Today

- Graphics and Exploratory Data Analysis
    - Classical Data Analysis vs. EDA
    - Box plot
    - Histogram
    - ECDF plot
    - Run chart
    - Lag plot
    - Scatter plot
    - Q-Q plot
    - ROC curves

Exploratory Data Analysis (EDA)

- EDA is an approach to analyse data that employs various techniques to :
  - Find structure in the data
  - Extract relevant variables
  - Detect outliers and anomalies
  - Test underlying assumptions

Exploratory Data Analysis (EDA)

- EDA vs. Classical Data Analysis

EDA:

- The focus is on the data
- Find structure and outliers
- Find models suggested by the data
- It is based on graphical techniques

Classical:

- The focus is on the model
- Estimate parameters of the model
- Generate predicted values from the model
- It uses statistical tests and regression models

Exploratory Data Analysis (EDA)

- EDA vs. Classical Data Analysis

EDA:

- No or few assumptions on the data
- It is suggestive, indicative and subjective

Classical:

- It depends on underlying assumptions (e.g. normality)
- It is rigorous, formal and objective

An Example of EDA

- A classical example*

Group 1		Group 2		Group 3		Group 4	
X	Y	X	Y	X	Y	X	Y
10.00	8.04	10.00	9.14	10.00	7.46	8.00	6.58
8.00	6.95	8.00	8.14	8.00	6.77	8.00	5.76
13.00	7.58	13.00	8.74	13.00	12.74	8.00	7.71
9.00	8.81	9.00	8.77	9.00	7.11	8.00	8.84
11.00	8.33	11.00	9.26	11.00	7.81	8.00	8.47
14.00	9.96	14.00	8.10	14.00	8.84	8.00	7.04
6.00	7.24	6.00	6.13	6.00	6.08	8.00	5.25
4.00	4.26	4.00	3.10	4.00	5.39	19.00	12.50
12.00	10.84	12.00	9.13	12.00	8.15	8.00	5.56
7.00	4.82	7.00	7.26	7.00	6.42	8.00	7.91
5.00	5.68	5.00	4.74	5.00	5.73	8.00	6.89

* Taken from the Engineering Statistics Handbook website

An Example of EDA

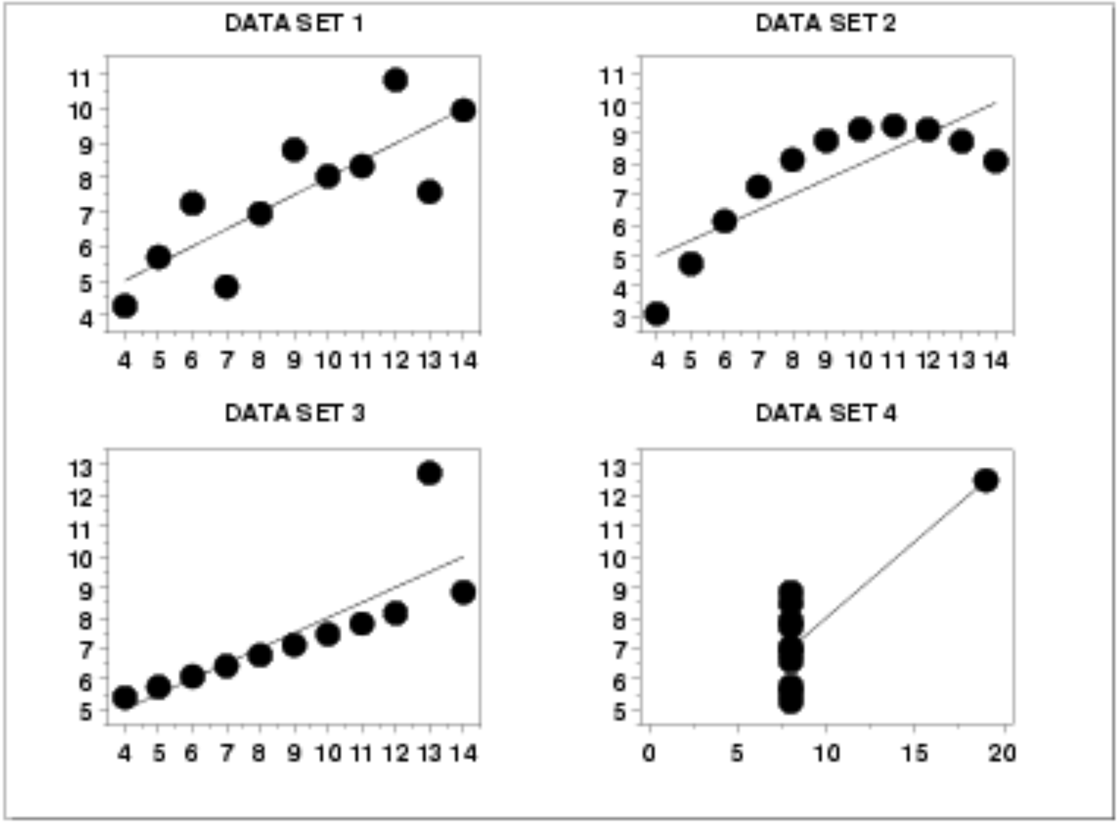
- Summary of data and best linear fit for Y as function of X (values are similar to all groups):

	Group 1	Group 2	Group 3	Group 4
N	11	11	11	11
Mean of X	9.0	9.0	9.0	9.0
Mean of Y	7.5	7.5	7.5	7.5
Intercept	3.0	3.0	3.0	3.0
Slope	0.5	0.5	0.5	0.5
Correlation	0.82	0.82	0.82	0.82

Conclusion: The four data sets look “equivalent”

An Example of EDA

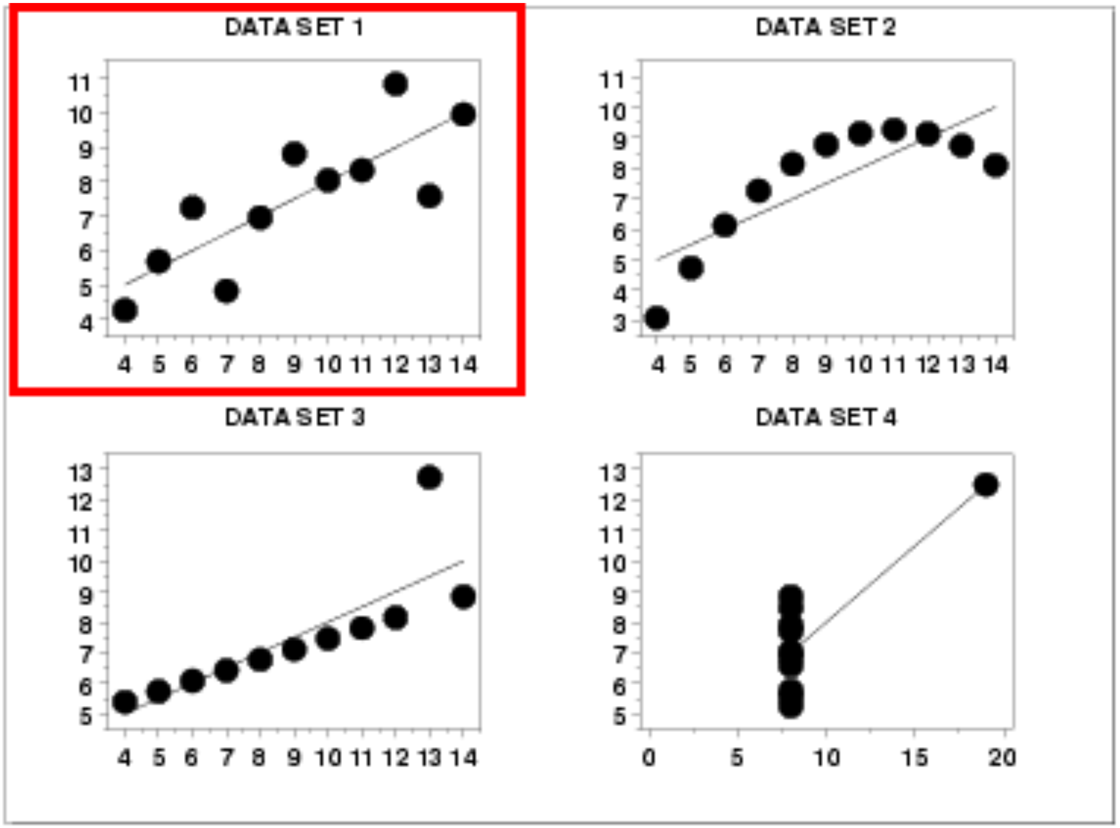
- Scatter plots:



Conclusion: The four data sets do not look “equivalent”.

An Example of EDA

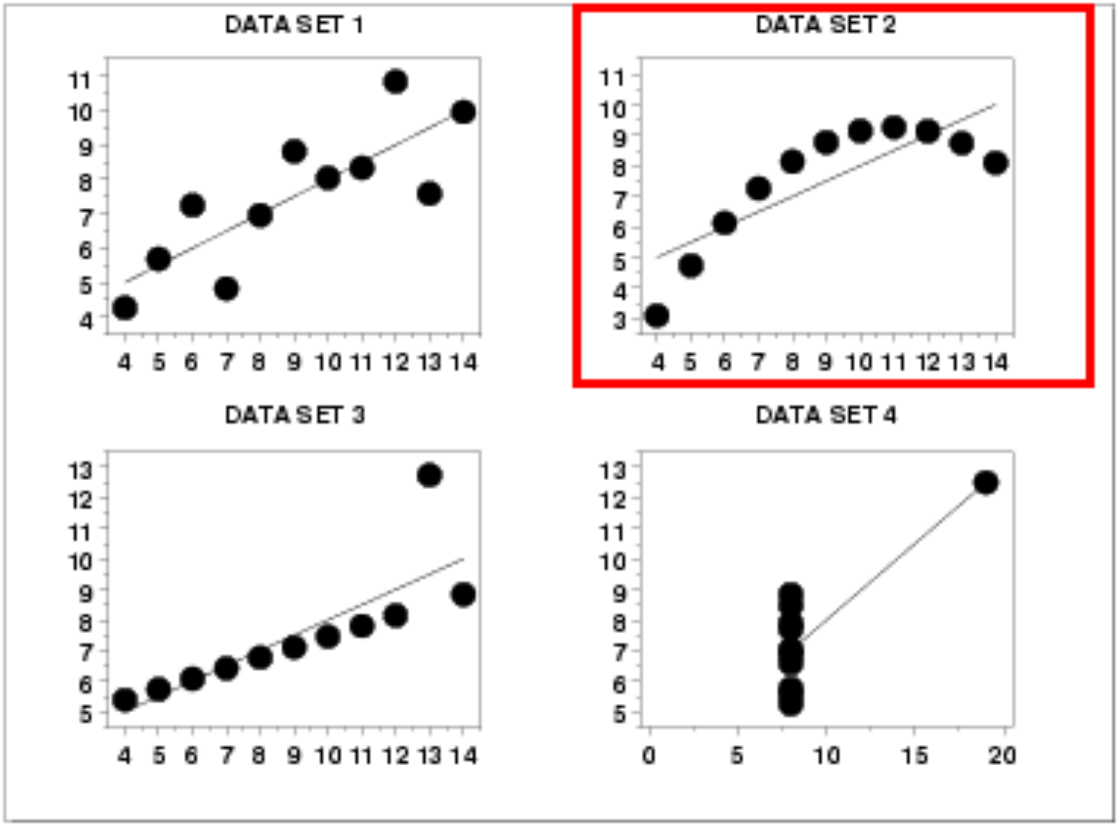
- Scatter plots:



Conclusion: The first data set is linear with some scatter

An Example of EDA

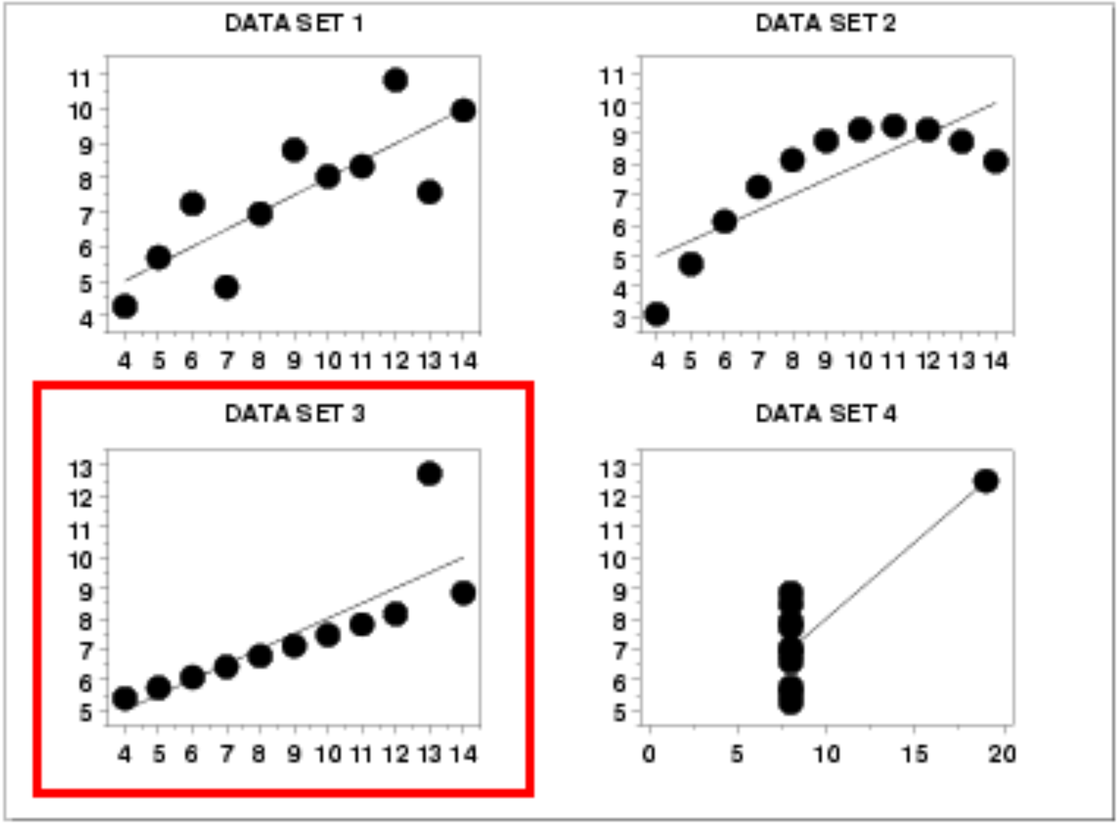
- Scatter plots:



