

EXPLORATORY DATA ANALYSIS

Exploratory data analysis

Goal of this lecture

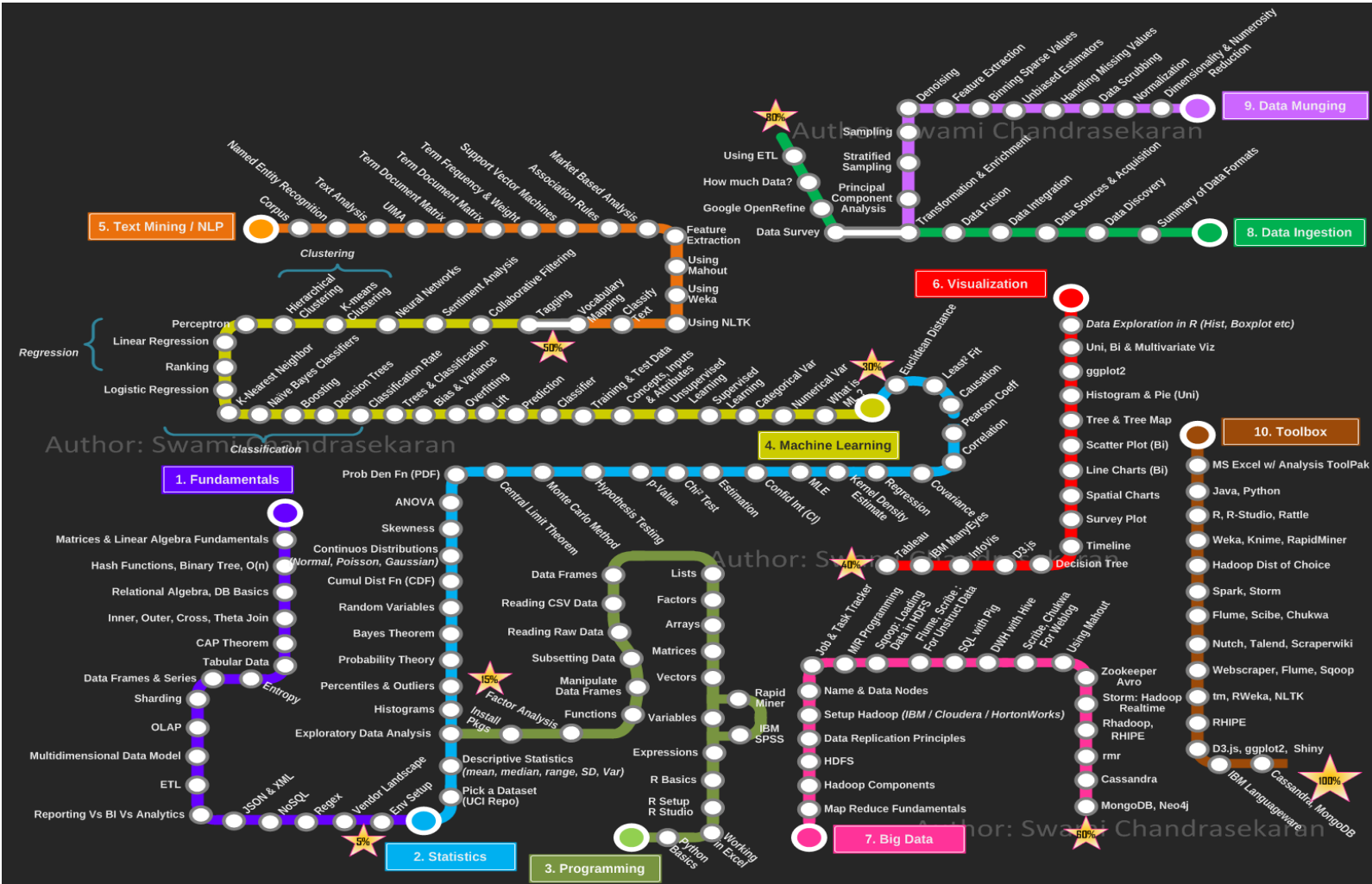
- ▶ Present an overview on the most popular data analysis paradigms
- ▶ Explain the role and benefits of exploratory data analysis
- ▶ Make students familiar with common graphical methods for exploratory data analysis
- ▶ Make students familiar with cluster analysis for exploratory data analysis

Data scientists in demand



Source: <http://www.indeed.com/jobtrends> as of June 2016

The road to data scientist



- Fundamentals
- Statistics
- Programming
- Machine learning
- Text mining
- Visualization
- Big data
- Data munging
- Toolbox

Exploratory data analysis

Overview

- ▶ Introduction: what is data analysis?
- ▶ Data analysis: taxonomies & paradigms
- ▶ Exploratory data analysis
 - ▶ Goal
 - ▶ Graphical methods
 - ▶ Introduction to cluster analysis

Exploratory data analysis

Introduction: what data analysis means?

► Some definitions

1. The process of evaluating data using analytical and logical reasoning to examine each component of the data provided. This form of analysis is just one of the many steps that must be completed when conducting a research experiment. Data from various sources is gathered, reviewed, and then analyzed to form some sort of finding or conclusion.
2. **Analysis of data is a process of inspecting, cleaning, transforming, and modeling data with the goal of discovering useful information, suggesting conclusions, and supporting decision-making.**
3. Data analysis is the systematic study of data in order to understand its meaning, organization, structure, relationships
4. Data analysis refers to the process of applying logical and statistical techniques to evaluate, condense and describe raw data with the sole intent of extracting useful information. There are several different forms of raw data, including observations, survey responses and measurements.

Exploratory data analysis

Overview

- ▶ Introduction: what is data analysis?
- ▶ Data analysis: taxonomies & paradigms
- ▶ Exploratory data analysis
 - ▶ Goal
 - ▶ Graphical methods
 - ▶ Introduction to cluster analysis



Exploratory data analysis

Data analysis: taxonomies

- ▶ Qualitative **vs** Quantitative
- ▶ Exploratory **vs** Confirmatory
- ▶ Descriptive **vs** Inferential
- ▶ Univariate **vs** Bivariate **vs** Multivariate



6 archetypical analysis

Descriptive
Exploratory
Inferential
Predictive
Causal
Mechanistic

Identified by **Jeffrey Leek**, *Assistant professor of Biostatistics at John Hopkins Bloomberg School of Public Health*

Exploratory data analysis

Overview

- ▶ Introduction: what is data analysis? 
- ▶ Data analysis: taxonomies & paradigms 
- ▶ Exploratory data analysis
 - ▶ Goal
 - ▶ Graphical methods
 - ▶ Introduction to cluster analysis

Exploratory data analysis

When EDA is useful?

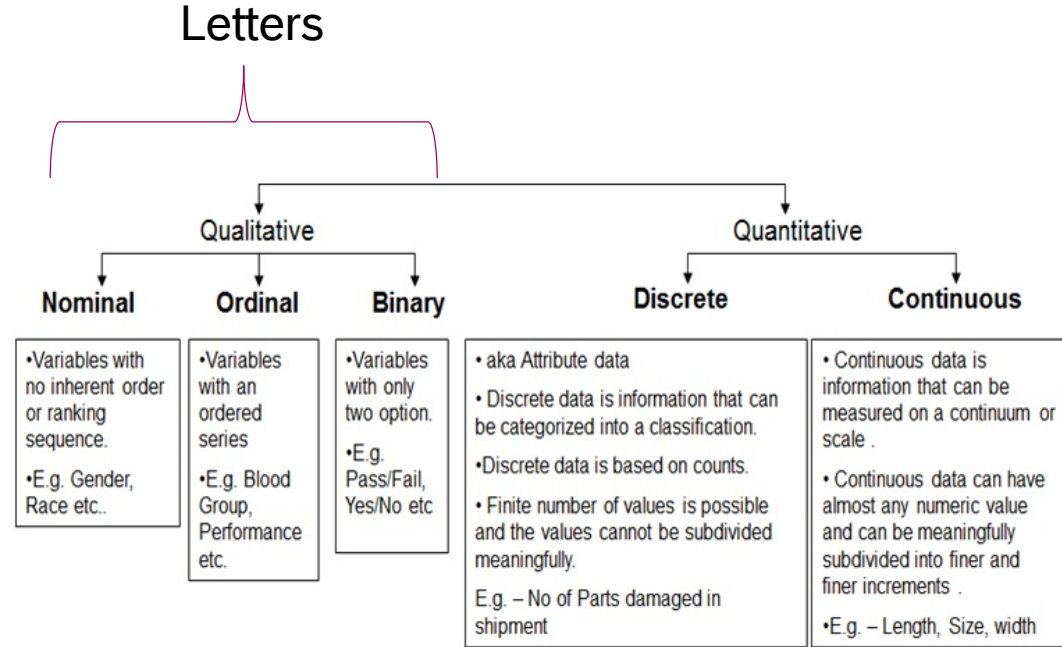
- ▶ Suggest hypotheses about the causes of observed phenomena
- ▶ Assess assumptions and provide support for the selection of appropriate statistical tools and techniques
- ▶ Provide a basis for further data collection through surveys or experiments

Exploratory data analysis

Some definitions

► Real life use-case, presentation of a dataset

↓	C1	C2	C3	C4	C5-T
	Sepal length	Sepal width	Petal length	Petal width	Species
1	5.1	3.5	1.4	0.2	I. setosa
2	4.9	3.0	1.4	0.2	I. setosa
3	4.7	3.2	1.3	0.2	I. setosa
4	4.6	3.1	1.5	0.2	I. setosa
5	5.0	3.6	1.4	0.2	I. setosa
6	5.4	3.9	1.7	0.4	I. setosa
7	4.6	3.4	1.4	0.3	I. setosa
8	5.0	3.4	1.5	0.2	I. setosa
9	4.4	2.9	1.4	0.2	I. setosa

