

# Data Preprocessing

# Why Preprocess Data?

---

- ❑ Raw data not ready to analyze
- ❑ Issues of data quality
- ❑ Conclusions drawn may be questionable or unreliable

# Measures for data quality

---

- ❑ Accuracy: is the data correct or wrong, accurate or not?
- ❑ Completeness: is there missing data?
- ❑ Consistency: are there conflicts in the data?
- ❑ Timeliness: is data old or recently updated?
- ❑ Believability: can you trust that the data is correct?
- ❑ Interpretability: how easily can the data be understood?

# Major Data Preprocessing Tasks

---

- ❑ **Data cleaning**

- ❑ Handle missing data, smooth noisy data, identify or remove outliers, and resolve inconsistencies

- ❑ **Data integration**

- ❑ Integration of multiple databases, data cubes, or files
- ❑ Often involves resolving conflicts between data sources

- ❑ **Data reduction and transformation**

- ❑ Speeds up analysis when data is *too* big
- ❑ E.g., can reduce rows (data points) or columns (attributes) of matrices

# Data Cleaning

---

- ❑ Data in the Real World Is Dirty: Lots of potentially incorrect data, e.g., faulty instruments, human or computer error, and transmission error
- ❑ Incomplete: lacking attribute values, lacking certain attributes of interest, or containing only aggregate data
  - ❑ e.g., *Occupation* = “ ” (missing data)
- ❑ Noisy: containing noise, errors, or outliers
  - ❑ e.g., *Salary* = “-10” (an error)

# Data Cleaning, continued

---

- ❑ Inconsistent: containing discrepancies in codes or names, e.g.,
  - ❑ Different data formats, e.g., rating “1, 2, 3” is now “A, B, C”
  - ❑ Different Scales/Units for Data Type ( £, \$, or €)
  - ❑ Discrepancy between duplicate records
- ❑ Intentional: (e.g., *disguised missing* data)
  - ❑ Defaults: Jan. 1 as everyone’s birthday?

# Incomplete (Missing) Data

---

- ❑ Data is not always available
  - ❑ E.g., many tuples have no recorded value for several attributes, such as customer income in sales data
- ❑ Missing data may be due to
  - ❑ Equipment malfunction
  - ❑ Inconsistent with other recorded data and thus deleted
  - ❑ Data were not entered due to misunderstanding
  - ❑ Certain data may not be considered important at the time of entry
  - ❑ Did not register history or changes of the data
- ❑ Missing data may need to be inferred

# How to Handle Missing Data?

---

- ❑ Ignore the tuple
  - ❑ Often not desirable, can cause data set to shrink dramatically
- ❑ Fill in the missing value manually
  - ❑ Tedious + infeasible?
- ❑ Fill in it automatically with
  - ❑ a global constant : e.g., “unknown”, a new class?!
  - ❑ the attribute mean
  - ❑ the attribute mean for all samples belonging to the same class: smarter
  - ❑ **the most probable value: inference-based such as Bayesian formula or decision tree**



# Handling Missing Data: Example

---

- ❑ Want to predict likely value for missing data
- ❑ Example: Student missing data for final course grade
  - ❑ This student is male, age 33, 4.0 GPA
  - ❑ Find similar people in the data and see what their value for final grade is
  - ❑ Fill missing spot with most likely final grade based on the other data

# Noisy Data

---

- ❑ **Noise:** random error or variance in a measured variable
- ❑ **Incorrect attribute values** may be due to various reasons
  - ❑ Faulty data collection instruments, Data entry problems, Data transmission problems, Technology limitation, Inconsistency in naming convention, ...
- ❑ **Other data problems**
  - ❑ Outliers
  - ❑ Duplicate records
  - ❑ Incomplete data
  - ❑ Inconsistent data