Conclusion: The second data set is non-linear

An Example of EDA

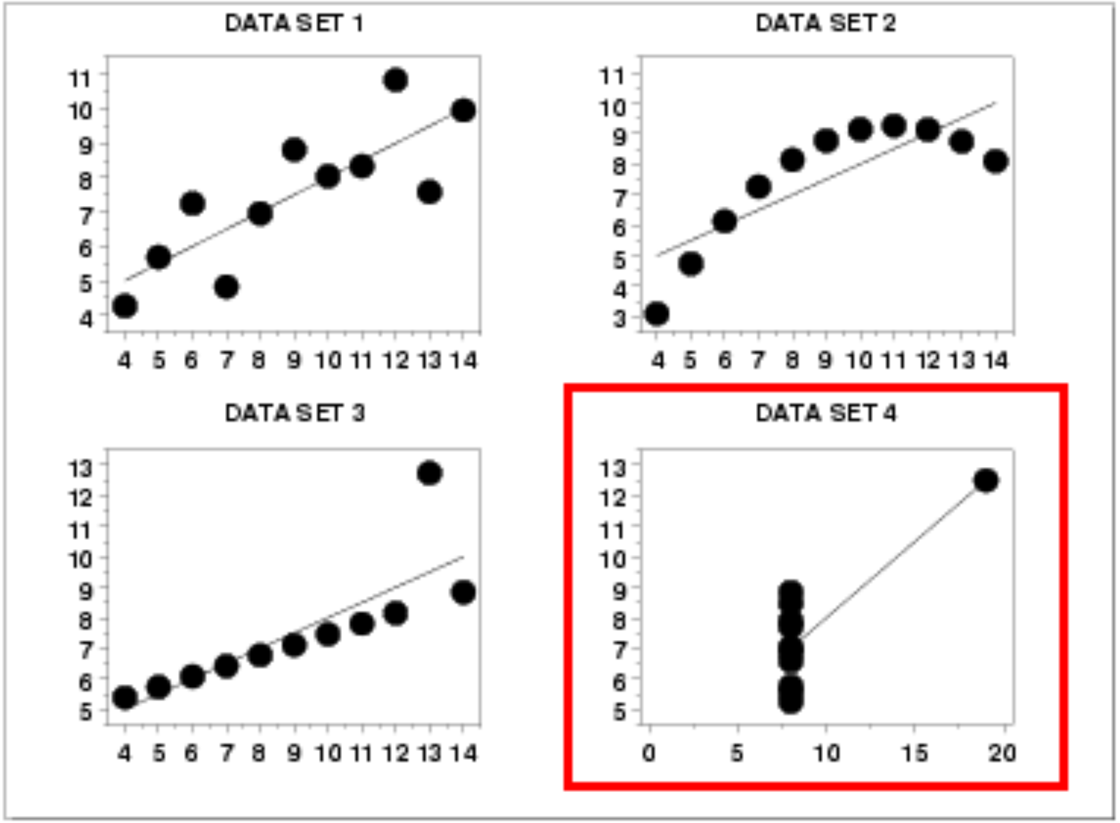
- Scatter plots:



Conclusion: The third data set has an outlier

An Example of EDA

- Scatter plots:



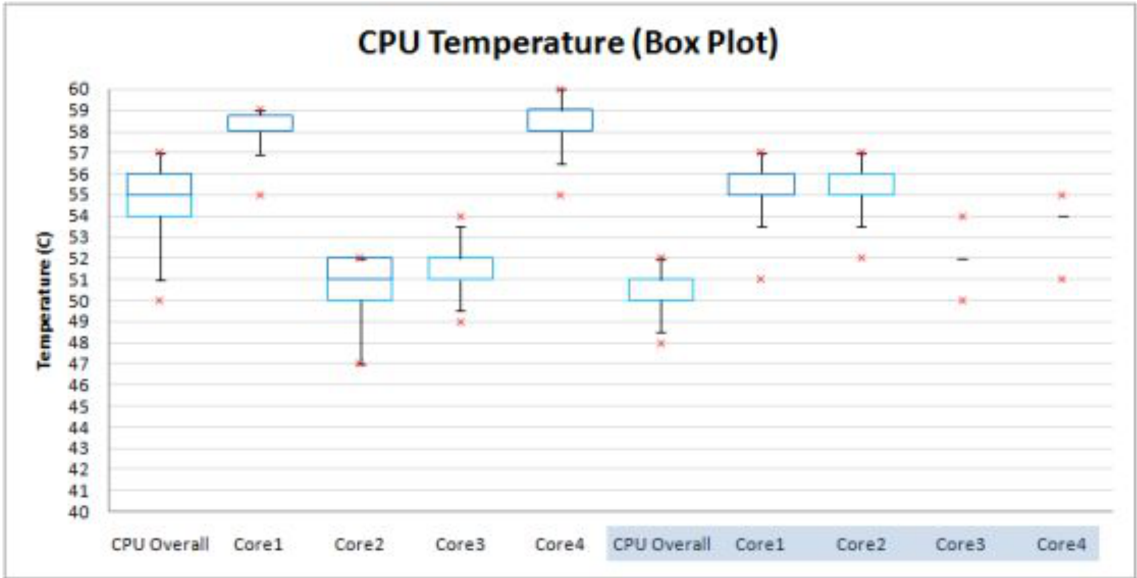
Conclusion: Bad experimental design in the fourth data set

An Example of EDA

- What to learn from the example?
  - Quantitative statistics are numeric summaries of the data
  - They focus on particular aspects of the data (location, correlation, etc.)
  - They are not wrong per se but depend of the underlying assumptions.
    Ex: linear regression cannot be applied to non-linear data (such as data set 2, 3 and 4).
  - EDA, through the use of scatter plots, gave further insight into the data.

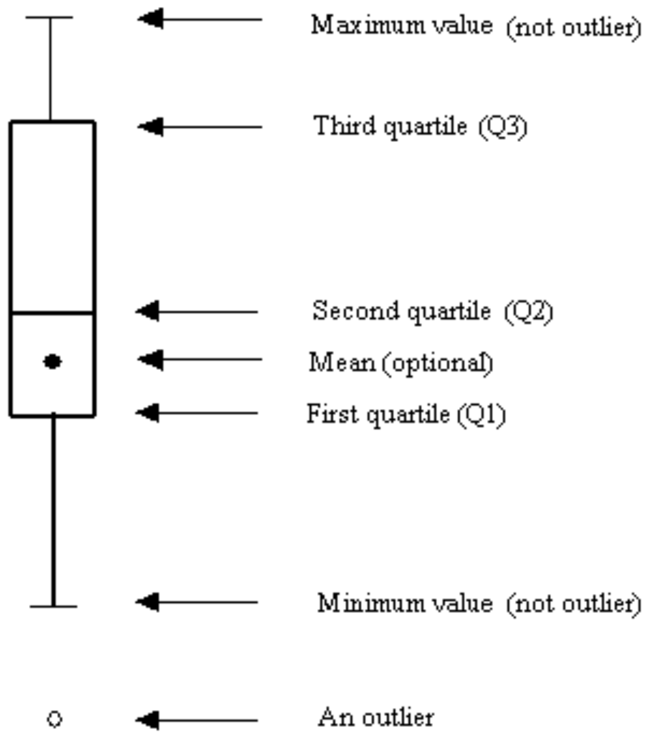
Graphical Techniques

- Find out the meaning!



Graphical Techniques

- Box Plot



- *Quartiles* split data into quarters
- *Interquartile range* is

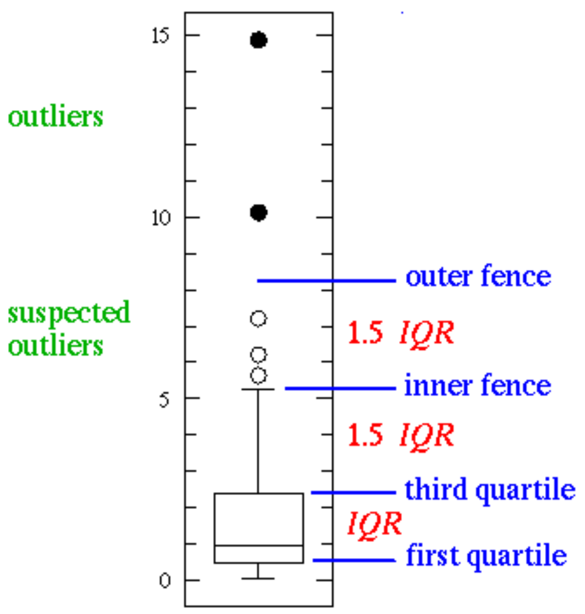
$$IQR = Q3 - Q1$$

- *Ouliers* are outside the interval

$$[Q1 - 1.5 \cdot IQR, Q3 + 1.5 \cdot IQR]$$

Graphical Techniques

- Box Plot



- *Mild outliers* are inside the intervals

$$[Q3 + 1.5 \cdot IQR, Q3 + 3 \cdot IQR]$$

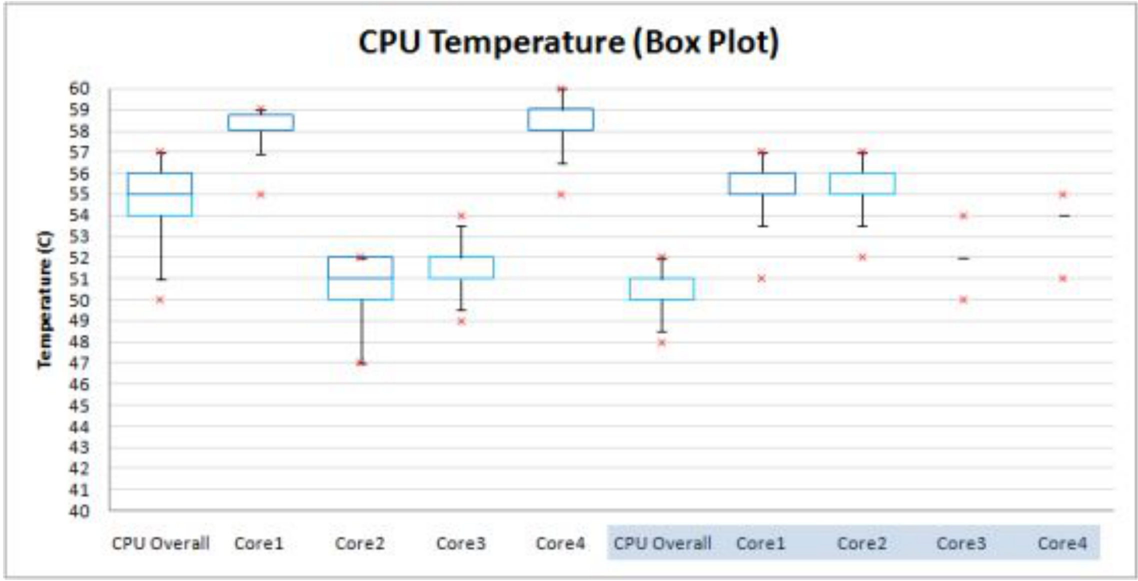
$$[Q1 - 3 \cdot IQR, Q1 - 1.5 \cdot IQR]$$

- *Extreme outliers* are outside the interval

$$[Q1 - 3 \cdot IQR, Q3 + 3 \cdot IQR]$$

Graphical Techniques

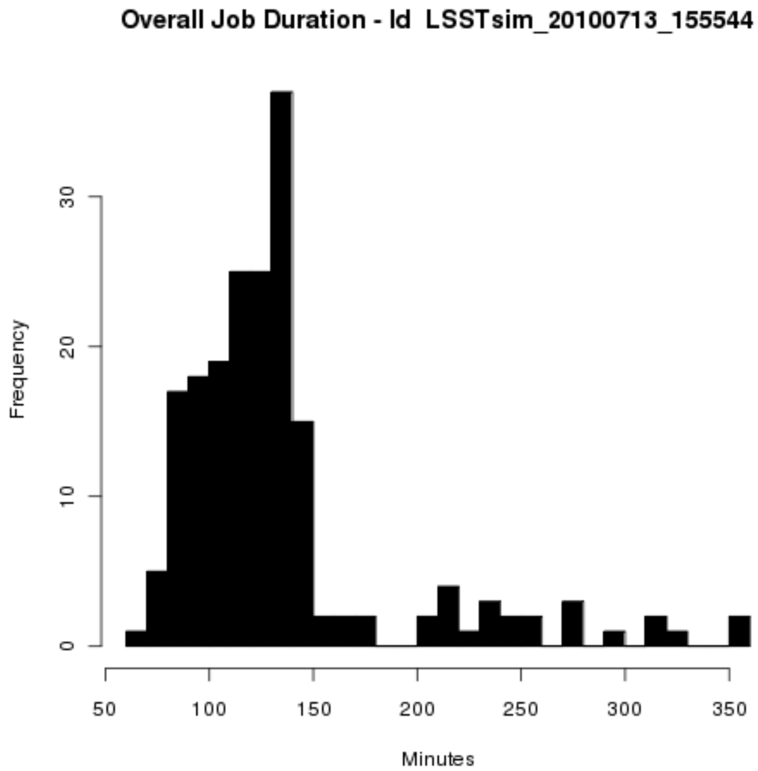
- Box Plot



- Does the location differ between groups?
- Does the variation differ between groups?
- Are there any outliers?

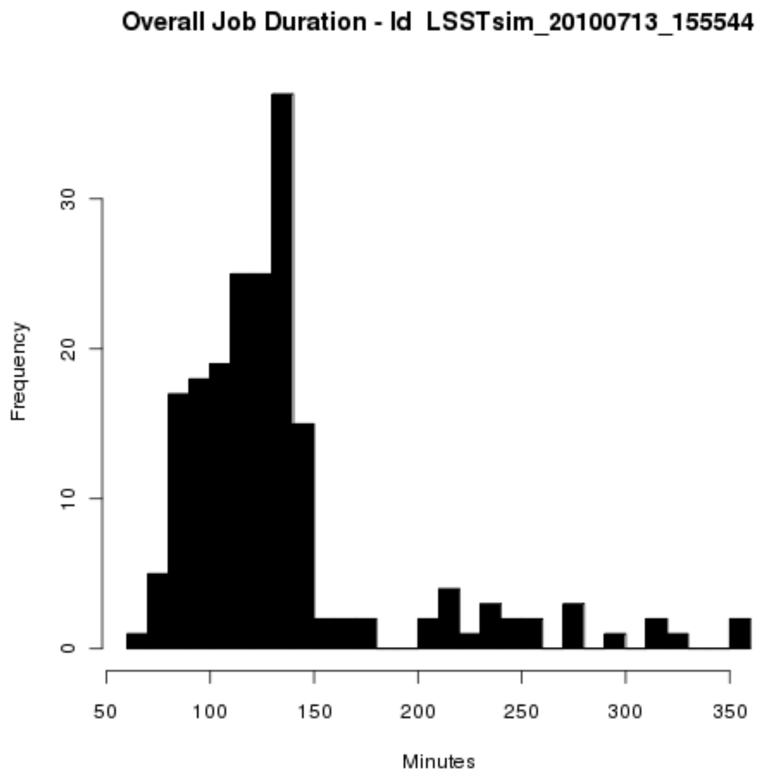
Graphical Techniques

- Find out the meaning!