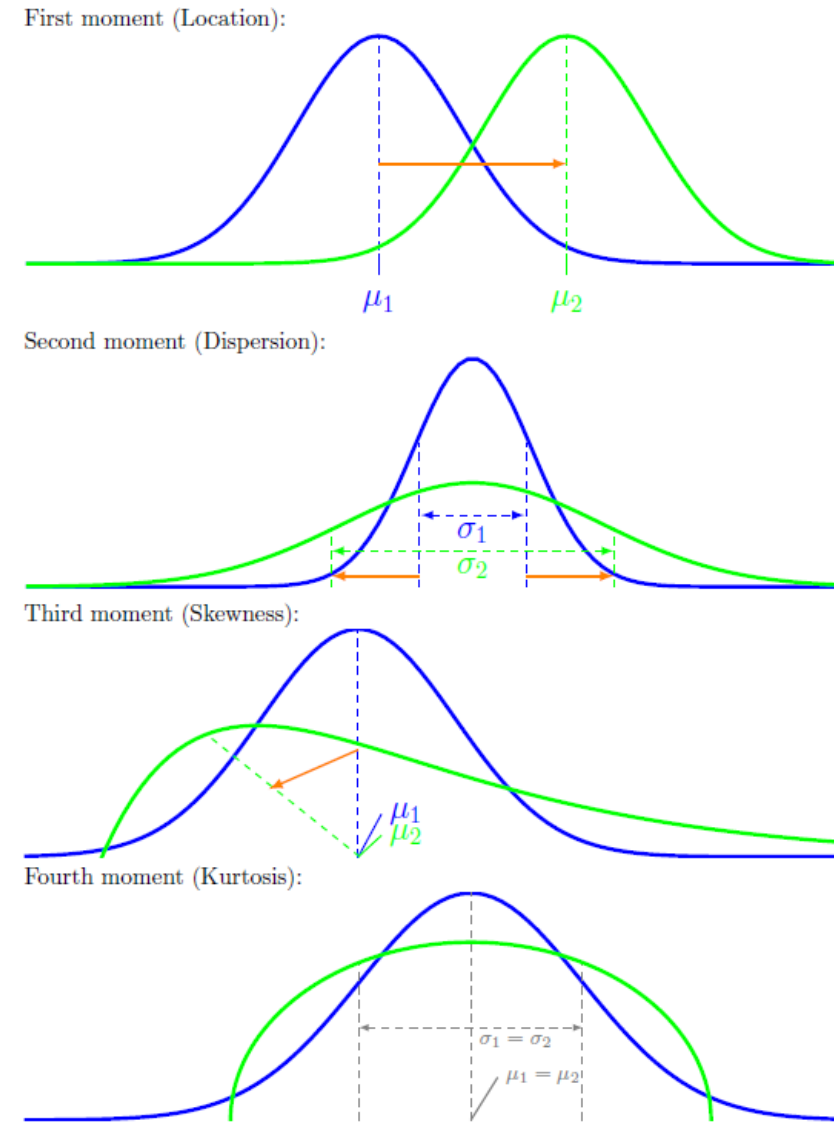


Numbers, can do
maths on it

Exploratory data analysis




Data summarization (descriptive analysis)

- ▶ Central tendency: mean, median, mode
- ▶ Spread: variance, inter-quantile range, range (min - max)
- ▶ Skewness (asymmetry)
- ▶ Kurtosis (heavy tails)
- ▶ Missing data
- ▶ Outliers



Exploratory data analysis

Overview

- ▶ Introduction: what is data analysis? 
- ▶ Data analysis: taxonomies & paradigms 
- ▶ Exploratory data analysis 
 - ▶ Goal
 - ▶ Graphical methods
 - ▶ Introduction to cluster analysis

Exploratory data analysis

Graphical methods for exploratory analysis

Univariate

- Barplots
- Pareto charts
- Histograms
- Boxplots
- Density plots
- Run plot
- Lag plot
- Autocorrelation plot
- Normal probability plot
- 4 plot

Bivariate

- QQ plot
- Scatterplot
- Smooth scatterplot

Multivariate

- Correlation matrix
- Heatmaps

Univariate graphical methods for EDA

Barplot

! Commonly associated with categorical variables

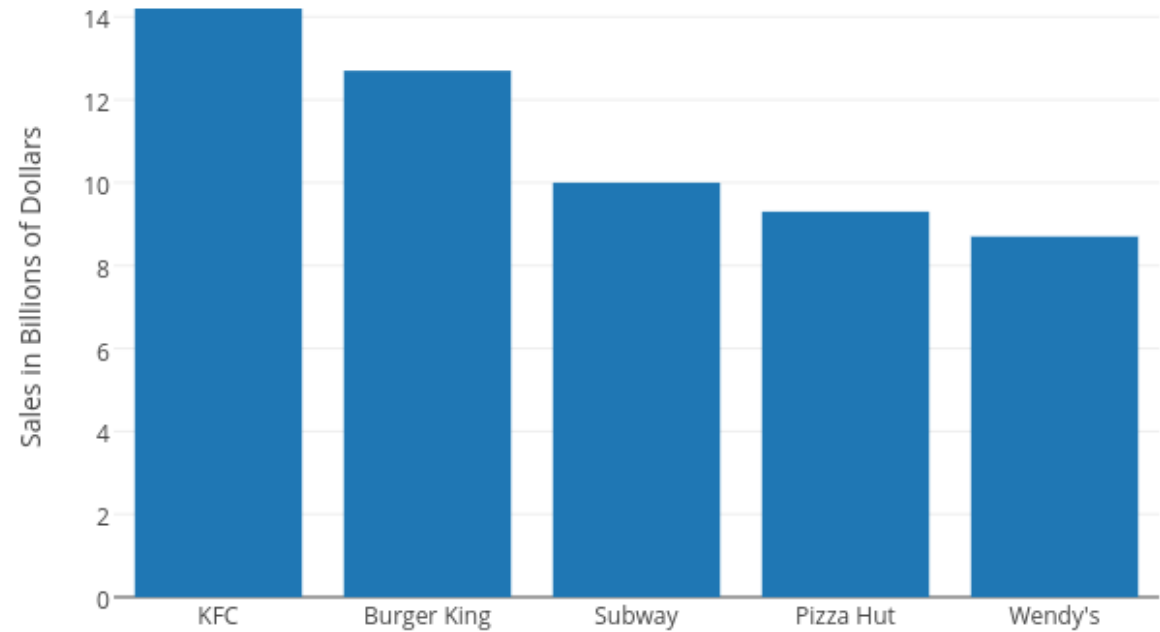
Definition:

Vertical axis: Aggregated values (e.g. counts, percentage) for each category
Horizontal axis: The factor of interest

Purpose:

Used to show comparisons between categories of data

Worldwide Sales of Fast Food



Univariate graphical methods for EDA

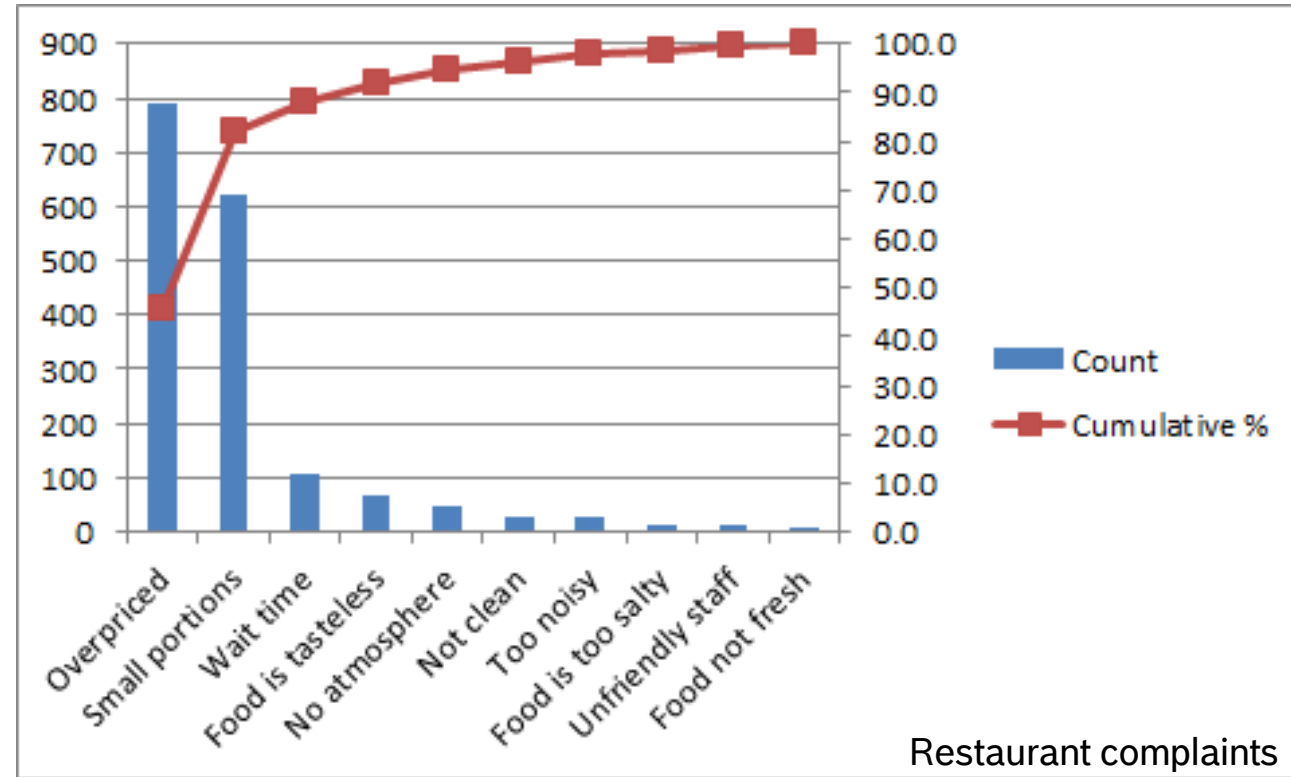
Pareto chart

! Commonly associated with categorical variables

Definition:

Same as a bar chart

In addition, it also contains a cumulative percentage of the values quantified for each category



