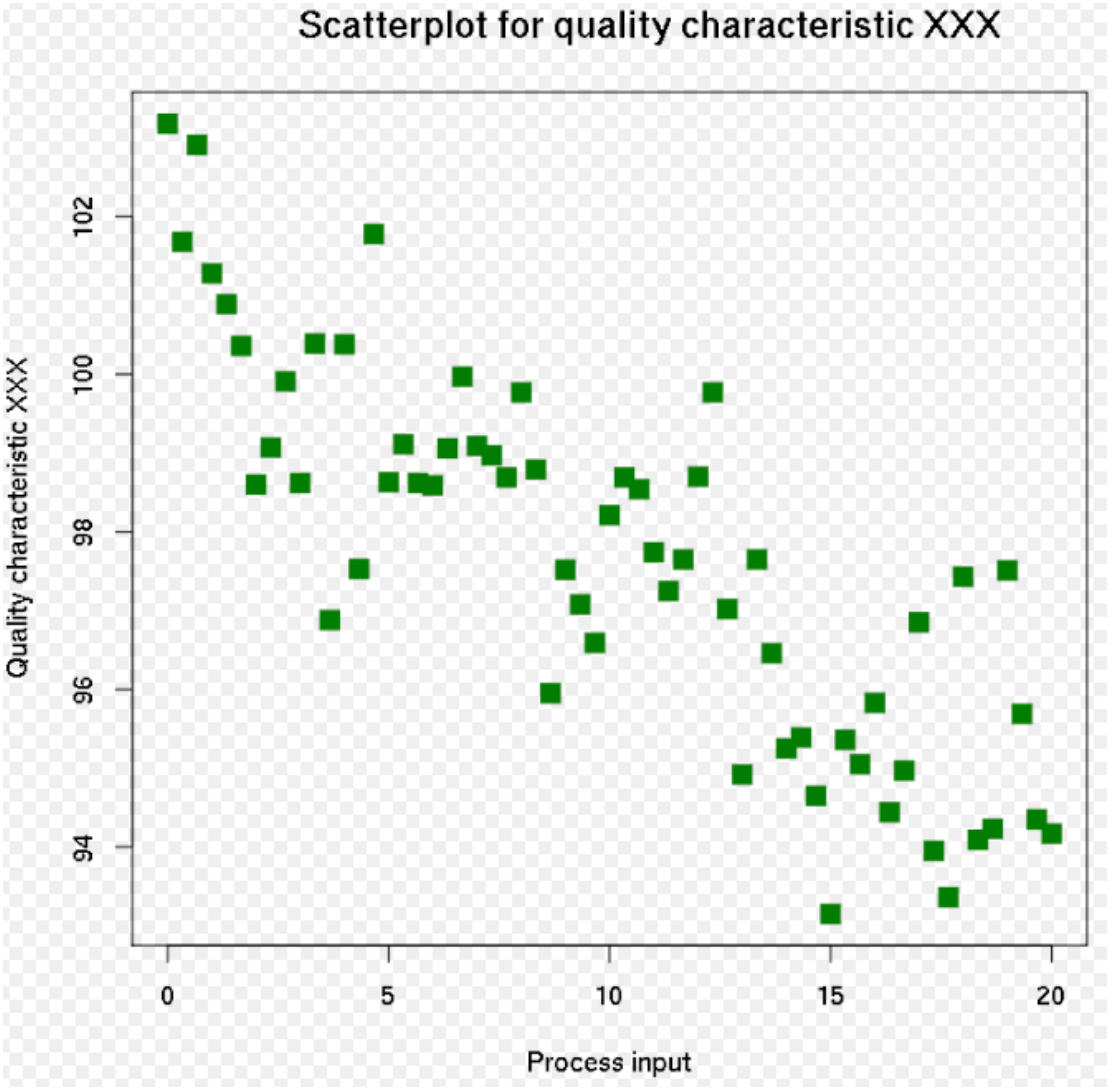
Box-plot

Visualization



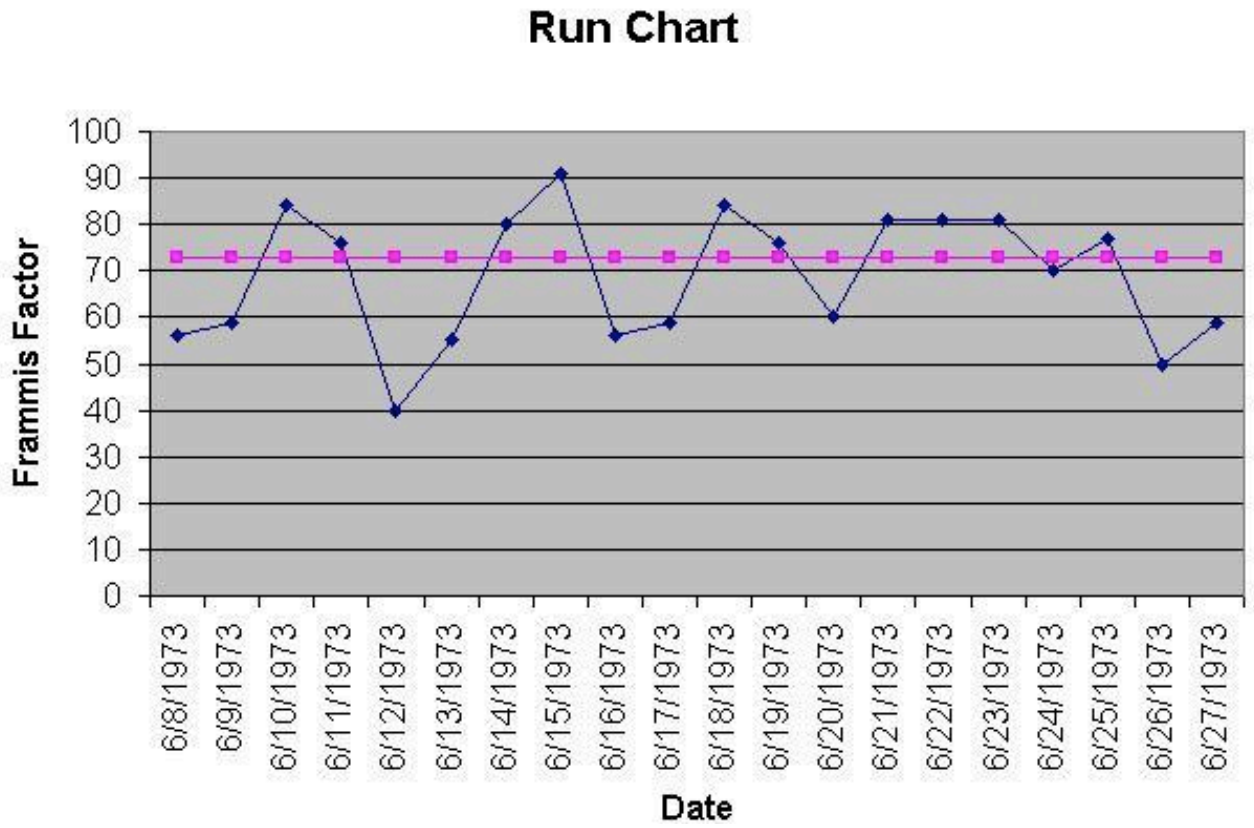
Violin Plot

Visualization



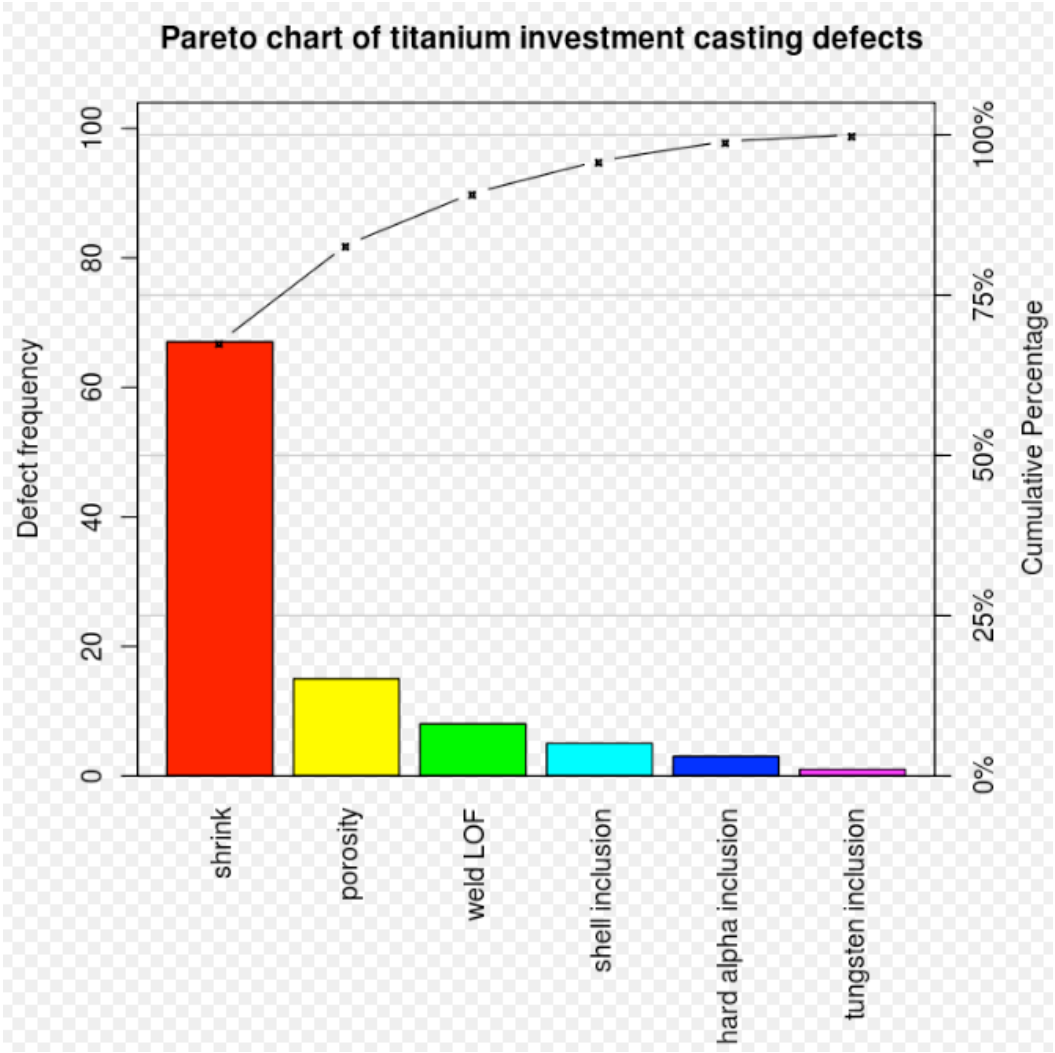
Scatter Plot

Visualization



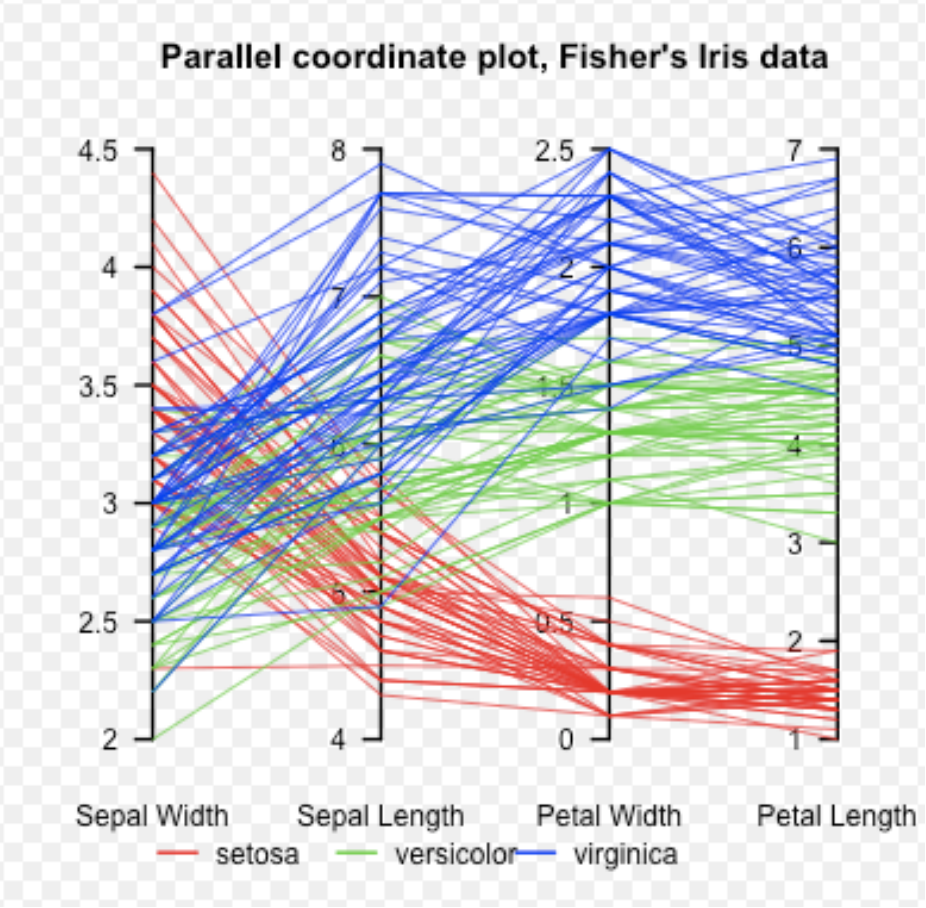
Run Chart

Visualization



Pareto Chart

Visualization

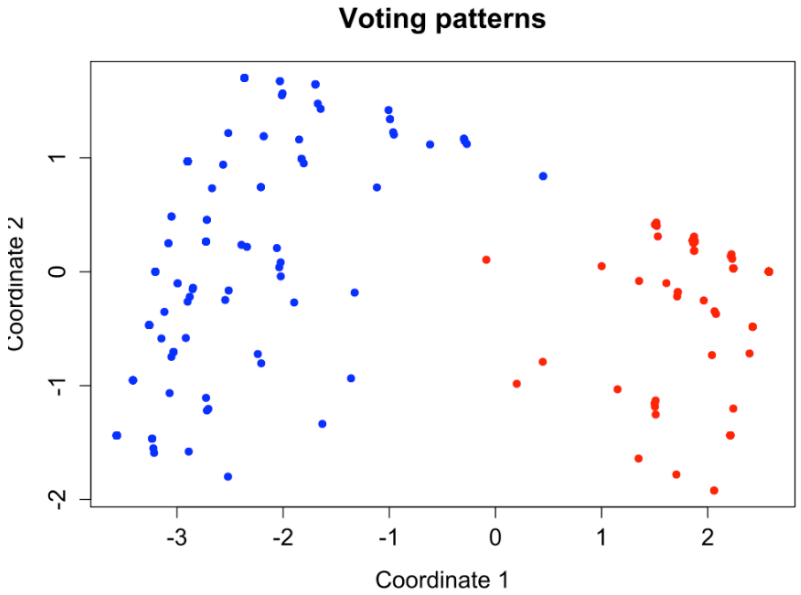


Parallel Coordinates

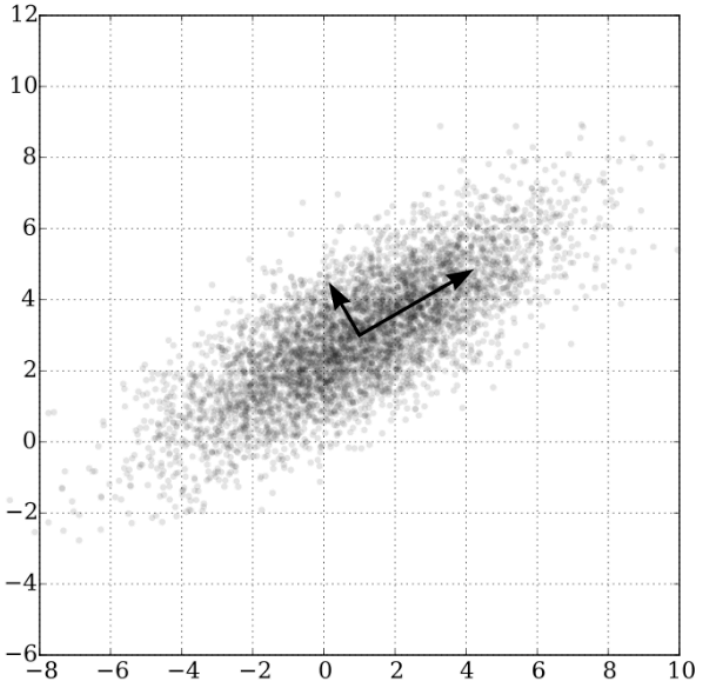
Visualization

Word Clouds

Dimensionality Reduction