Graphical Techniques

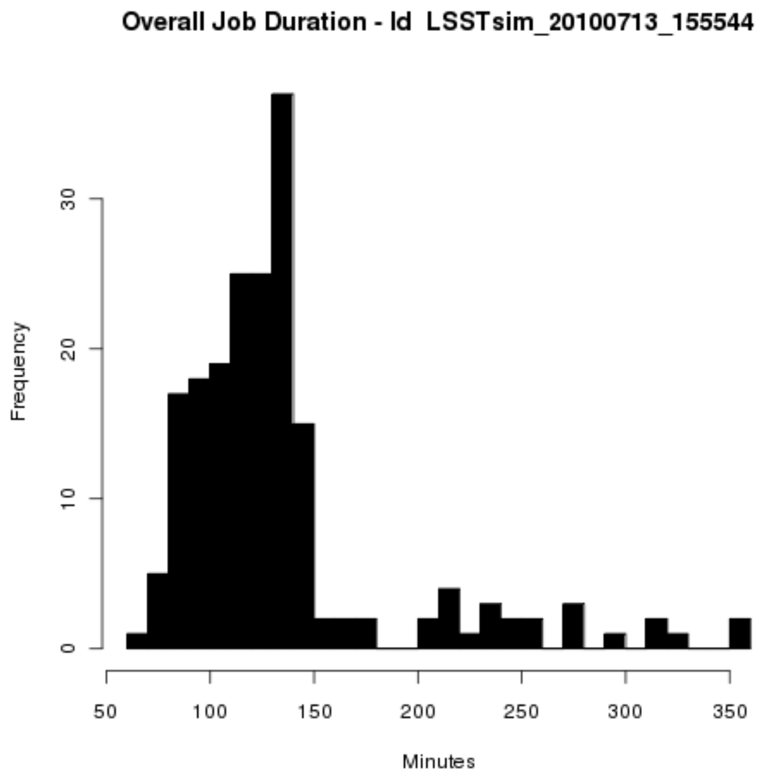
- Histogram



- Divide range into equal-sized bins
- Count the number of values that fall into each bin
- Divide the count in each bin by the total number of observations (optional)

Graphical Techniques

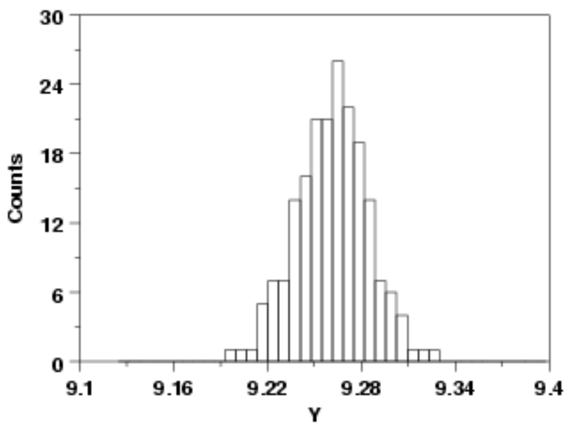
- Histogram

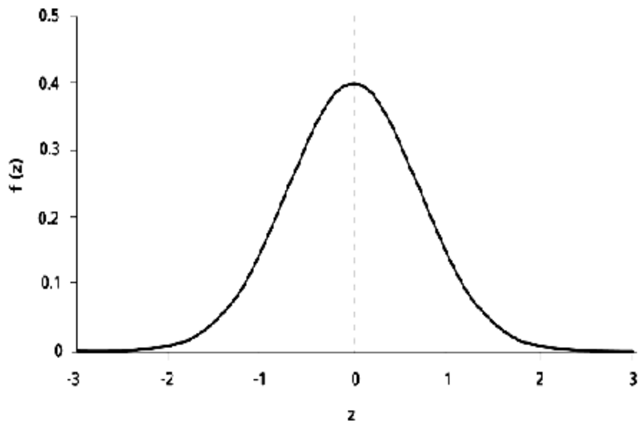


- What kind of distribution do the data come from?
- Where are the data located?
- How spread out are the data?
- Are the data symmetric or skewed?
- Are there outliers in the data?

Graphical Techniques

- Histogram

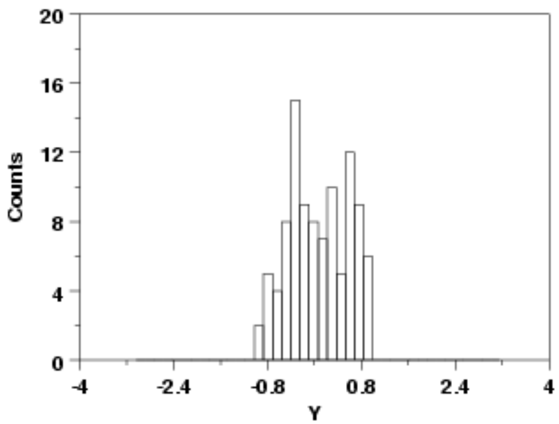


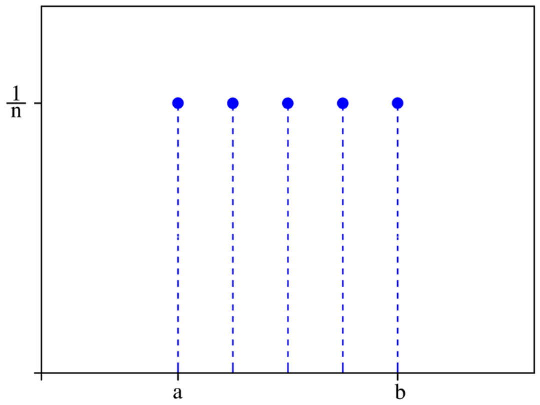


- Symmetric, moderated-tailed histogram (top plot)
- It suggests a normal distribution (bottom plot)
- Next step: Check normality

Graphical Techniques

- Histogram

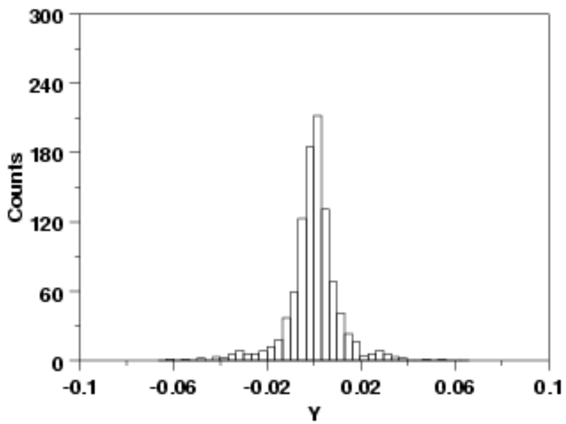


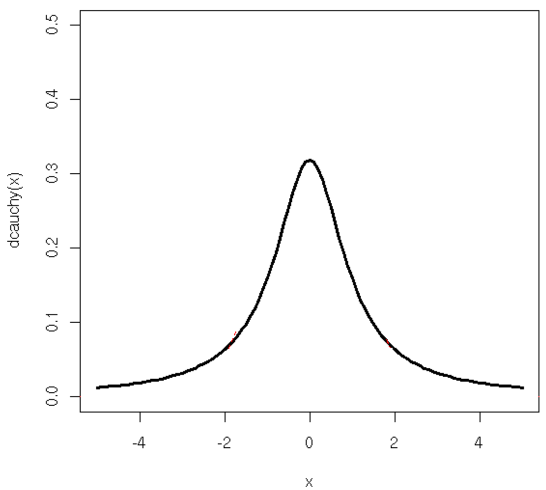


- Symmetric, short-tailed histogram (top plot)
- It suggests an uniform distribution (bottom plot)
- Next step: Check if the data follows an uniform distribution

Graphical Techniques

- Histogram

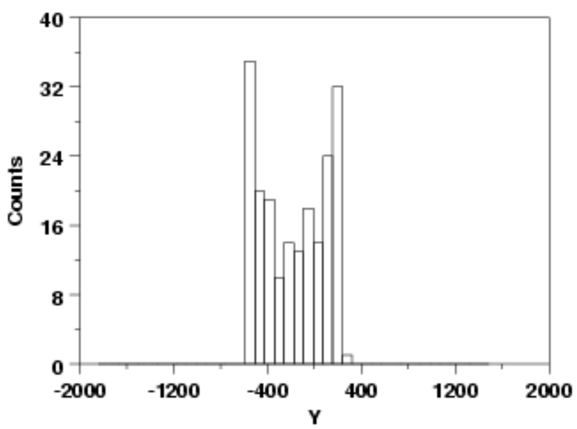


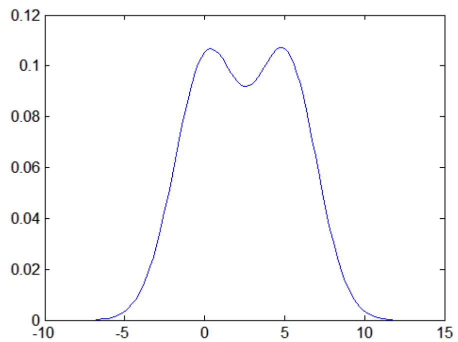


- Symmetric, long-tailed histogram (top plot)
- It suggests an Cauchy distribution (bottom plot)
- Next step: Check if the data follows a Cauchy distribution

Graphical Techniques

- Histogram

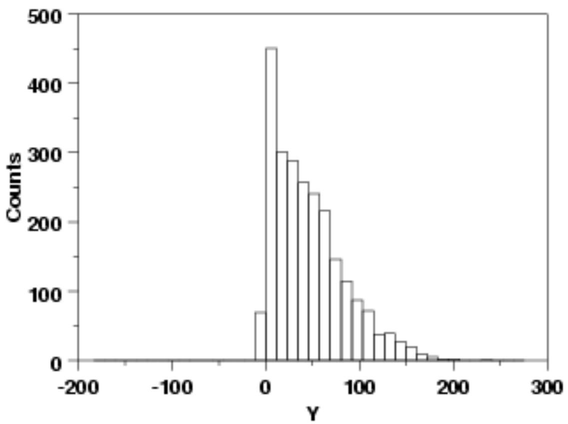


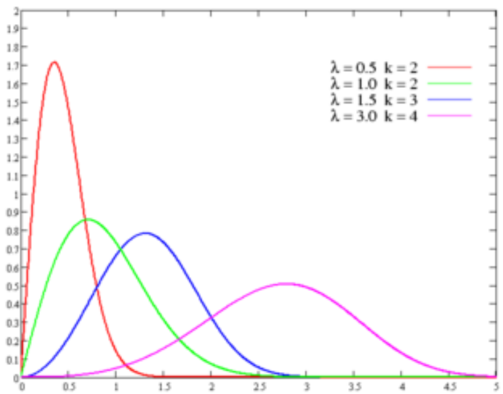


- Symmetric, bimodal histogram (top plot)
- It suggests a mixture of two distributions (bottom plot)

Graphical Techniques

- Histogram

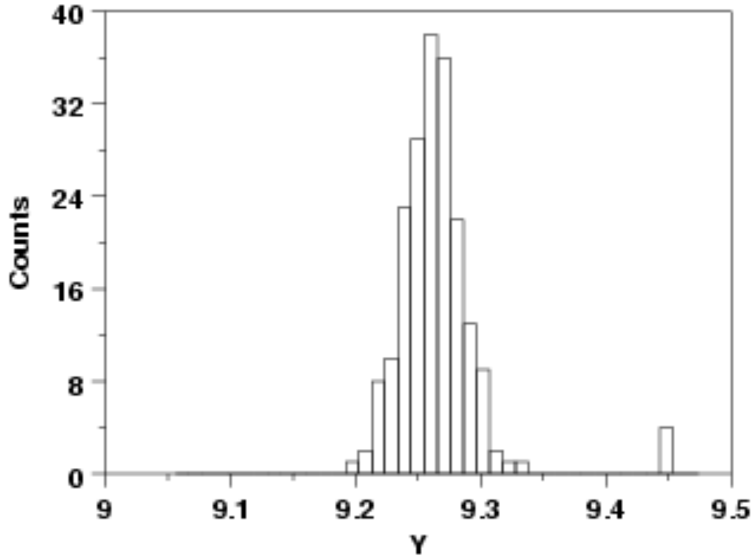




- Right-skewed histogram (top plot)
- It suggests an extreme-value distribution, e.g. Weibull (bottom plot)
- Same reasoning applies to left-skewed data

Graphical Techniques

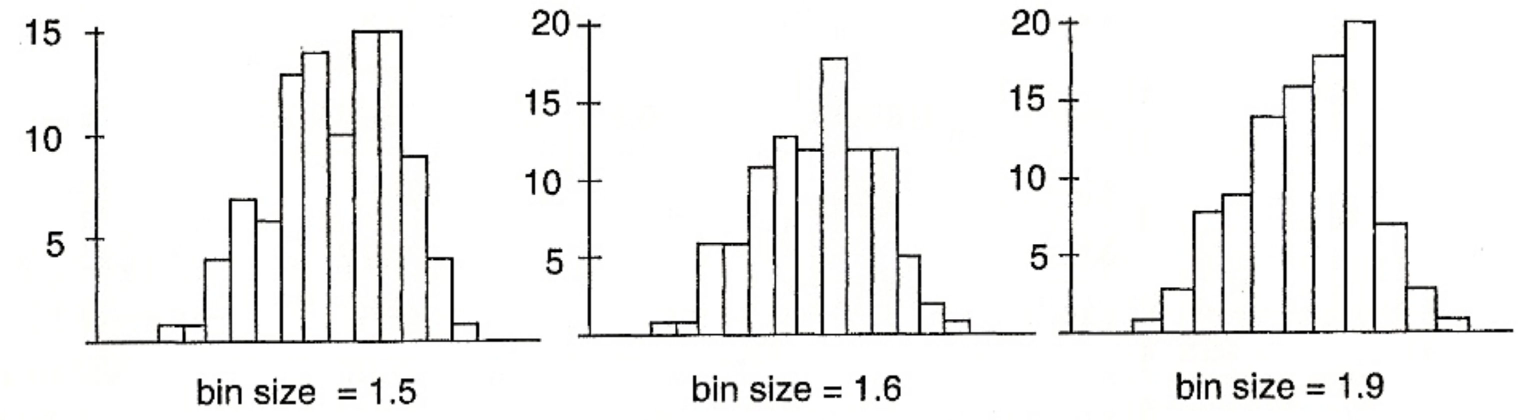
- Histogram



Data with outliers (in the right)