Univariate graphical methods for EDA

Boxplot

! Commonly associated with quantitative variables

! Can also be used as a bivariate tool, but one variable is always quantitative

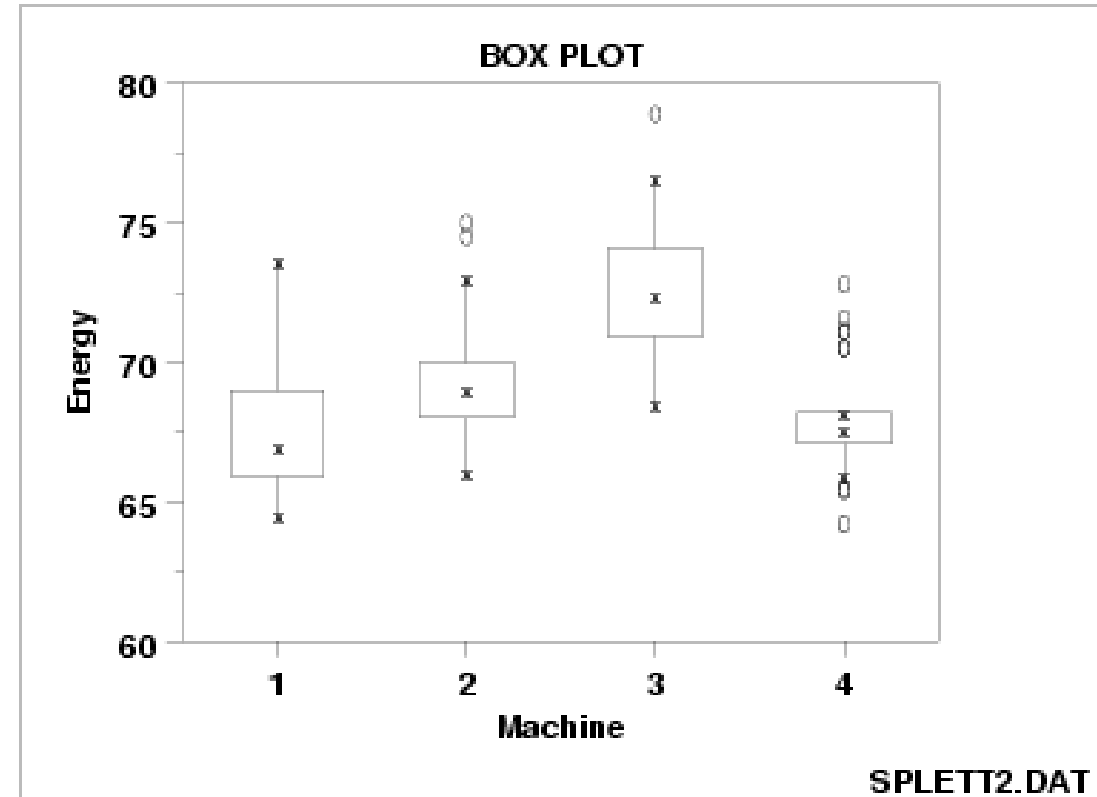
Definition:

Vertical axis: Response variable

Horizontal axis: The factor of interest

Purpose:

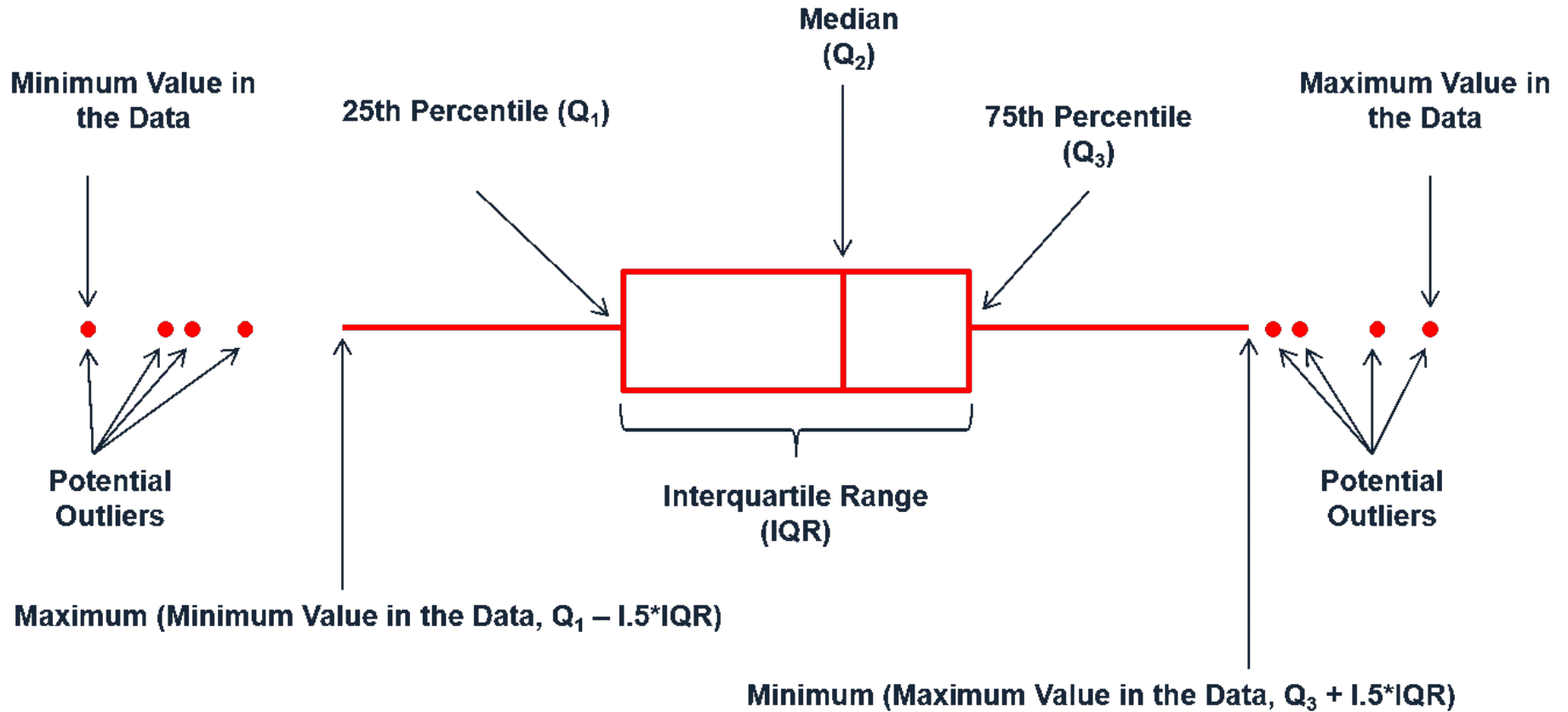
To graphically summarize the distribution of a univariate data set.



Univariate graphical methods for EDA

Boxplot

Description:

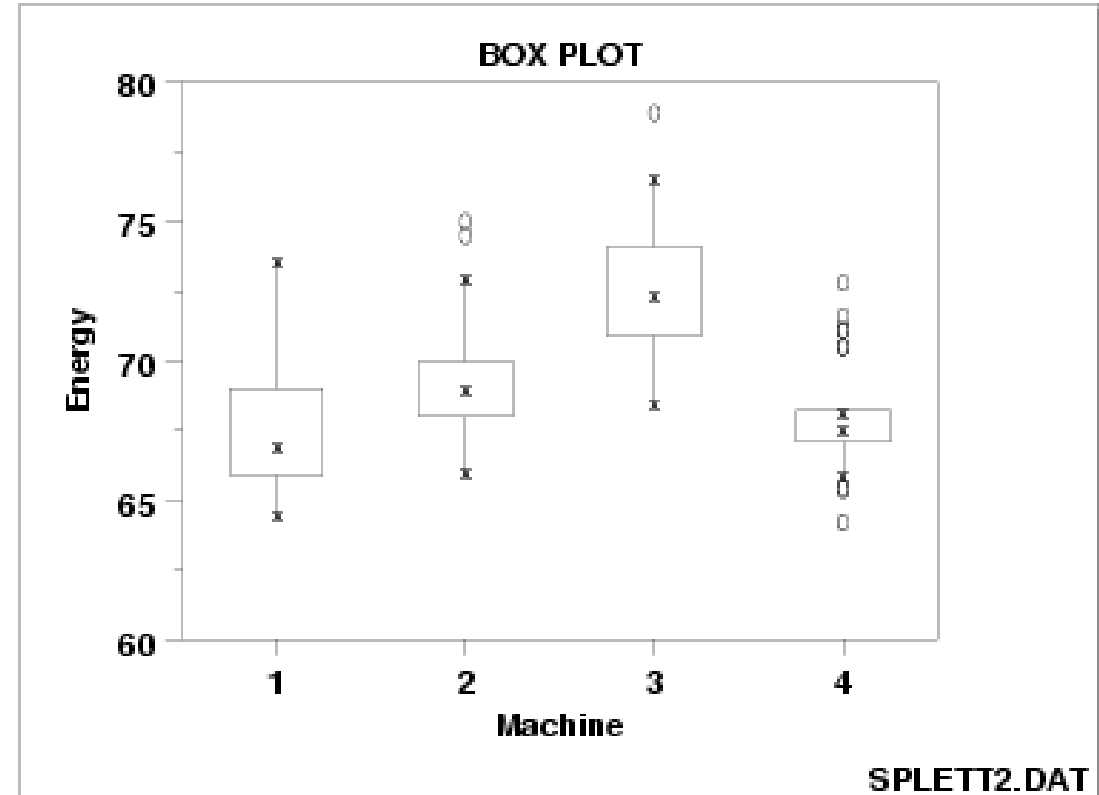


Univariate graphical methods for EDA

Boxplot

Questions addressed:

1. Is a factor significant?
2. Does the location differ between subgroups?
3. Does the variation differ between subgroups?
4. Are there any outliers?