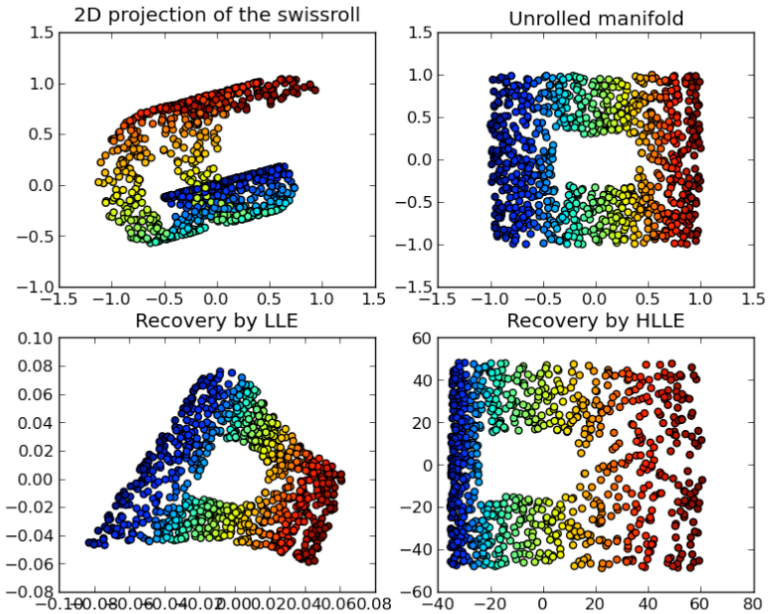


Multidimensional Scaling



Principal Component Analysis

Nonlinear dimensionality reduction



Hands-on Exercises in RStudio Histograms

- Go to script or go to: (Cholesterol and Air-passengers)

- https://www.datacamp.com/community/tutorials/make-histogram-basic-r

Hands-on Exercises in RStudio Boxplot

- Go to script or go to:

- https://www.statmethods.net/graphs/boxplot.html

Hands-on Exercises in RStudio
Violin plot

Two examples:

- Go to script or go to:

- http://www.sthda.com/english/wiki/ggplot2-violin-plot-quick-start-guide-r-software-and-data-visualization

Hands-on Exercises in RStudio Scatterplot 1

- Go to script or go to:

- https://www.statmethods.net/graphs/scatterplot.html

Hands-on Exercises in Rstudio Scatterplot 2

- Continue with same script or go to:

- http://www.sthda.com/english/wiki/scatter-plots-r-base-graphs

Hands-on Exercises in Python

- **Data Visualization:**

https://www.tutorialspoint.com/python_data_science/python_chart_properties.htm

- Style
- Boxplots
- Heatmaps
- Scatterplots:
- Bubble charts:
- 3D charts:

Word Cloud with Python

Wine Dataset

- Go to Notebook or:

https://www.datacamp.com/community/tutorials/wordcloud-python

Francisco J. Cantú-Ortiz, PhD

Professor of Computer Science and Artificial Intelligence
Tecnológico de Monterrey
Enago-Academy Advisor for Strategic Alliances

E-mail: fcantu@tec.mx, fjcantor@gmail.com
Cel: +52 81 1050 8294, SNI-2 CVU: 9804
Personal Page: http://semtech.mty.itesm.mx/fcantu/
Facebook: fcantu; Twitter: @fjcantor; Skype: fjcantor
Orcid: 0000-0002-2015-0562
Scopus ID:6701563520
Researcher ID: B-8457-2009
https://www.researchgate.net/profile/Francisco_Cantu-Ortiz
https://scholar.google.com.mx/citations?hl=es&user=45-uuK4AAAAJ
https://itesm.academia.edu/FranciscoJavierCantuOrtiz
Ave. Eugenio Garza Sada No. 2501, Monterrey N.L., C.P. 64849, México