Graphical Techniques

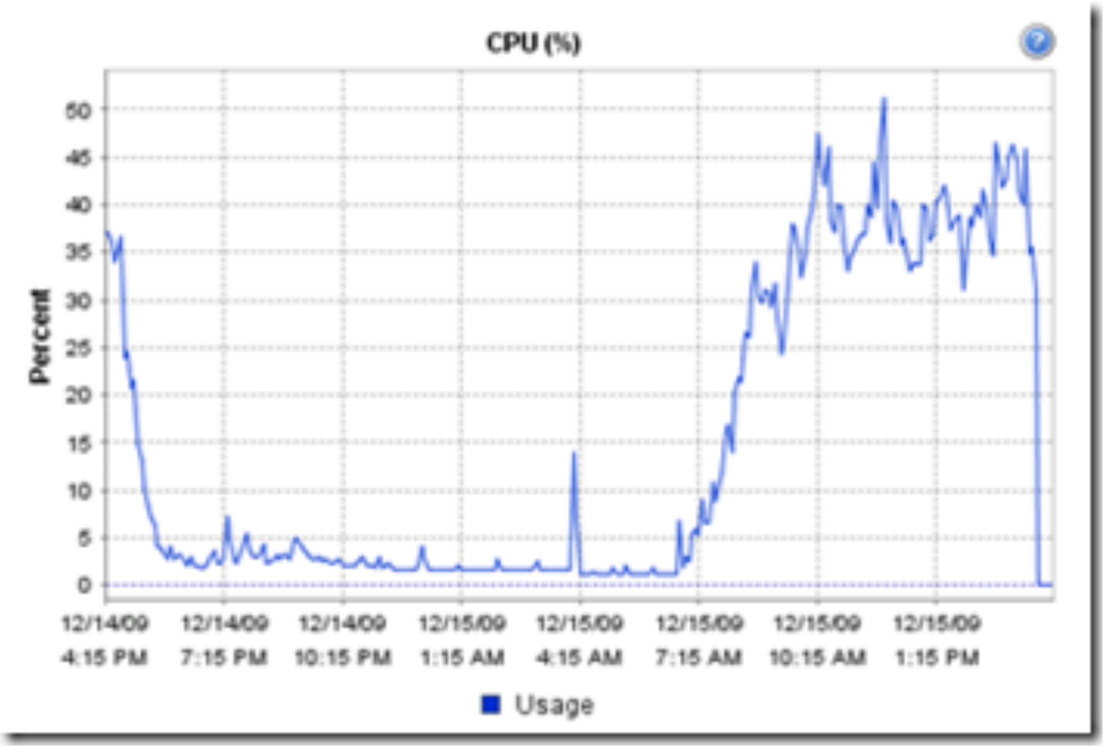
- Histogram



The size of the bins may affect the shape of the distribution

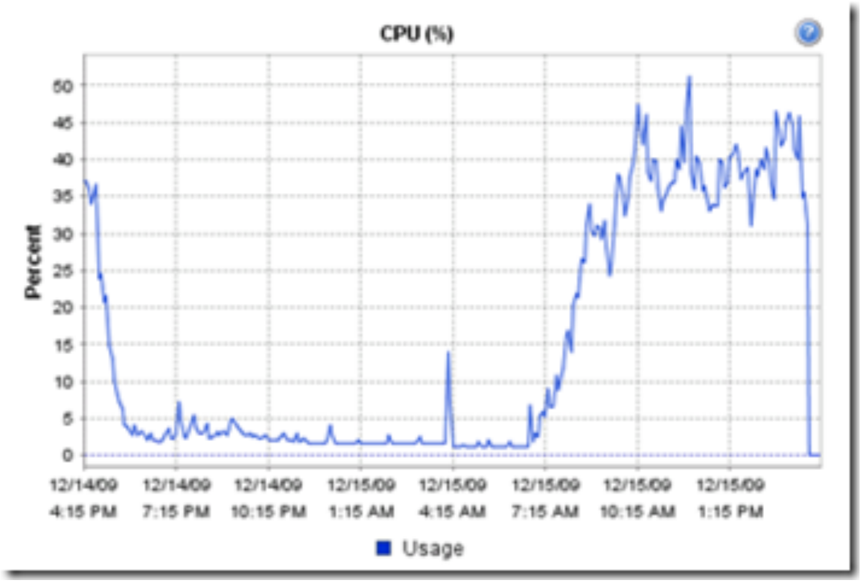
Graphical Techniques

- Find out the meaning!



Graphical Techniques

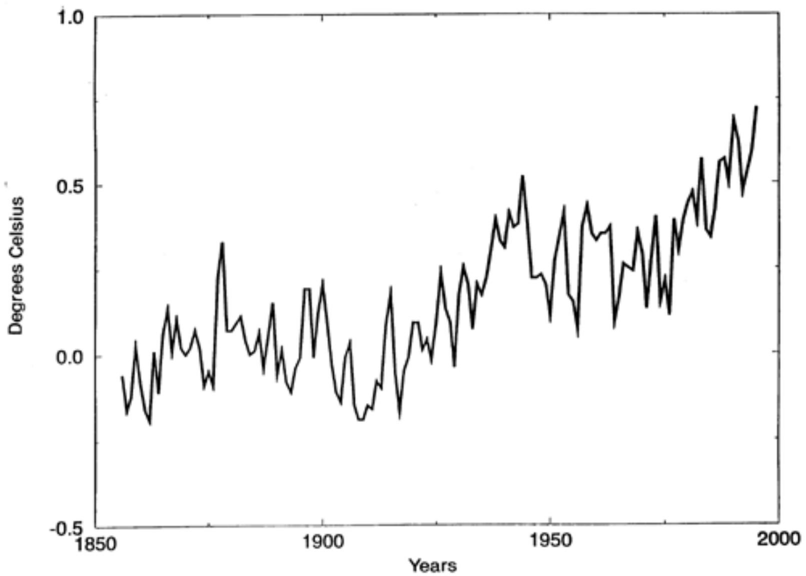
- Run Chart



- Used for time series representation
- Are there any shifts in location?
- Are there any shifts in variation?
- Are there any outliers?

Graphical Techniques

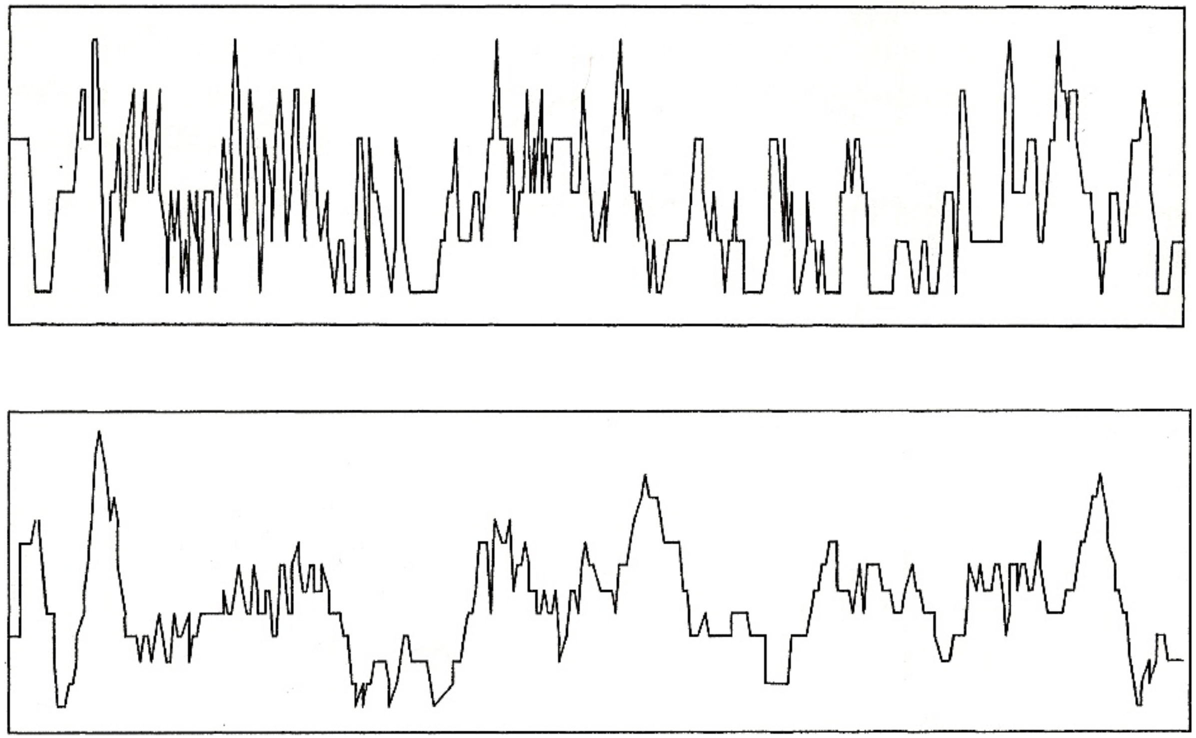
- Run Chart



- Used for time series representation
- Are there any shifts in location?
- Are there any shifts in variation?
- Are there any outliers?

Graphical Techniques

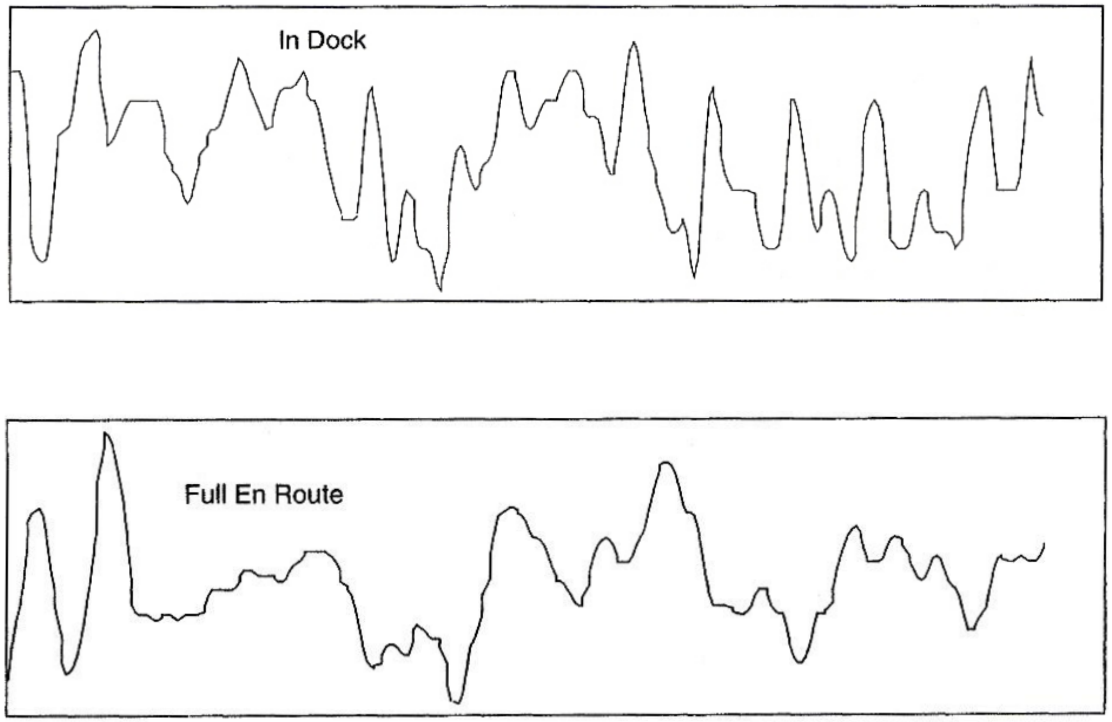
- Run Chart Example (Cohen 1995)



Run chart for the number of ships in dock (top) and full in route (bottom) over time.