Univariate graphical methods for EDA

Histogram

!Only associated with quantitative variables

Definition:

Vertical axis: Frequency or count

Horizontal axis: Response variable

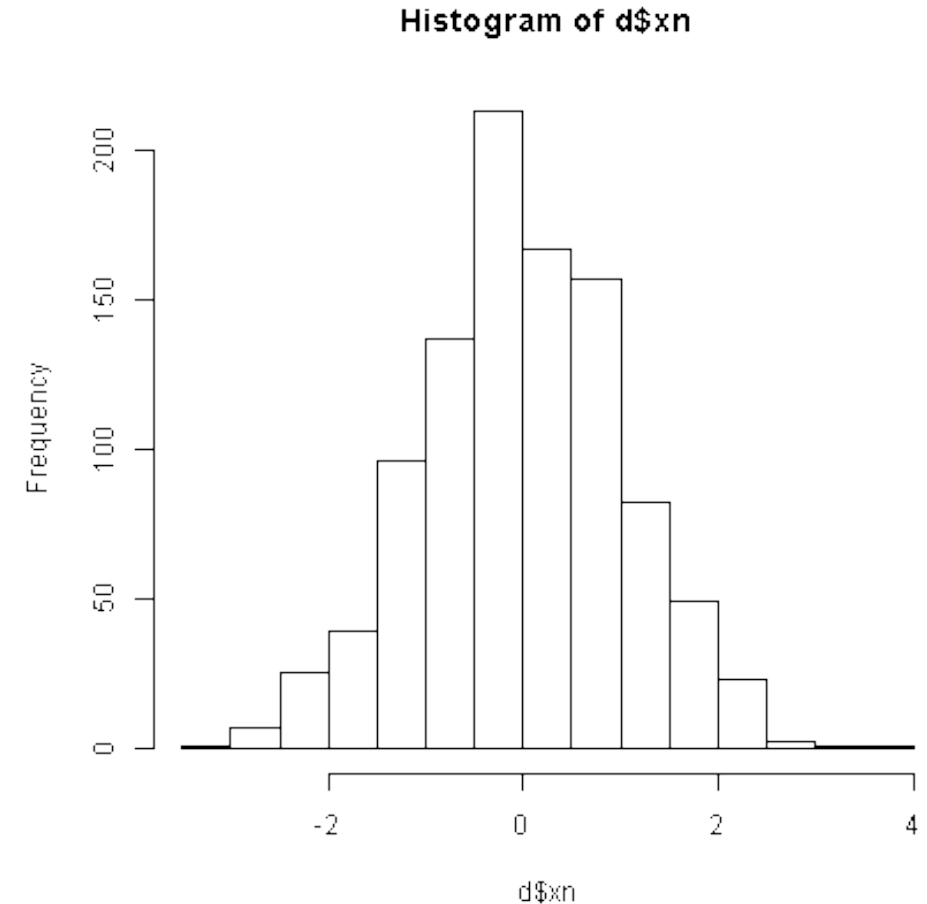
Purpose:

To graphically summarize the distribution of a univariate data set.

Important parameter: the bin size.

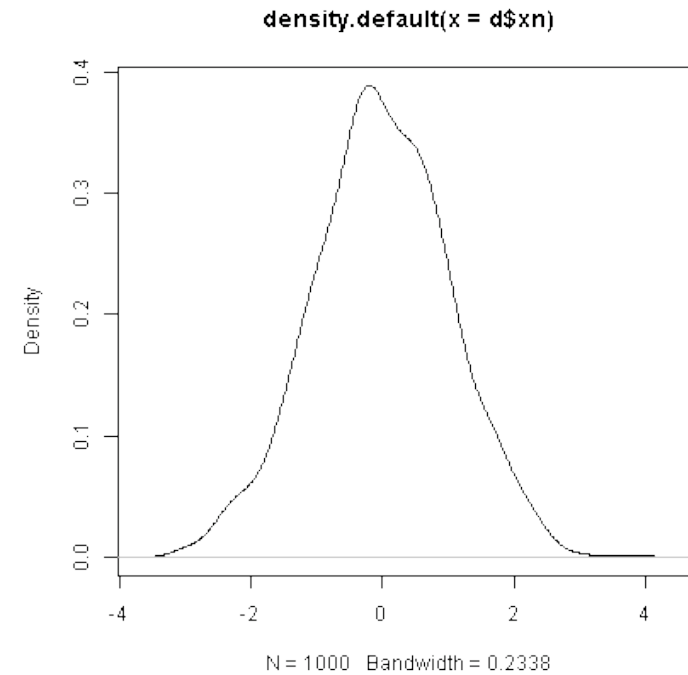
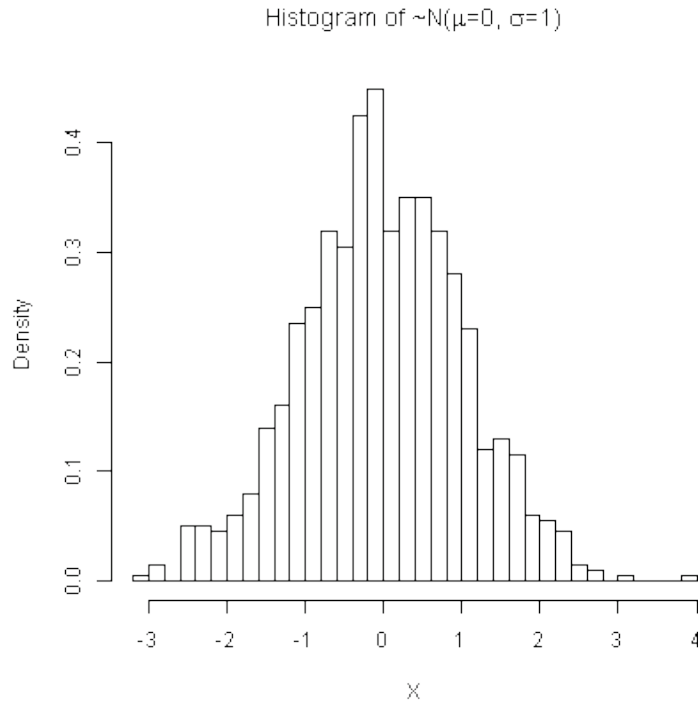
A number of theoretically derived rules have been proposed by Scott & Freedman

(<https://www.fmrib.ox.ac.uk/datasets/techrep/tr00mj2/tr00mj2/node24.html>)



Univariate graphical methods for EDA

Density plots



Univariate graphical methods for EDA

Run plots

!Only associated with quantitative variables

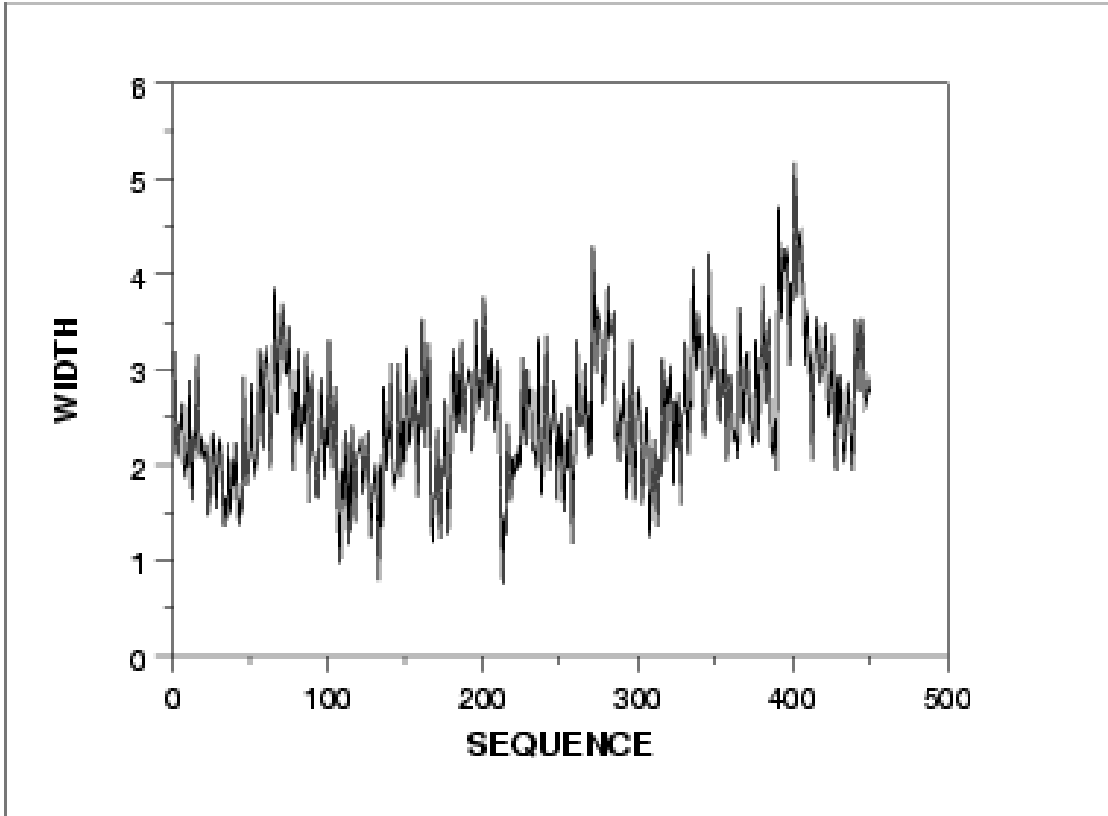
Definition:

Vertical axis: Response variable Y_i

Horizontal axis: Index i ($i = 1, 2, 3, \dots$)

Questions addressed:

- 1.Are there any shifts in location?
- 2.Are there any shifts in variation?
- 3.Are there any outliers



Univariate graphical methods for EDA

Lag plots

!Only associated with quantitative variables

Definition:

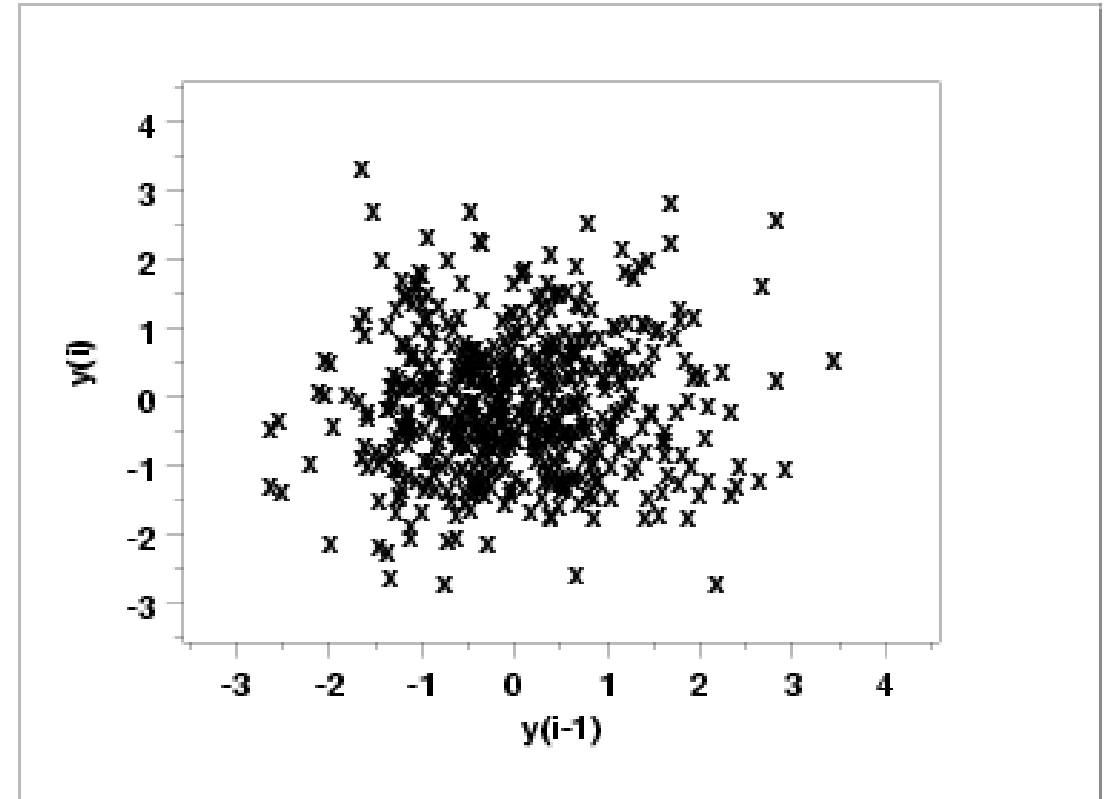
A lag is a fixed time displacement.

Vertical axis: Y_i for all i

Horizontal axis: Y_{i-1} for all i

Questions addressed:

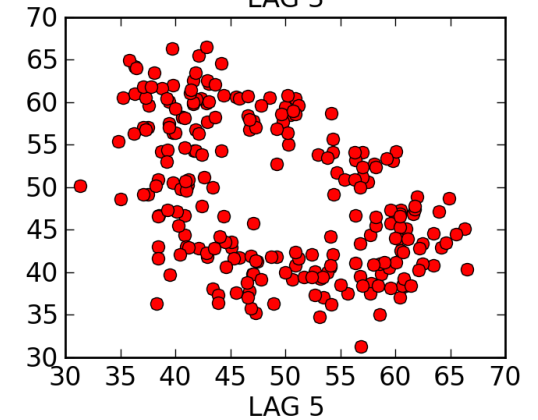
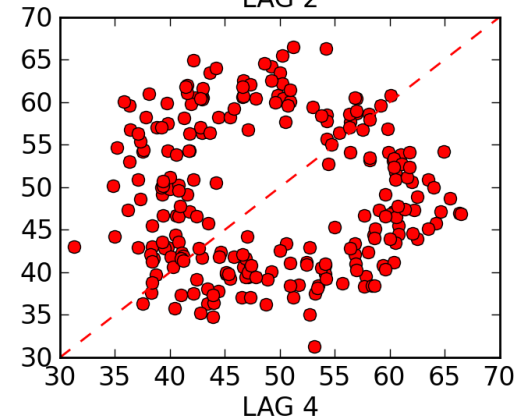
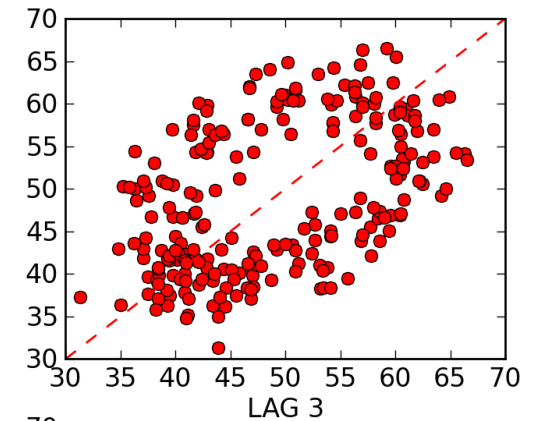
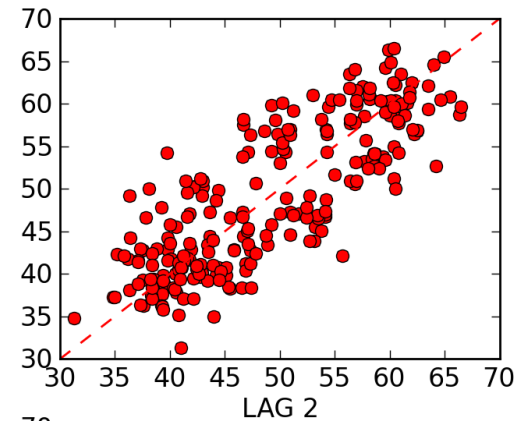
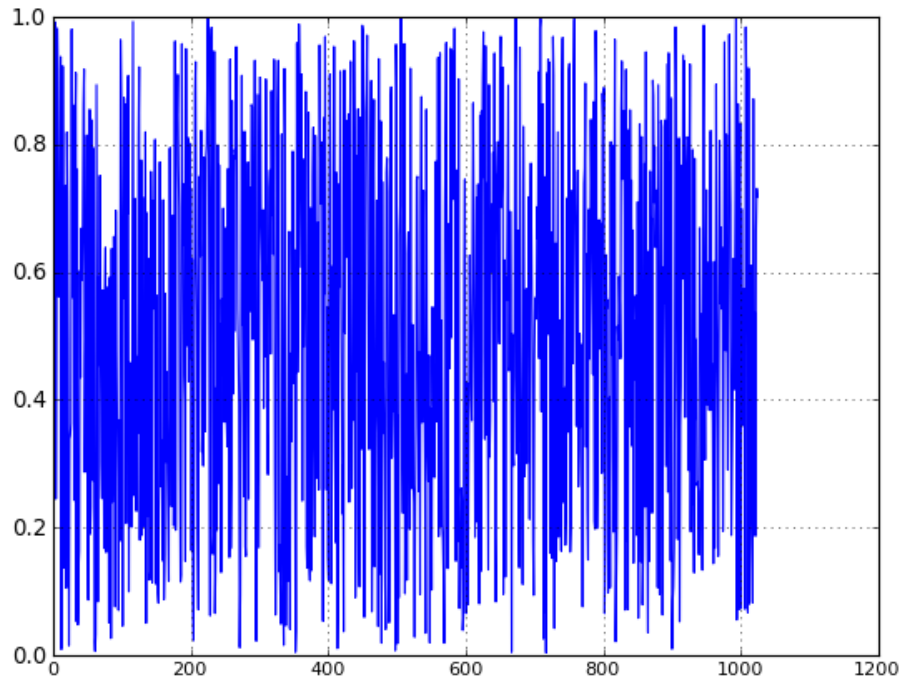
1. Are the data random?
2. Is there serial correlation in the data?
3. What is a suitable model for the data?
4. Are there outliers in the data?



Univariate graphical methods for EDA

Lag plots

Example non random data



Univariate graphical methods for EDA

Autocorrelation plot

!Only associated with quantitative variables

Definition:

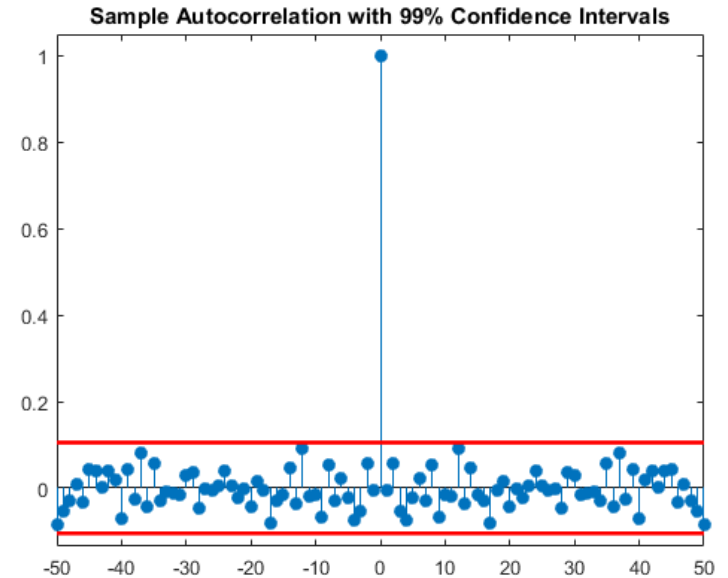
Vertical axis: Autocorrelation coefficient

Horizontal axis: Time lag h ($h = 1, 2, 3, \dots$)

Purpose:

To test for randomness

Autocorrelation plots are also used in the model identification stage for fitting ARIMA models.