# How to Handle Noisy Data?

---

- ❑ Want to detect and (possibly) remove outliers
  - ❑ **Binning**
    - ❑ Sort data and partition into bins
    - ❑ Can smooth by bin means, bin median, bin boundaries, etc.
  - ❑ **Regression**
    - ❑ Smooth by fitting the data into regression functions
  - ❑ **Clustering**
    - ❑ Group data so that that points in the same cluster are more similar to each other than to those in other clusters
  - ❑ **Semi-supervised:** Combined computer and human inspection
    - ❑ Detect suspicious values and have humans check

# Data Cleaning as a Process

---

- ❑ Tools and guidelines exist to help with data cleaning
- ❑ **Not a one-pass task**
  - ❑ Often requires multiple rounds of identifying problems and resolving them

# Data Integration

---

- ❑ Data integration – What is it?
  - ❑ Combining data from multiple sources into a coherent store
- ❑ **Schema integration:**
  - ❑ e.g., A.cust-id  $\equiv$  B.cust-#
  - ❑ Integrate metadata from different sources
- ❑ **Entity identification:**
  - ❑ Identify real world entities from multiple data sources, e.g., Bill Clinton = William Clinton

# Data Integration – Why?

---

- Why data integration?
  - Clarifies data inconsistencies/Noise
    - Example: Age and Date of Birth.
      - Database 1 (Google): 02/26/1908; Age 38,
      - Database 2(Wikipedia): 02/26/1980; Age 38
        - Data from Database 2 clarifies the error in Year of Birth
  - Fills in Important Attributes for Analysis
    - Merging from more than 1 dataset provides more important information.
  - Speeds up Data Mining
    - One Master Schema can be mined rather than each of the 10 one-by-one

# Data Integration- Challenges

---

- ❑ What problems will you face?
  - ❑ Schema differences
    - ❑ Column is called “PersonAge” from Customer Table
    - ❑ Column is called “CustomerAge” from Person Table
  - ❑ Data Value Representation Conflicts
    - ❑ Database 1 -> “William Clinton”
    - ❑ Database 2 -> “Bill Clinton”
  - ❑ Bad Data
    - ❑ Typo; Wrong recording
    - ❑ Different Scales/Units for Data Type ( £, \$, or €)

# Data Integration - Handling Noise

---

- ❑ Detecting data value conflicts
  - ❑ For the same real world entity, attribute values from different sources are different
  - ❑ Possible reasons: no reason, different representations, different scales, e.g., metric vs. British units
- ❑ Resolving conflict information
  - ❑ Take the mean/median/mode/max/min
  - ❑ Take the most recent
  - ❑ Truth finding (Advanced): consider the source quality



# Data Integration - Handling Redundancy

---

- ❑ Redundant data often occurs when multiple databases are integrated
  - ❑ *Object identification / Entity Matching*: The same attribute or object may have different names in different databases
  - ❑ *Derivable data*: One attribute may be a “derived” attribute in another table, e.g., annual revenue
- ❑ What’s the problem?
  - ❑  $Y = 2X \rightarrow Y = X_1 + X_2 \quad Y = 3X_1 - X_2 \quad Y = -1291X_1 + 1293X_2$ 
    - ❑ Y equal to 2X in one DB, Y equal to sum of > 1 variable in another.
- ❑ Redundant attributes may be detected by correlation analysis and covariance analysis

# Example: stock market

Yahoo! Finance

Day's Range: 93.80-95.71

Nasdaq

## Green Mountain Coffee Roasters, (NasdaqGS: GMCR )


After Hours: 95.13 ↓ -0.01 (-0.02%) 4:07PM EDT

Last Trade:	<b>95.14</b>
Trade Time:	<b>4:00PM EDT</b>
Change:	<span style="color: green;">↑ 1.69 (1.81%)</span>
Prev Close:	<b>93.45</b>
Open:	<b>94.01</b>
Bid:	<b>95.03 x 100</b>
Ask:	<b>95.94 x 100</b>
1y Target Est:	