Graphical Techniques

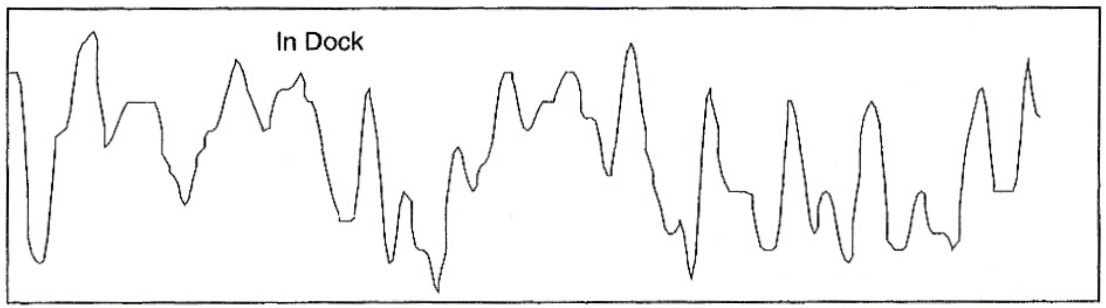
- Run Chart Example

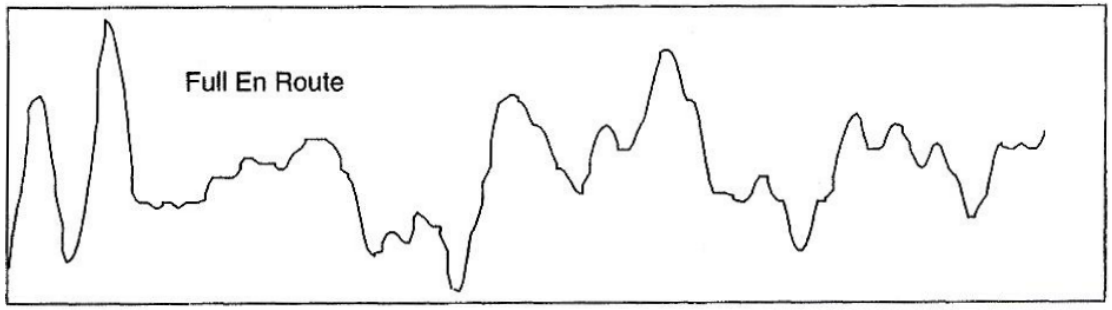


Time series smoothing (e.g. averaging values in a time window)

Graphical Techniques

- Run Chart Example

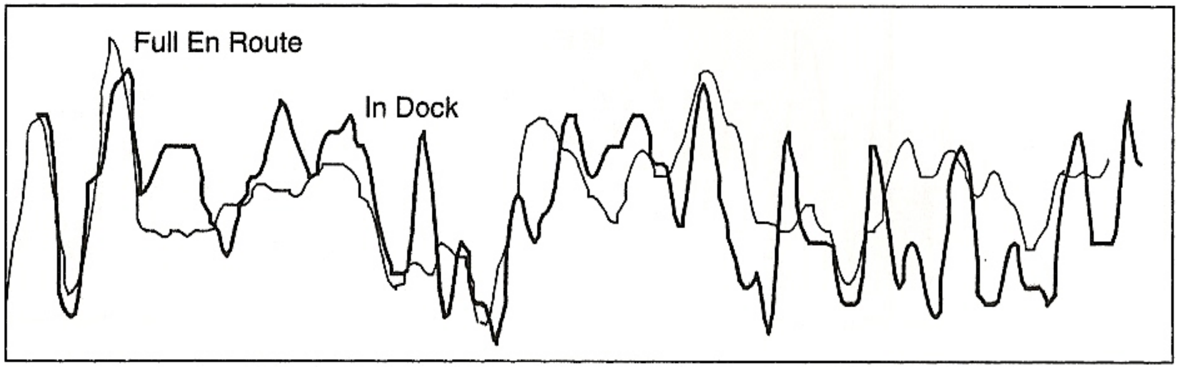




Time series shifting

Graphical Techniques

- Run Chart Example

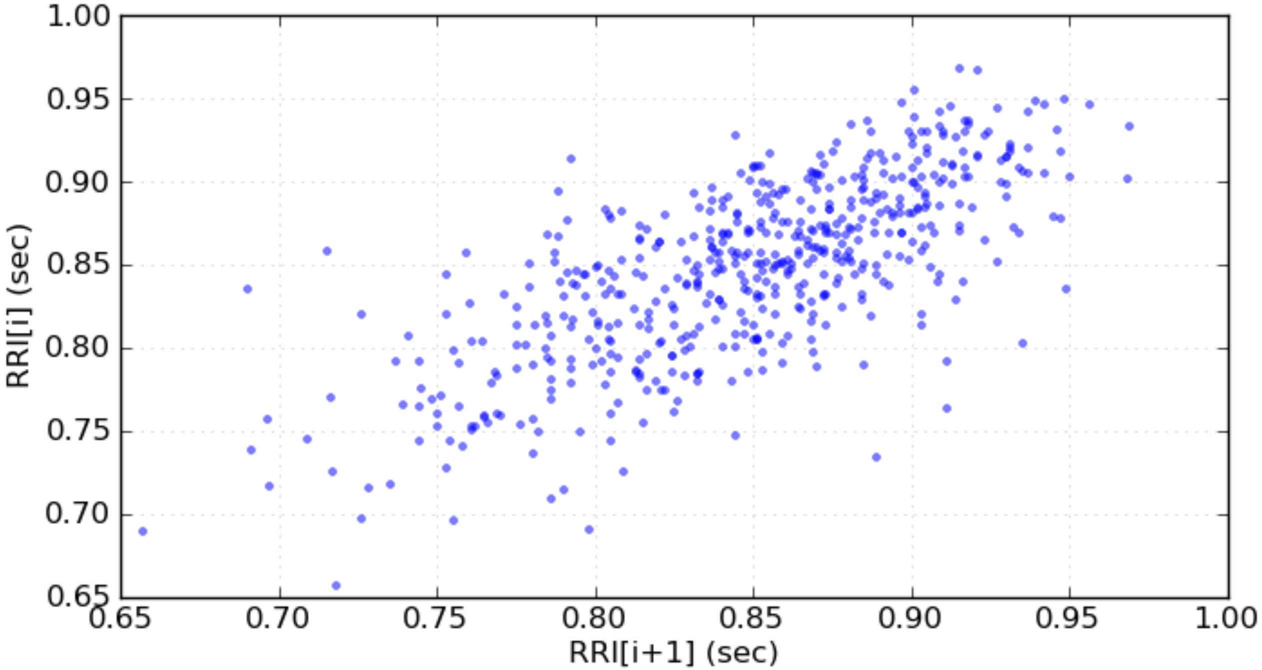


Time series overlapping: time series in dock antecipates the full en route series quite well.

The number of ships in dock predicts the number of ships in route in a few days.

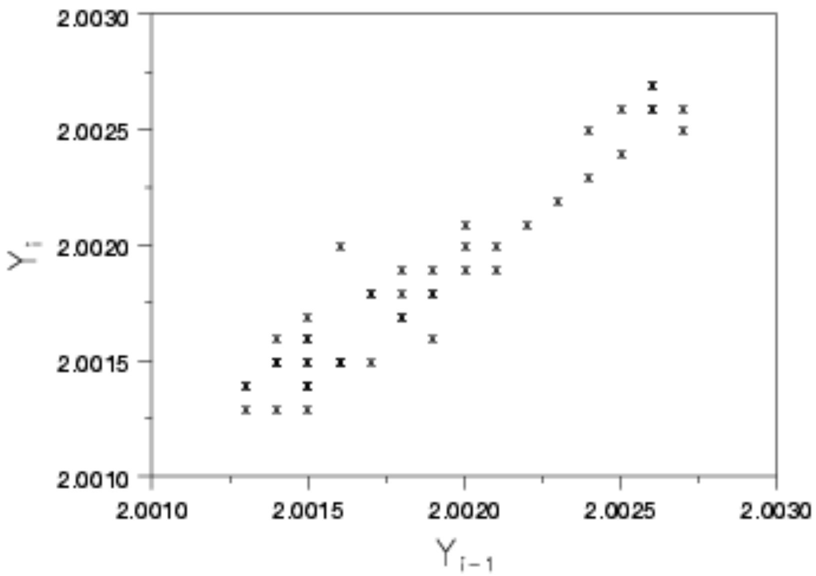
Graphical Techniques

- Find out the meaning!



Graphical Techniques

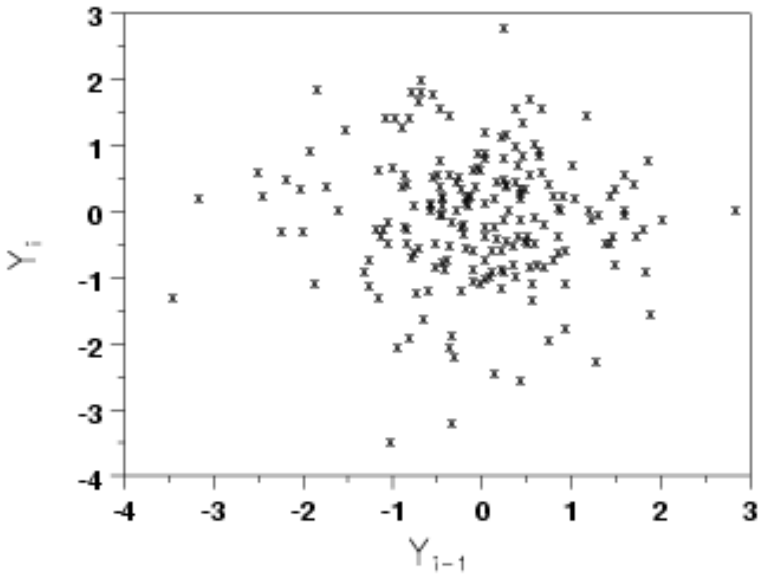
- Lag plot



- Indicates the randomness of a time series
- Are the data random?
- Is there serial correlation in the data?
- Are there outliers in the data?

Graphical Techniques

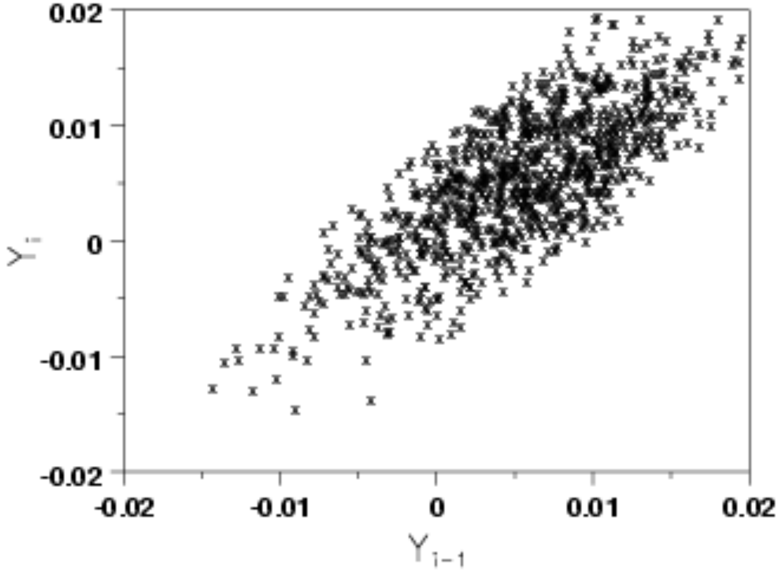
- Lag plot



- The data are random
- No correlation
- No outliers

Graphical Techniques

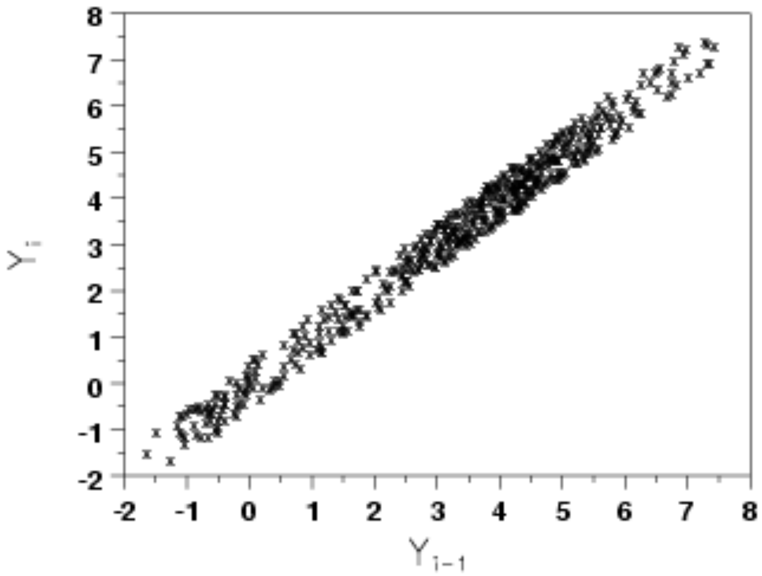
- Lag plot



- Weak positive serial correlation
- No outliers
- Suggest autoregressive model

Graphical Techniques

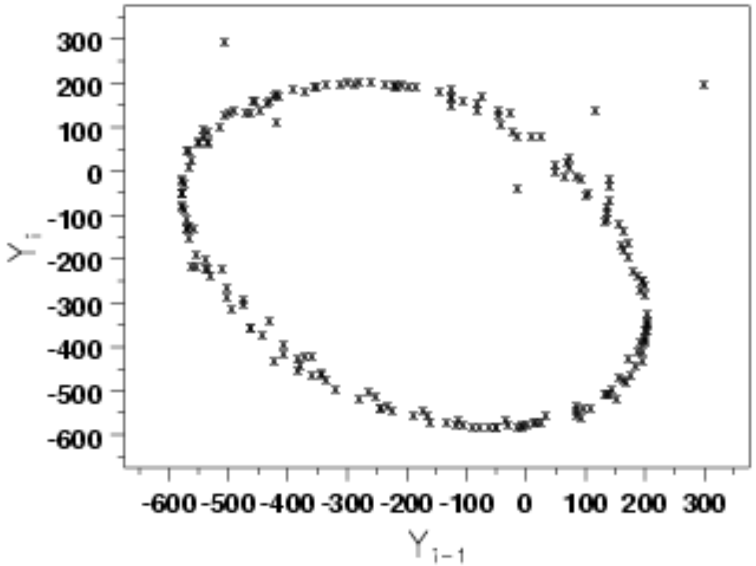
- Lag plot



- Strong positive serial correlation
- No outliers
- Suggest autoregressive model

Graphical Techniques

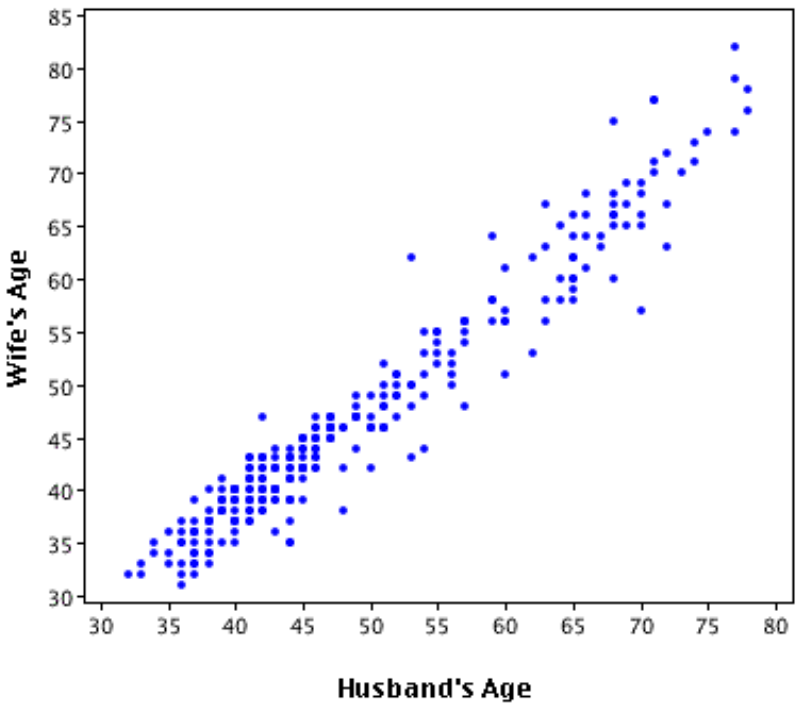
- Lag plot



- Sinusoidal model
- Presence of outliers

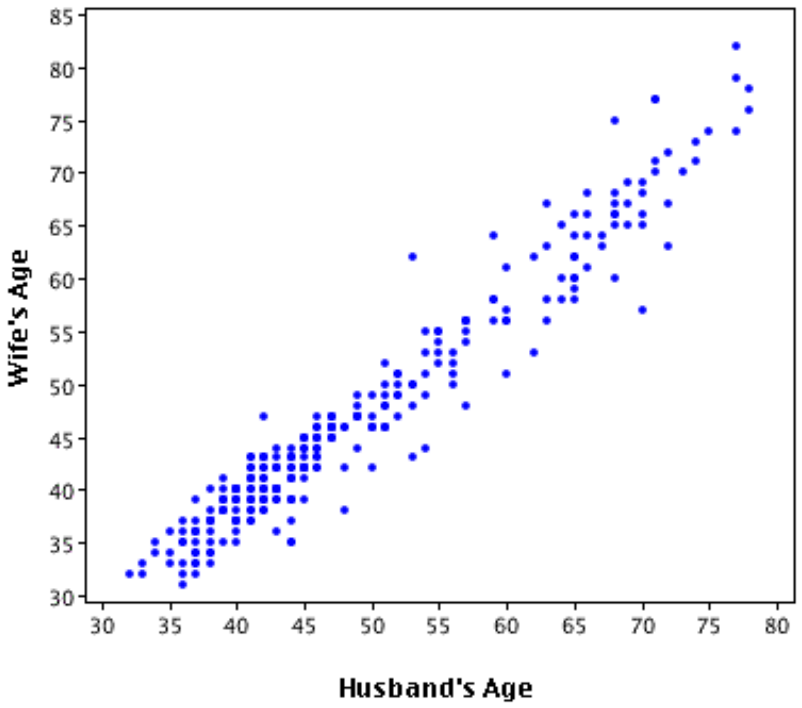
Graphical Techniques

- Find out the meaning!