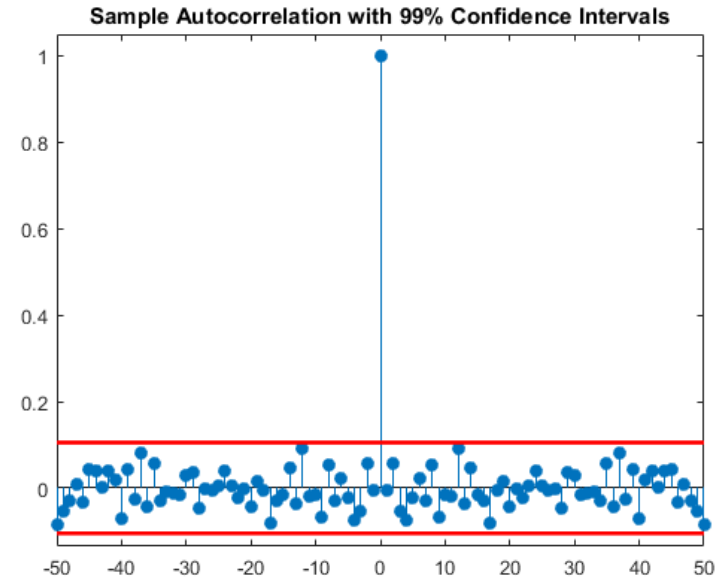
Univariate graphical methods for EDA

Autocorrelation plot

Questions addressed:

1. Are the data random?
2. Is an observation related to an adjacent observation?
3. Is an observation related to an observation twice-removed? (etc.)
4. Is the observed time series white noise?
5. Is the observed time series sinusoidal?
6. Is the observed time series autoregressive?
7. What is an appropriate model for the observed time series?
8. Is the model

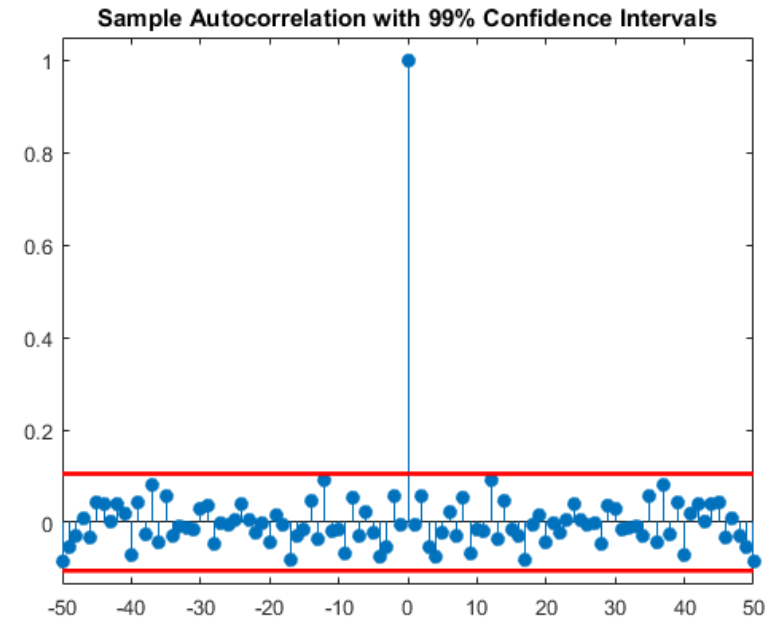
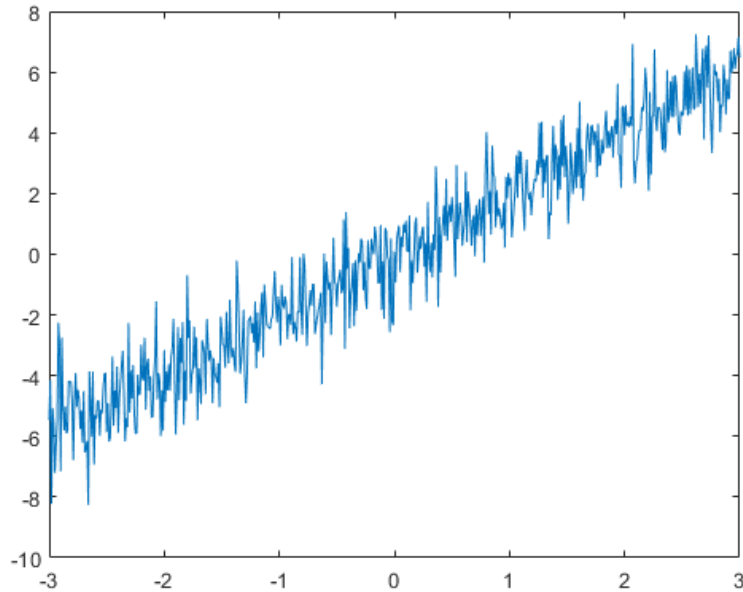
$Y = \text{constant} + \text{error}$
valid and sufficient?



Univariate graphical methods for EDA

Autocorrelation plot

Example random data



Univariate graphical methods for EDA

Normal probability plot

!Only associated with quantitative variables

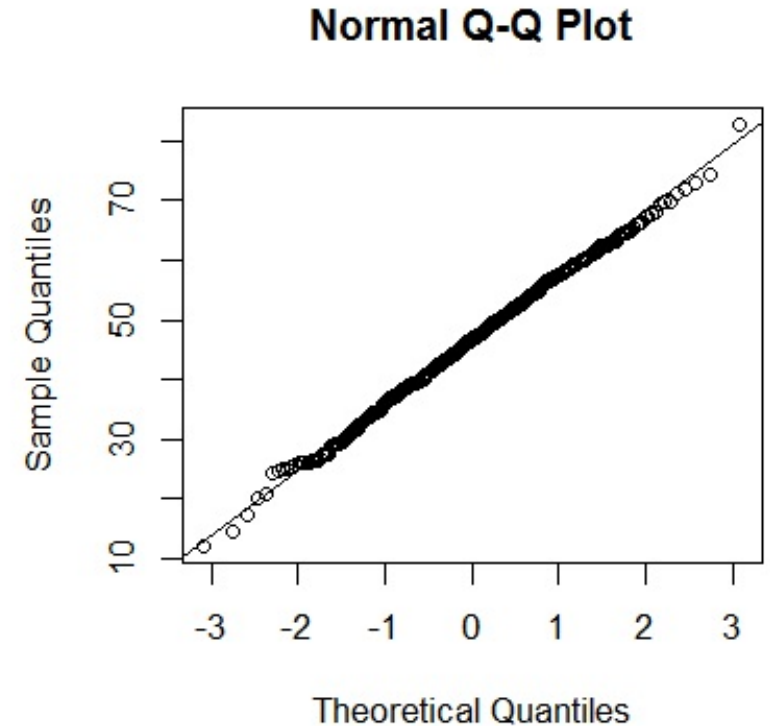
Definition:

Vertical axis: Observed quantiles

Horizontal axis: Normal theoretical quantiles

Questions addressed:

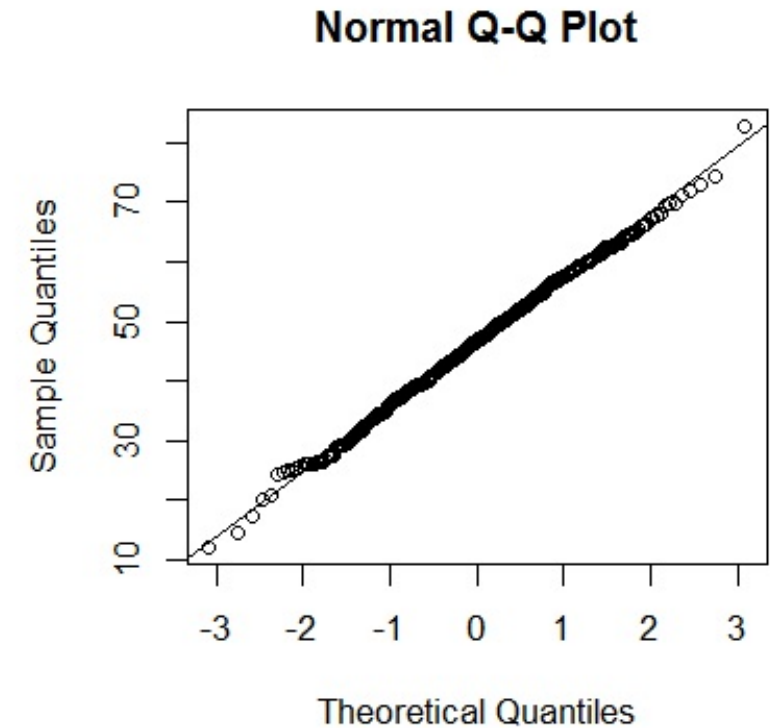
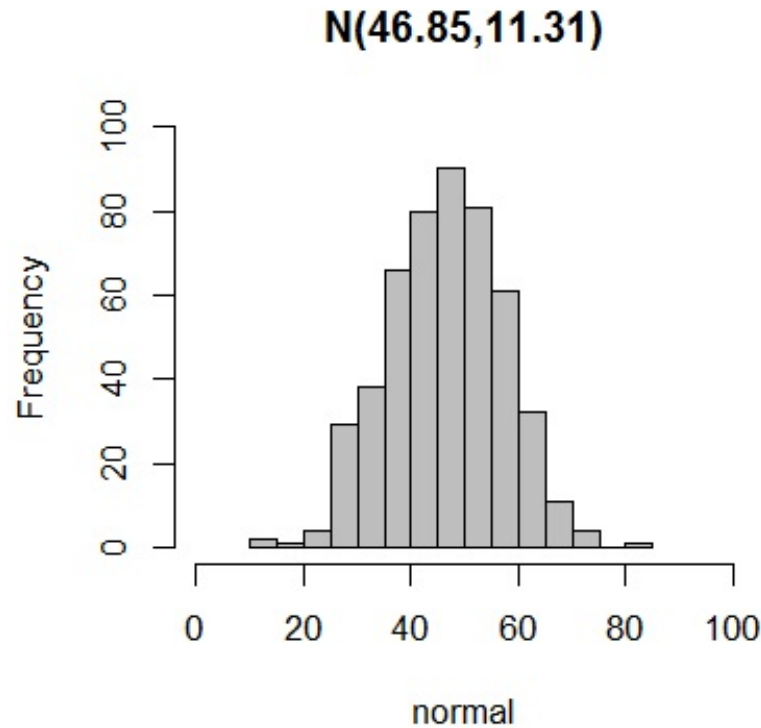
1. Are the data normally distributed?
2. What is the nature of the departure from normality (data skewed, shorter than expected tails, longer than expected tails)?