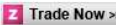
Day's Range:	<b>93.80 - 95.71</b>
52wk Range:	<b>25.38 - 95.71</b>
Volume:	<b>2,384,075</b>
Avg Vol (3m):	<b>2,512,070</b>
Market Cap:	<b>13.51B</b>
P/E (ttm):	<b>119.82</b>
EPS (ttm):	<b>0.79</b>
52wk Range:	<b>25.38-95.71</b>
52 Wk:	<b>25.38-93.72</b>


Last Sale	<b>\$ 95.14</b>
Change Net / %	<span style="color: green;">1.69 ▲ 1.81%</span>
Best Bid / Ask	\$ 95.03 / \$ 95.94
1y Target Est:	\$ 95.00
Today's High / Low	\$ 95.71 / \$ 93.80
Share Volume	2,384,175
50 Day Avg. Daily Volume	2,751,062
Previous Close	\$ 93.45
52 Wk High / Low	\$ 93.72 / \$ 25.38
Shares Outstanding	152,785,000
Market Value of Listed Security	\$ 14,535,964,900
P/E Ratio	120.43
Forward P/E	63.57
Earnings Per Share	\$ 0.79
Annualized Dividend	N/A
Ex Dividend Date	N/A
Dividend Payment Date	N/A
Current Yield	N/A
Beta	0.82
NASDAQ Official Open Price:	\$ 94.01
Date of NASDAQ Official Open Price:	Jul. 7, 2011
NASDAQ Official Close Price:	\$ 95.14
Date of NASDAQ Official Close Price:	Jul. 7, 2011

# Example: stock market

SALVEPAR (SY)  GET QUOTE Search InvestCenter >

Recent Quotes > My Watchlist > Top Indices >

**SALVEPAR**  Trade Now >>

 **-0.8900 (-1.212%)** at 72.55 EUR



70 in Volume

Add to:  
My  
Watchlist

Data as of  
04:18 AM  
EDT Jul 7,  
2011

Quote News Profile Research Community



**SYBASE (SY)**

 Like  Like 1

SOURCE: NYSE

As of July 29, 2010 4:04 pm. Quotes are delayed by at least 15 minutes

+0.01

 **\$64.98**  Change 209,960 \$64.97

Last Trade +0.02% Volume Prev. Close

Change (%)

**SY** 64.98 +0.00 (0.00%)

Click Here to Receive Instant E-mail and RSS Alerts

Trade SY now with **\$3.95** STOCK TRADES

**SALVEPAR (SY)**

29 Aug 2011 - 22 Feb 2012



©FinancialContent.com

Sep Nov 2012

Stock Details

Last Trade:	64.98
Change:	+0.00 (0.00%)
Prev Close:	64.98
Open:	14.73
Days Range:	64.98 - 64.98
52 Week Range:	33.54 - 66.00
Volume:	88168
P/E:	31.54
EPS:	2.06

# Example: stock market

**NASDAQ** One-click options strategies on Trade  
Trade free for 60 days + get up to \$600 cash. ▶

QUICK FIND: ETFs | Tools | After Hours | Global Indices | Earn a Degree | Company List

Home | Quotes & Research | Extended Trading | Market Activity | News

add symbol  
edit symbol list  
symbol lookup

Symbol List Views  
Flash Quotes  
Info Quotes  
Stock Details  
Real-Time Quotes  
Summary Quotes  
After Hours Quotes  
Pre-market Quotes  
Historical Quotes  
Options Chain  
CHARTS  
Basic Charts  
Interactive Charts  
COMPANY NEWS  
Company Headlines  
Press Releases  
Sentiment  
STOCK ANALYSIS  
Analyst Research  
Guru Analysis

Home > Quotes > Stock Quote > TTI

Trade Free for 60 days + Get up to \$600 with Trade Architect from TTI

TTI

Save my stocks for next time Investor Tools Tracking T

⚠ Cookies disabled? Please note that beginning 5/13/2011, you must have cookies. Please contact [isfeedback@nasdaq.com](mailto:isfeedback@nasdaq.com) with any questions or concerns.