Graphical Techniques

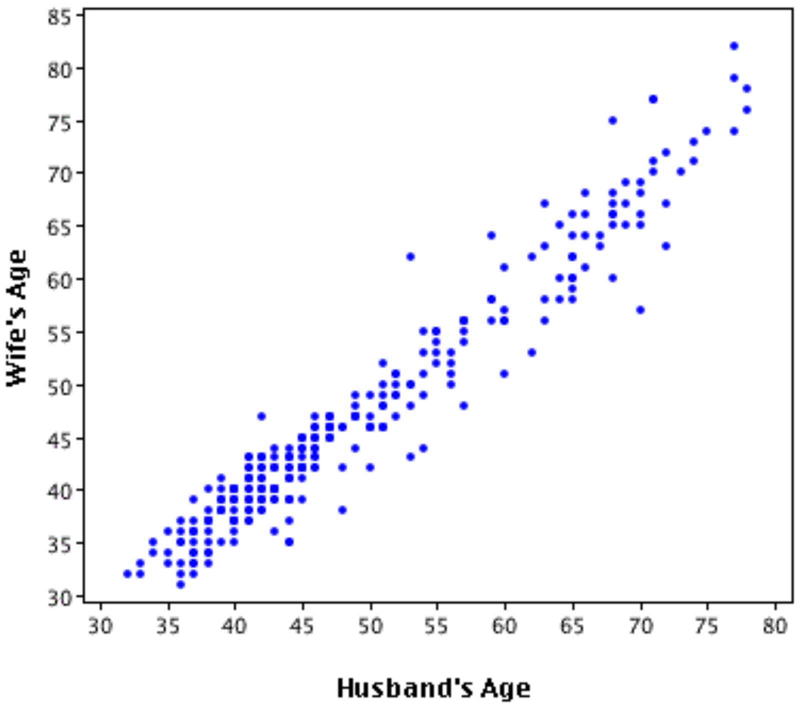
- Scatter plot



- Plot of the values of Y versus the corresponding values of X.
- Useful to reveal relationships or association between two or more variables.

Graphical Techniques

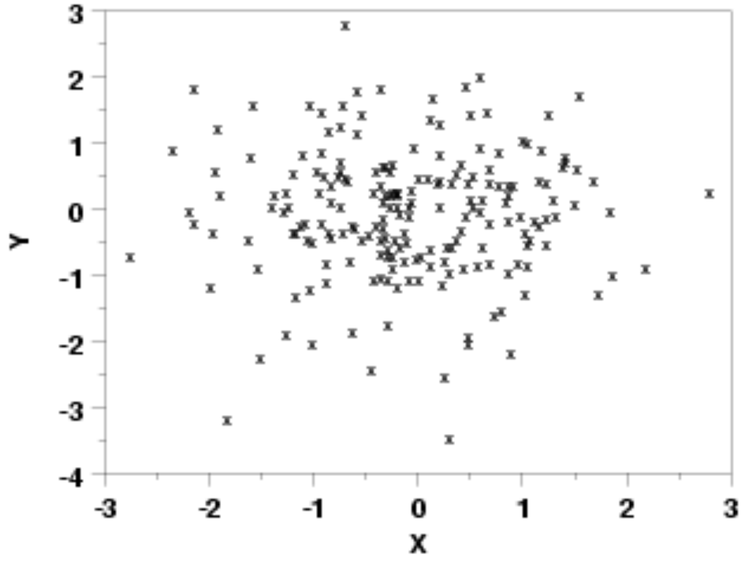
- Scatter plot



- Are variables X and Y related?
- Are variables X and Y linearly related?
- Are variables X and Y non-linearly related?
- Are there outliers?

Graphical Techniques

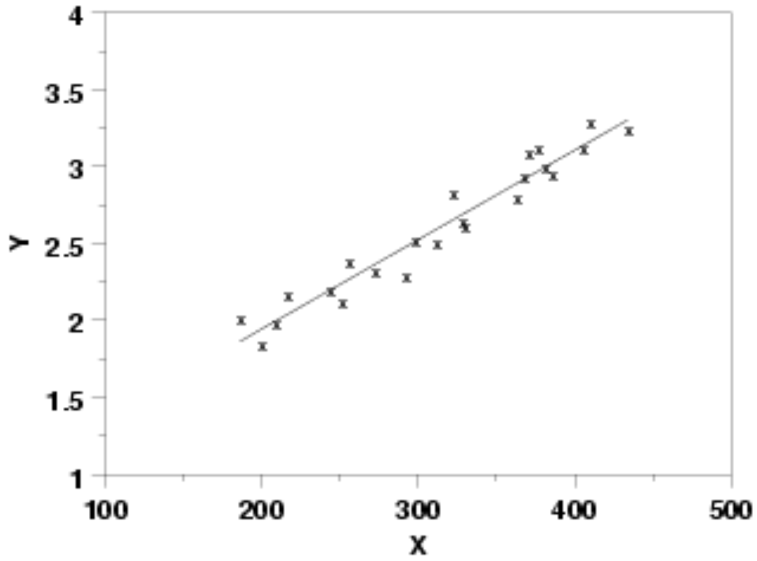
- Scatter Plot



No relationship

Graphical Techniques

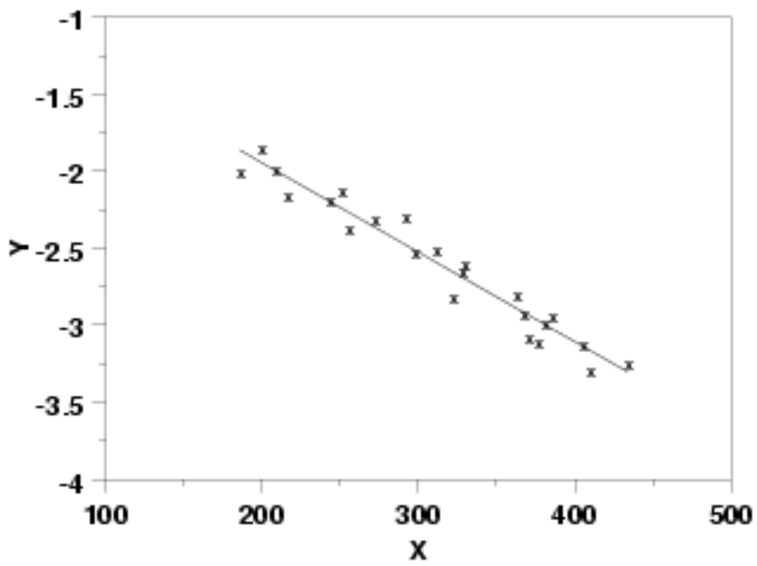
- Scatter Plot



Strong linear relationship
(positive correlation)

Graphical Techniques

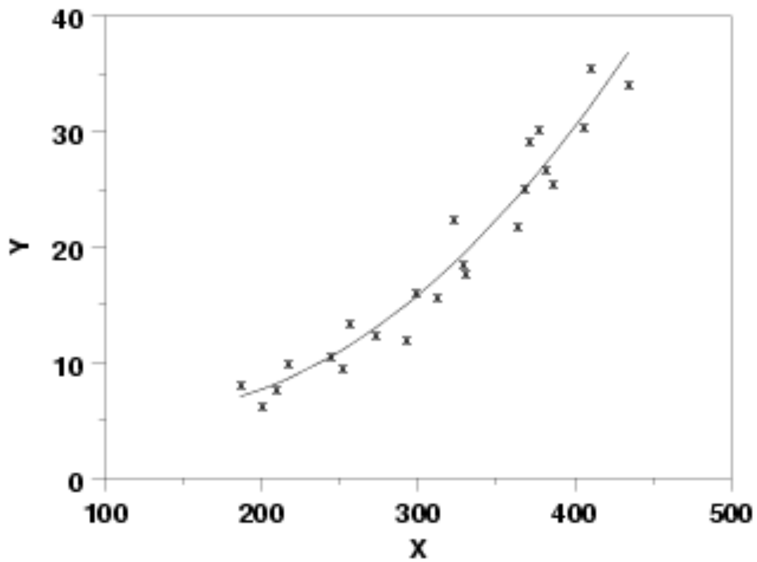
- Scatter Plot



Strong linear relationship
(negative correlation)

Graphical Techniques

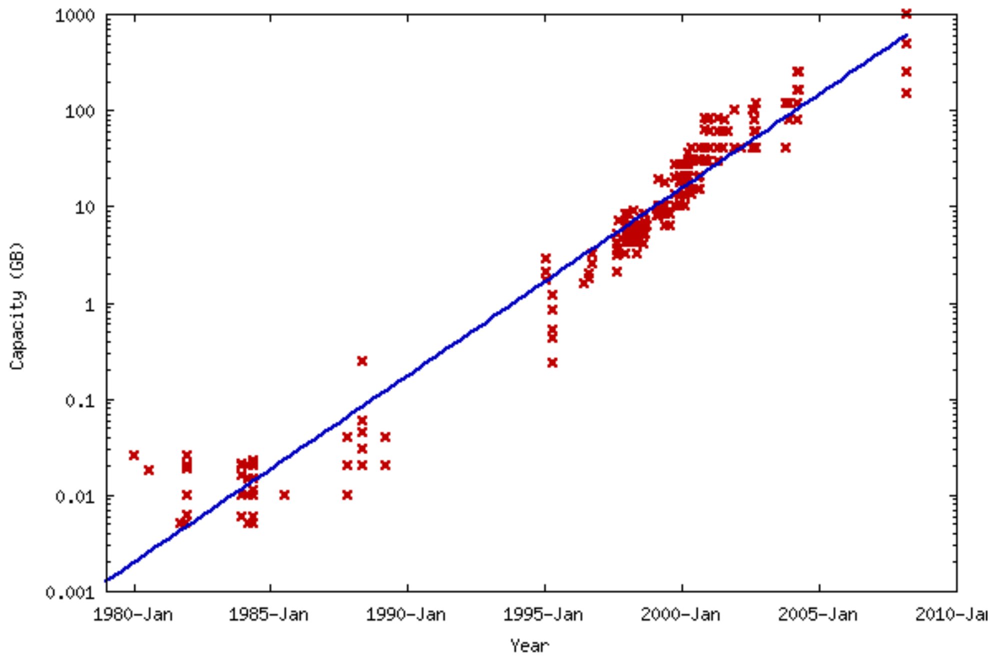
- Scatter Plot



Quadratic relationship

Graphical Techniques

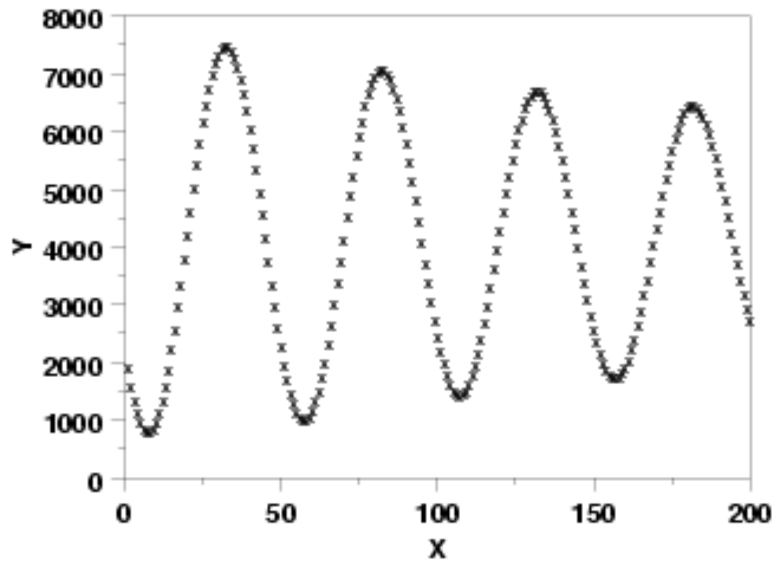
- Scatter Plot



Exponential relationship
(see logarithm scale in the y axis)