Univariate graphical methods for EDA

Normal probability plot

Example normal probability plot for normal distributed data



Univariate graphical methods for EDA

4 plot

Run sequence plot

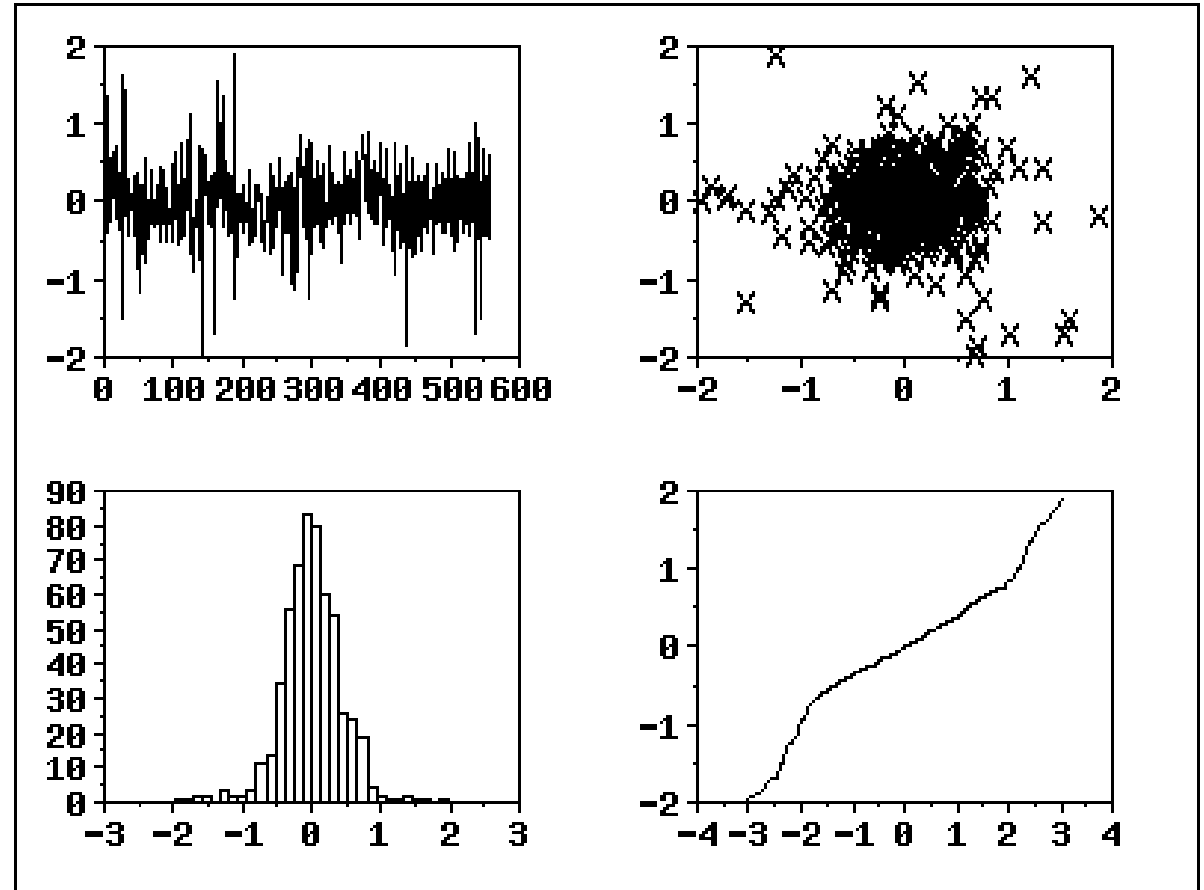
Lag plot

Histogram

Normal probability plot

Used to test the following assumptions

1. Fixed Location
2. Fixed Variation
3. Randomness
4. Fixed Distribution



Exploratory data analysis

Graphical methods for exploratory analysis

Univariate

- Barplots
- Pareto charts
- Histograms
- Boxplots
- Density plots
- Run plot
- Lag plot
- Autocorrelation plot
- Normal probability plot
- 4 plot

Bivariate

- QQ plot
- Scatterplot
- Smooth scatterplot

Multivariate

- Correlation matrix
- Heatmaps

Bivariate graphical methods for EDA

QQ plots

!Only associated with quantitative variables

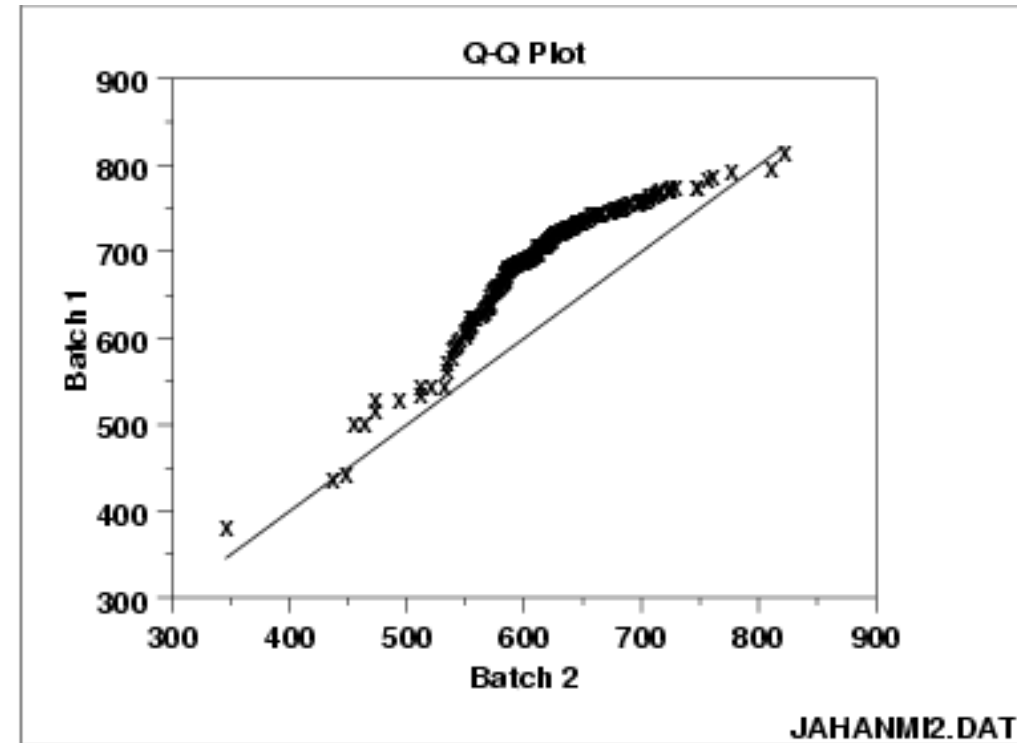
Definition:

Vertical axis: Estimated quantiles from data set 1

Horizontal axis: Estimated quantiles from data set 2

Questions addressed:

1. Do two data sets come from populations with a common distribution?
2. Do two data sets have common location and scale?
3. Do two data sets have similar distributional shapes?
4. Do two data sets have similar tail behavior?



Bivariate graphical methods for EDA

Scatterplots

!Only associated with quantitative variables

Definition:

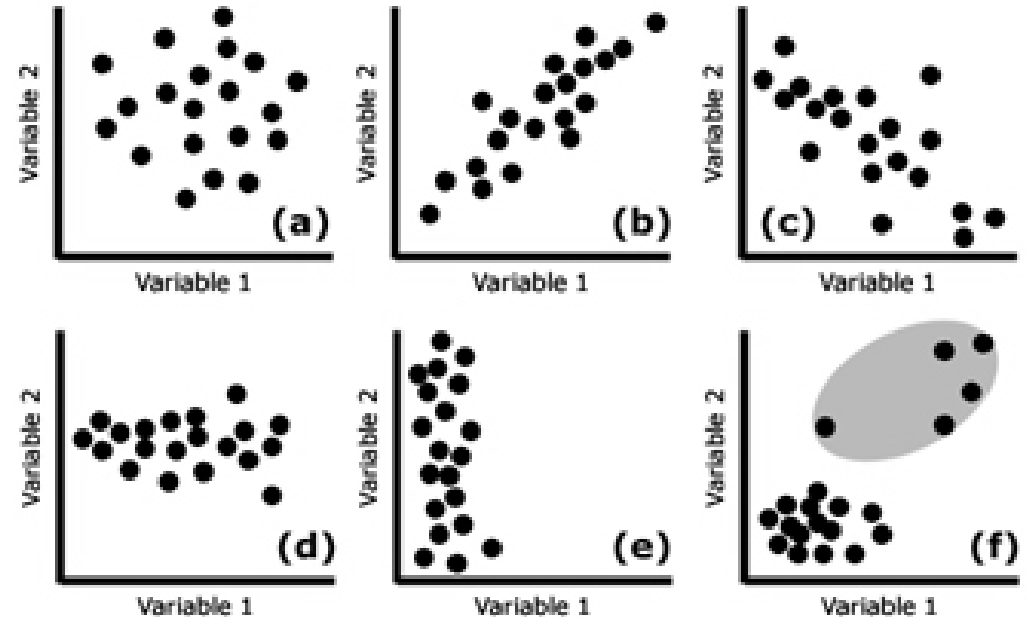
Vertical axis: variable Y --usually the response variable

Horizontal axis: variable X --usually some variable we suspect may be related to the response

Questions addressed:

- 1.Are variables X and Y related?
- 2.Are variables X and Y linearly related?
- 3.Are variables X and Y non-linearly related?
- 4.Does the variation in Y change depending on X ?
- 5.Are there outliers?