Graphical Techniques

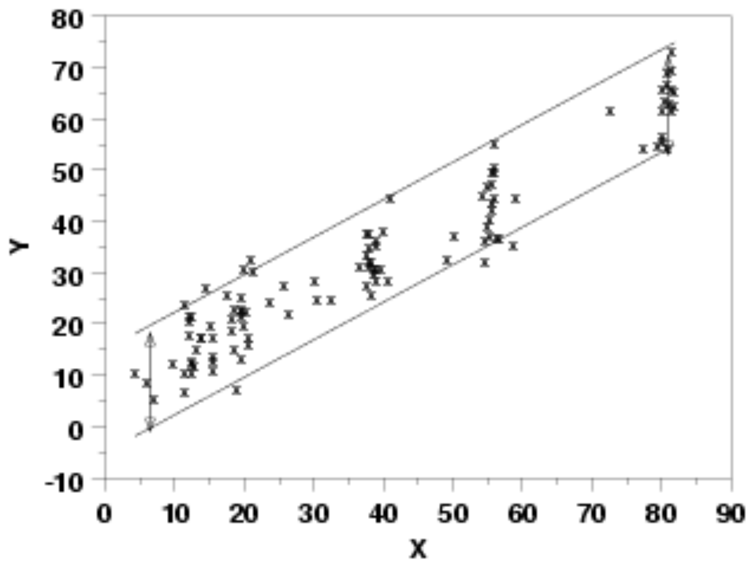
- Scatter Plot



Sinusoidal relationship

Graphical Techniques

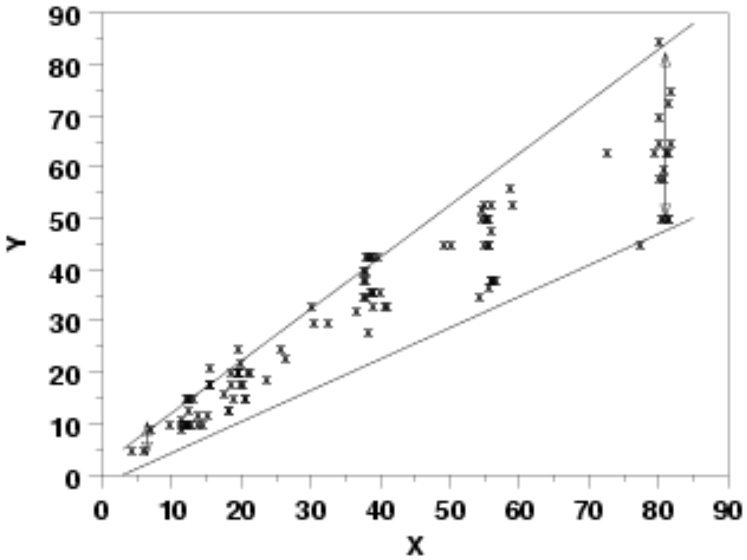
- Scatter Plot



Linear relationship with constant variation

Graphical Techniques

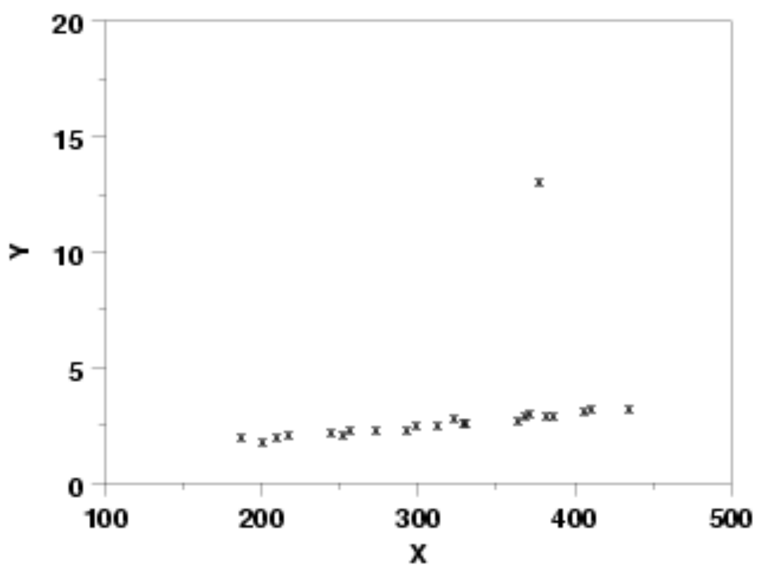
- Scatter Plot



Linear relationship with non-constant variation

Graphical Techniques

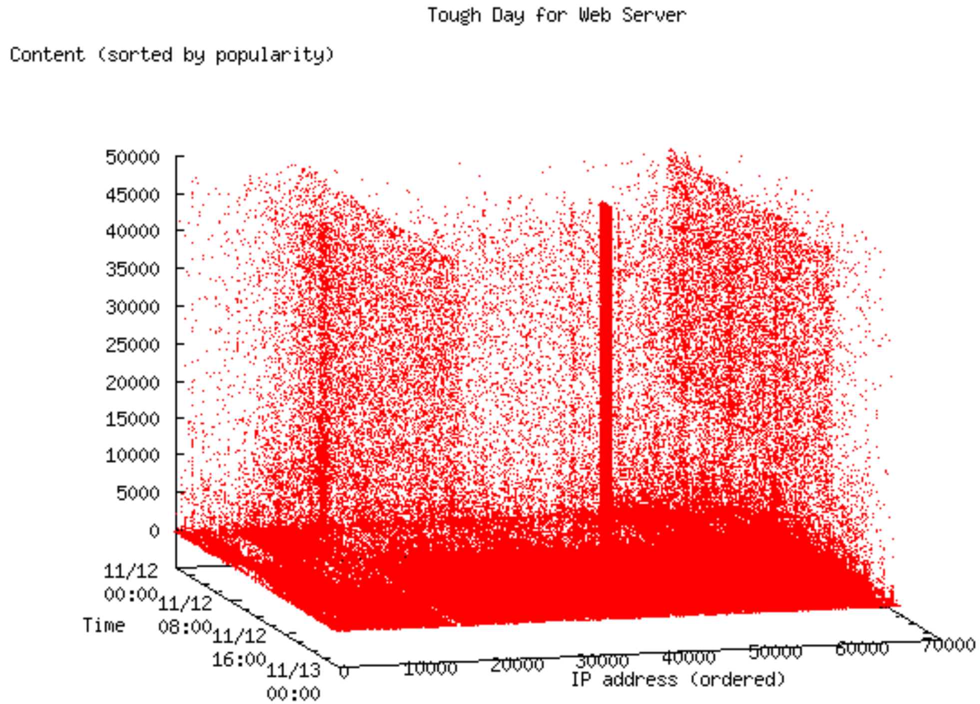
- Scatter Plot



Linear relationship with an outlier

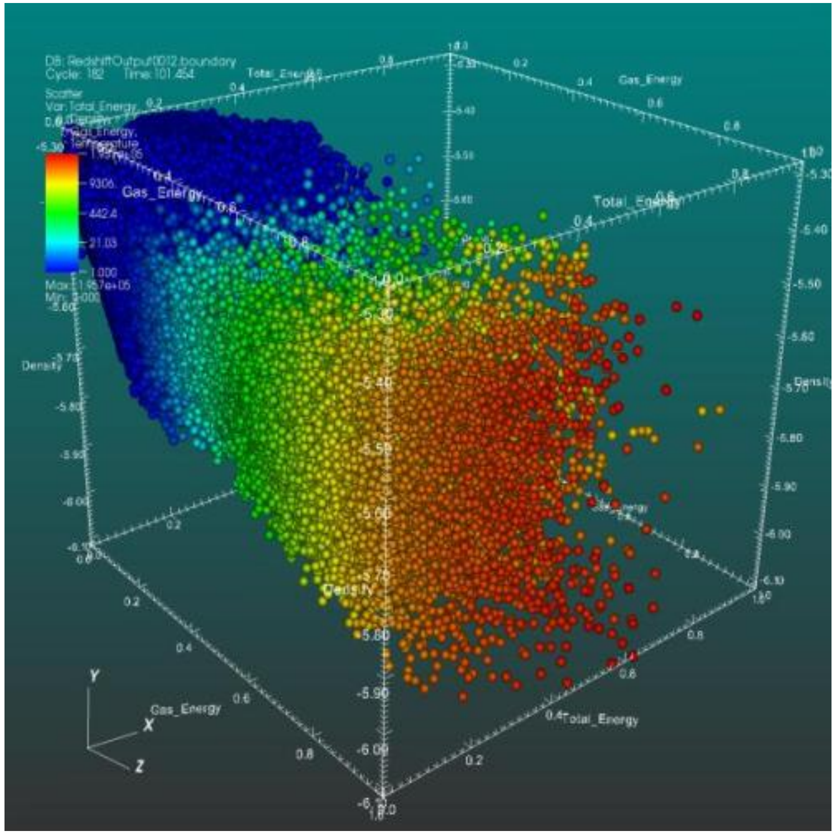
Graphical Techniques

- A 3D Scatter plot



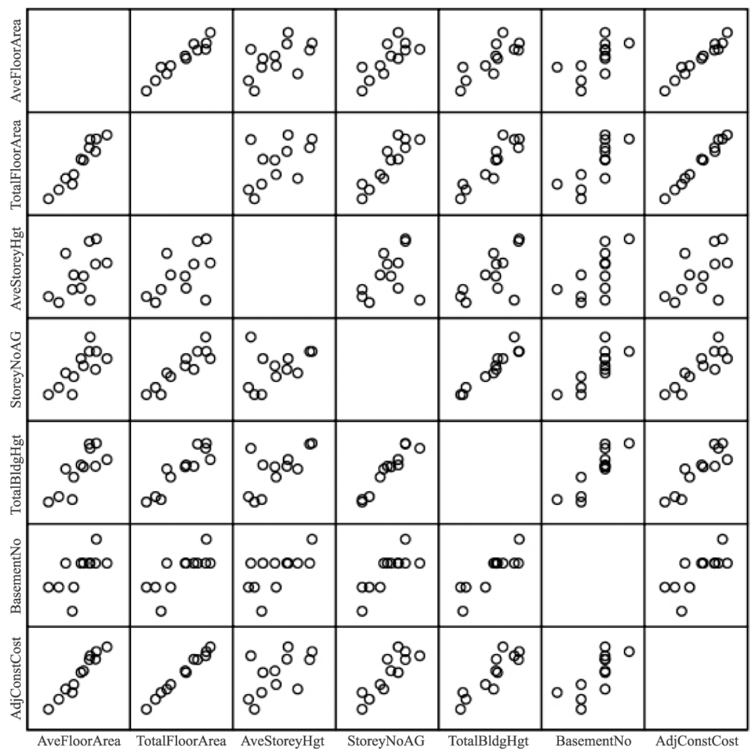
Graphical Techniques

- A 4D Scatter plot



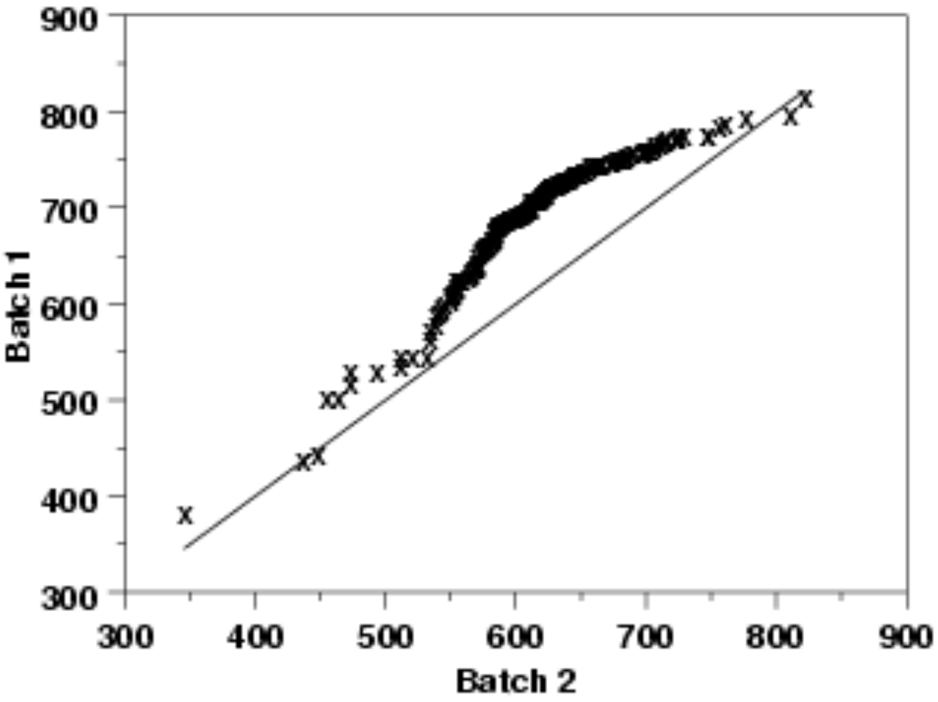
Graphical Techniques

- Scatter plot matrix



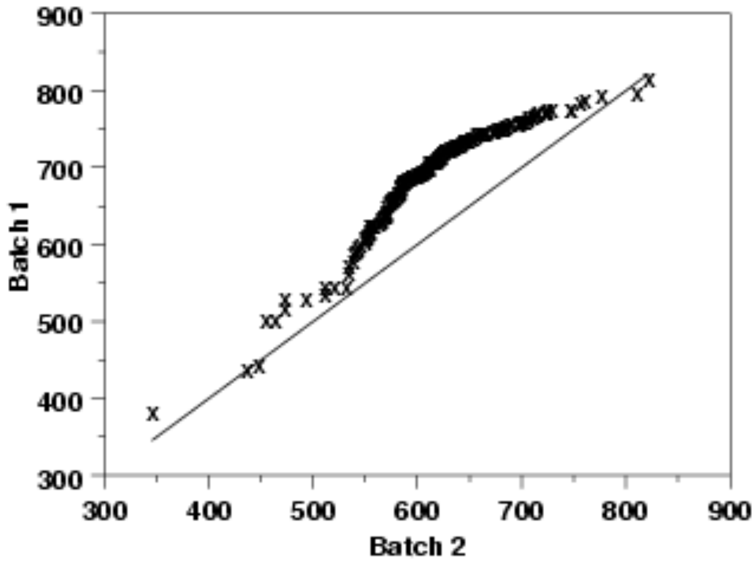
Graphical Techniques

- Find out the meaning!