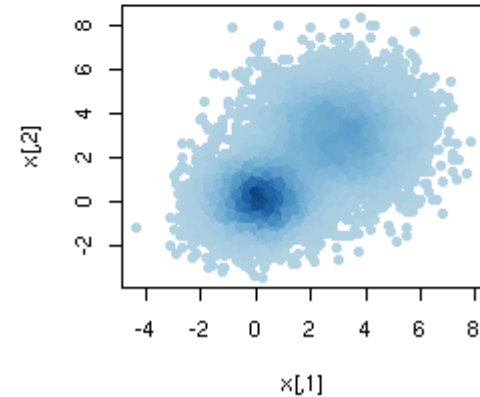
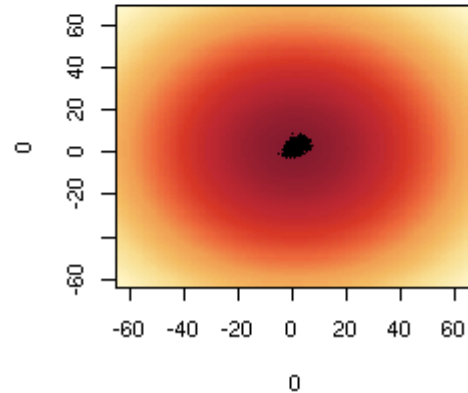
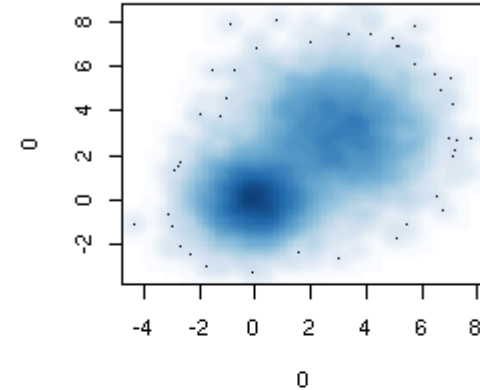
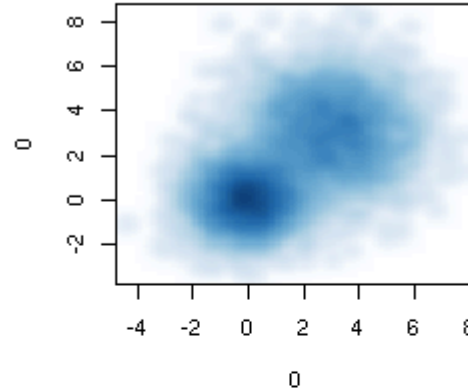
Example scatter plots



Bivariate graphical methods for EDA

Smooth scatterplots

If you have a lot of points...



Exploratory data analysis

Graphical methods for exploratory analysis

Univariate

- Barplots
- Pareto charts
- Histograms
- Boxplots
- Density plots
- Run plot
- Lag plot
- Autocorrelation plot
- Normal probability plot
- 4 plot

Bivariate

- QQ plot
- Scatterplot
- Smooth scatterplot

Multivariate

- Correlation matrix
- Heatmaps

Multivariate graphical methods for EDA

Correlation matrix

!Associated with quantitative or qualitative variables depending on the way correlation is calculated

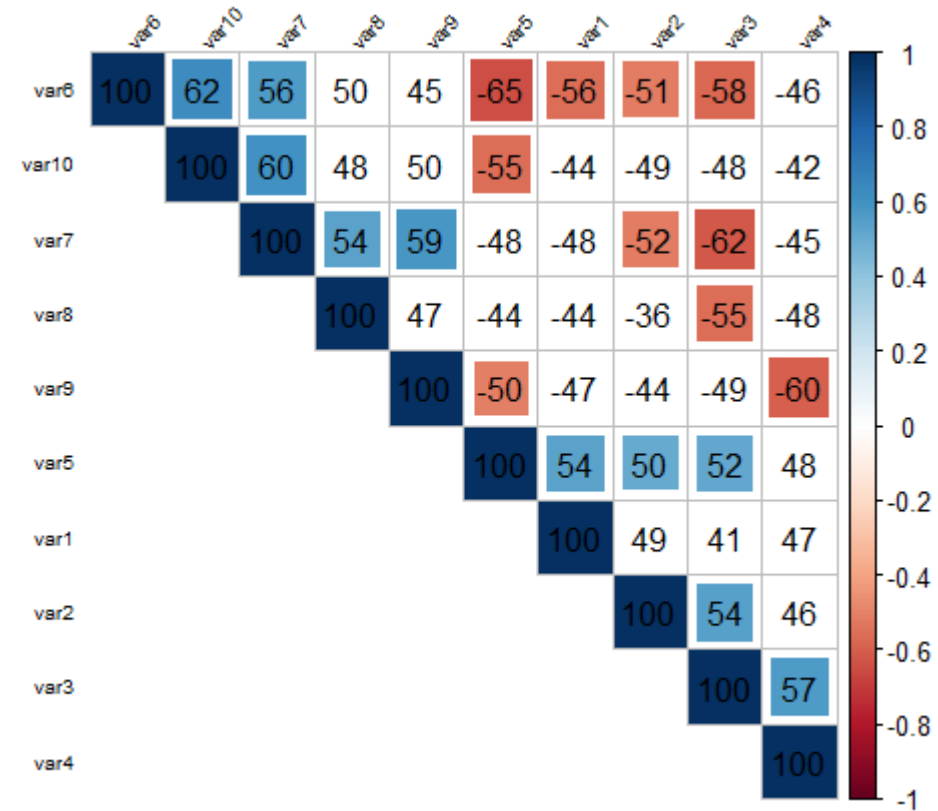
Definition:

Correlation test is used to evaluate the association between two or more variables

Correlation matrix is symmetrical

Questions addressed:

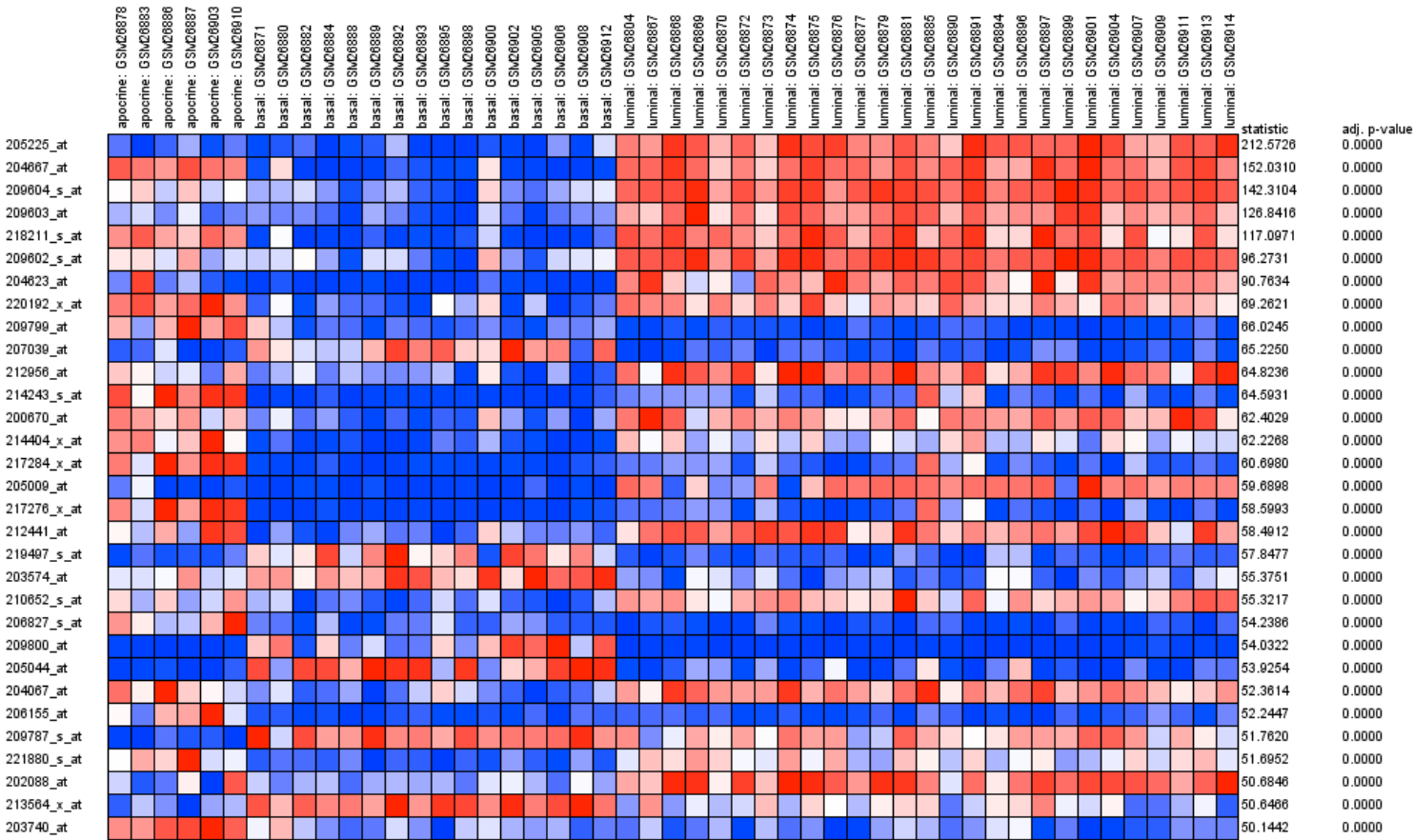
1. Is there correlations between variables ?



Multivariate graphical methods for EDA





Heat maps

Graphical representation of data where the individual values contained in a matrix are represented as colors.



Exploratory data analysis

Overview

- ▶ Introduction: what is data analysis? 
- ▶ Data analysis: taxonomies & paradigms 
- ▶ Exploratory data analysis
 - ▶ Goal 
 - ▶ Graphical methods 
 - ▶ Introduction to cluster analysis