Graphical Techniques

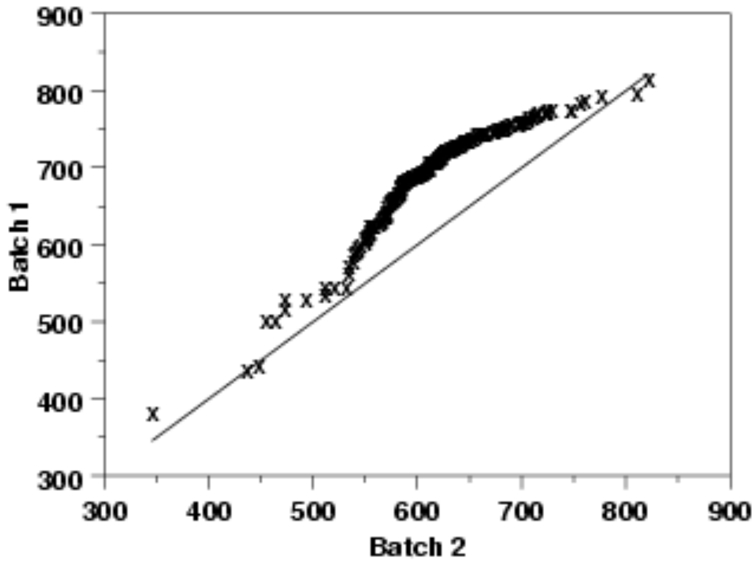
- Q-Q plot



- Plot quantiles of the first data set against the quantiles of the second data set
- 45-degree reference line is plotted for reference

Graphical Techniques

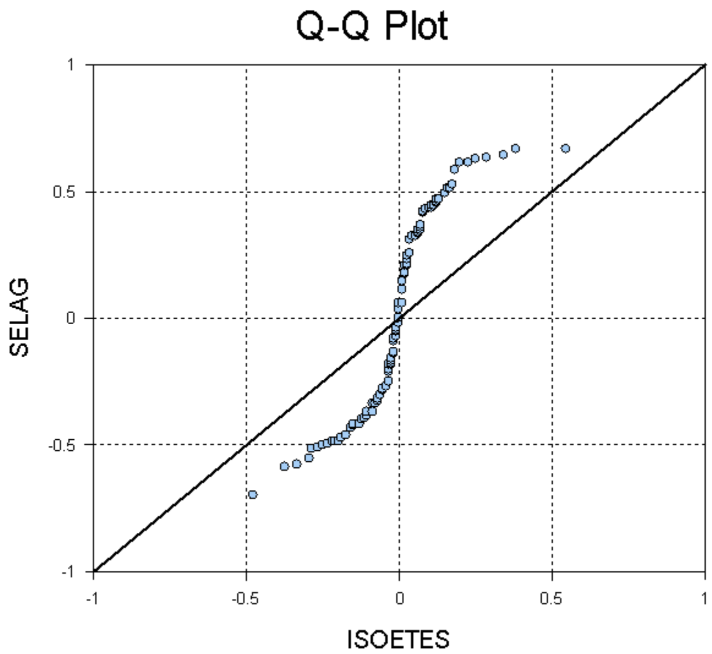
- Q-Q plot



- Do two data sets come from populations with a common distribution?
- In this case, no

Graphical Techniques

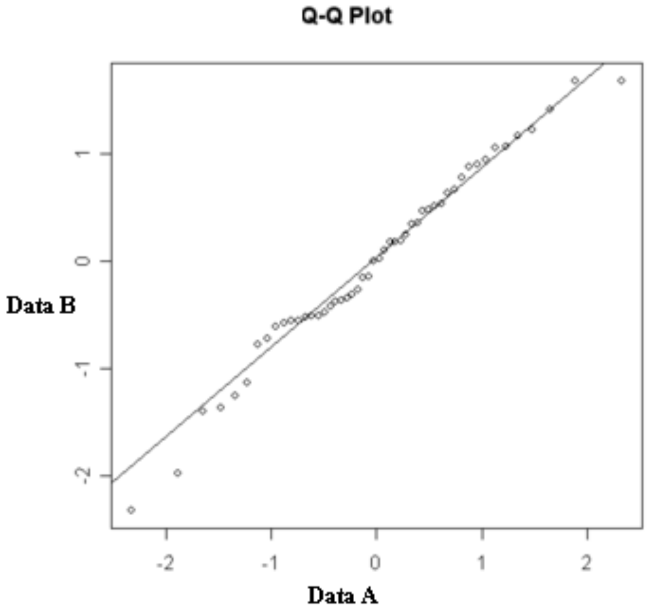
- Q-Q plot



- Do two data sets come from populations with a common distribution?
- In this case, no

Graphical Techniques

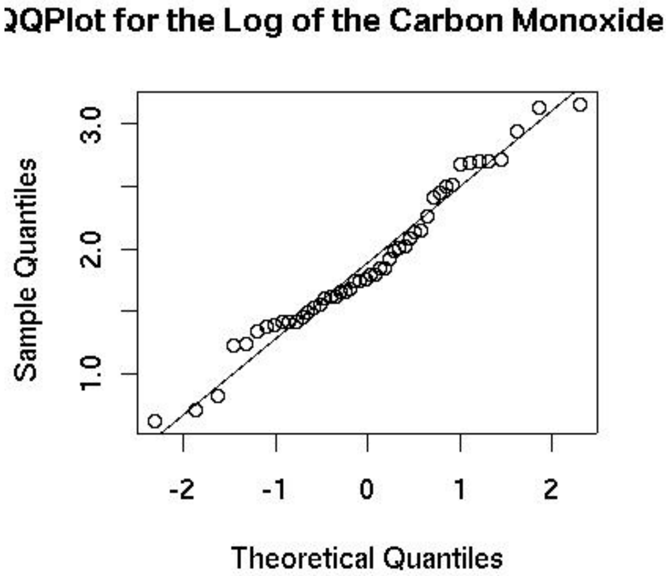
- Q-Q plot



- Do two data sets come from populations with a common distribution?
- In this case, maybe yes.
- Next step: Check with 2-sample K-S Test

Graphical Techniques

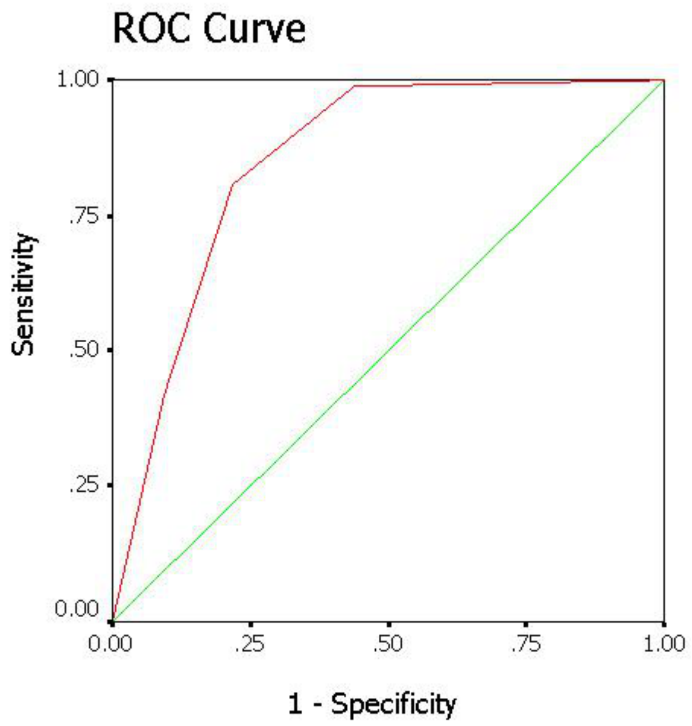
- Normal Q-Q plot



- Does the data set come from a population with a normal distribution?
- Most used way of checking normality before applying statistical tests
- A step further: Check normality with K-S Test

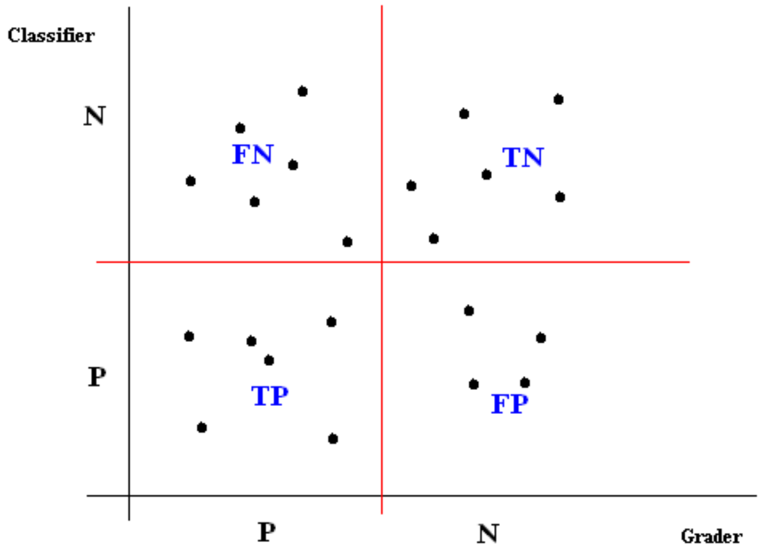
Graphical Techniques

- Find out the meaning!



Graphical Techniques

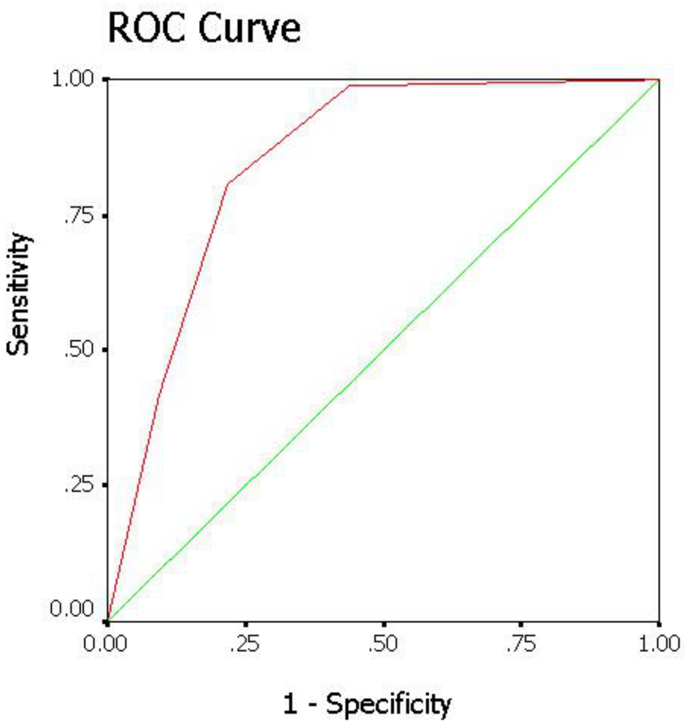
- ROC curves



- It evaluates the performance of classifiers with respect to a reference.
- Sensibility $= \dfrac{\text{TP}}{\text{TP+FN}}$
- Specificity $= \dfrac{\text{TN}}{\text{TN+FP}}$
- 1 - Specificity $= \dfrac{\text{FP}}{\text{TN+FP}}$

Graphical Techniques

- ROC curves



- ROC curves are drawn as the discriminator threshold is varied
- The larger the area under ROC curve, the better is the classifier
- Exception if the ROC curves are intersecting.

Graphical Techniques

- Other plots: Error Bar Plots

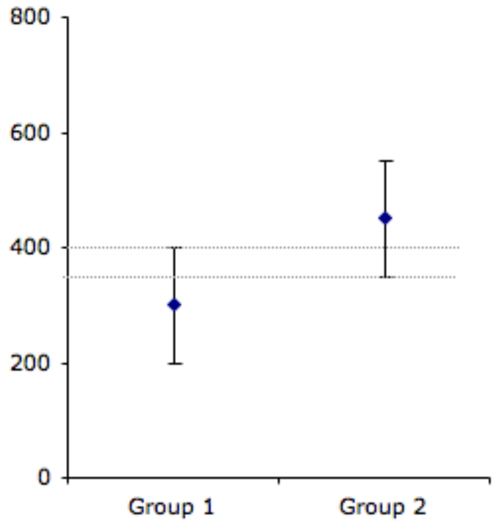
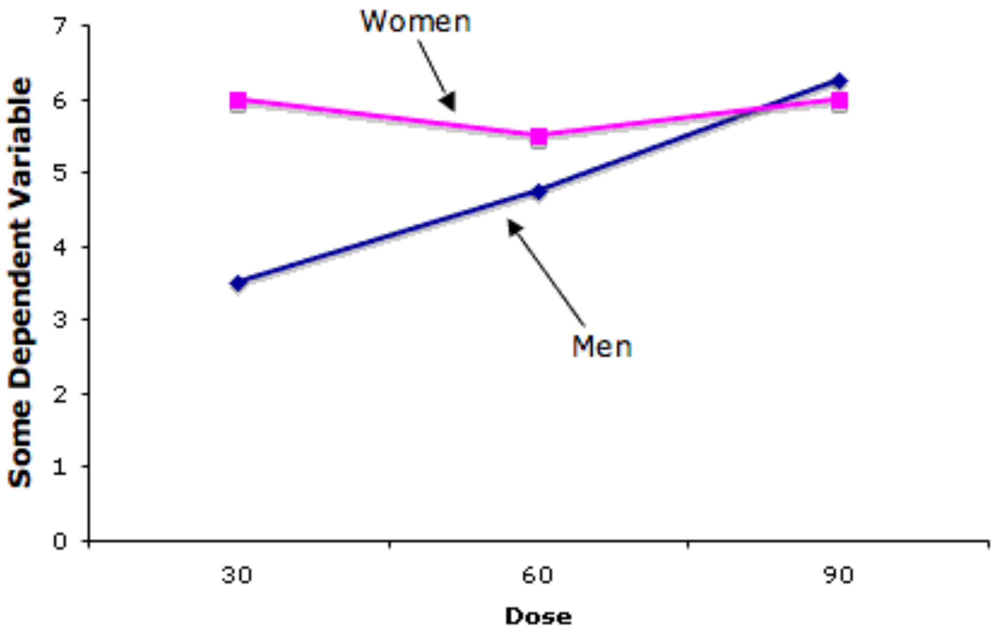


Figure 1: Mean reaction time (ms) and 95% confidence intervals for Group 1 (n=36) and Group 2 (n=34).

- It plots the means and some sort of error, e.g., confidence intervals (next lecture).
- Allows to detect significant differences between groups of observations.

Graphical Techniques

- Other plots: Interaction Plots



- It plots the means for each level of each factor (e.g., dose and gender).
- Allows to detect interaction between variables in an experimental design (next lecture).

Conclusions

- The main focus of EDA is on the data
- EDA relies strongly on graphical techniques to find structure and patterns in the data
- It suggests hypotheses to test
- But classical data analysis is still preferred as an ultimate step in research.