**TTI: Stock Quote & Summary Data**

**\$ 13.11** 0.51 ▲ 4.05% TTI TTI

Jul. 7, 2011 Market Closed  
Update Quotes: On. Updates every 7 Seconds.

for TTI Commentary for TTI Price Charts Company Financials

Last Sale	\$ 13.11
Change Net / %	0.51 / 4.05%
1y Target Est.	\$ 16.00
Today's High / Low	\$ 13.11 / \$ 12.67
Share Volume	480,067
Previous Close	\$ 12.60
52 Wk High / Low	\$ 16 / \$ 8
Shares Outstanding	76,821,000
Market Value of Listed Security	\$ 1,007,123,310
P/E Ratio	NE
Forward P/E (1yr)	19.69
Earnings Per Share	\$ -0.68
Annualized Dividend	N/A

**UPDOWN**  
Beat the market. Earn real money. Zero risk.

HOME | TRADING | STOCKS | COMMUNITY | CO

Overview | Market News | Top Stock Picks

GET QUOTE Sponsore

**TETRA TECHNOLOGIES (TTI) 1**

76.82B

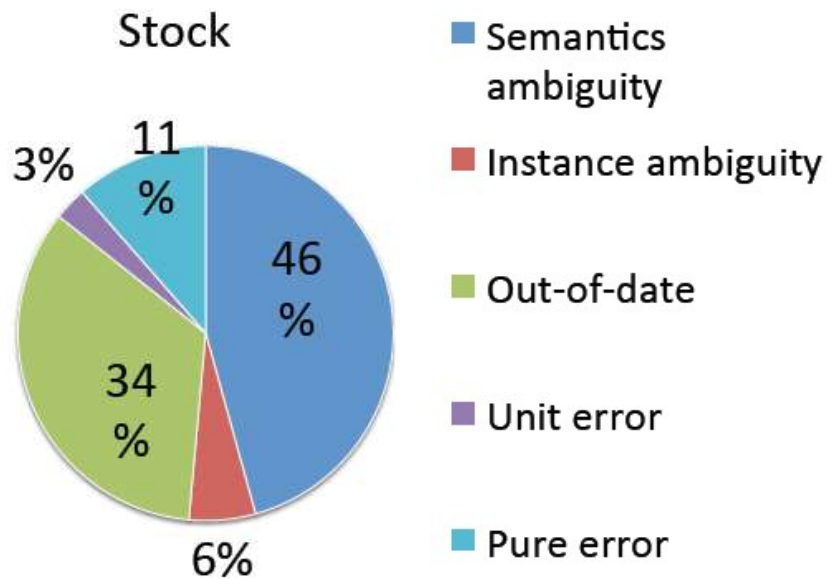
Overview Trade TTI Stock Picks Tweets

**TTI \$13.11 \$0.51 (4.05%)**

You need to upgrade your Flash Player

Today	5d	1m	3m	1y	5y	10y
Last:	\$13.11					
Prev Close:	\$12.60					
Open:	\$12.82					
Change:	\$0.51 (4.05%)					
Vol:	472,608					
Avg Volume:	559,308					
EPS:	-					
High:	\$13.15					
Low:	\$12.67					
Mkt Cap:	\$968M					
52Wk High:	\$16.00					
52Wk Low:	\$8.00					
Shares:	76.82B					
PE Ratio:	-					

# Example: stock market



Source	Accuracy	Coverage
<i>Google Finance</i>	.94	.82
<i>Yahoo! Finance</i>	.93	.81
<i>NASDAQ</i>	.92	.84
<i>MSN Money</i>	.91	.89
<i>Bloomberg</i>	.83	.81

Xian Li, Xin Luna Dong, Kenneth Lyons, Weiyi Meng, and Divesh Srivastava. Truth finding on the Deep Web: Is the problem solved? In *VLDB*, 2013.

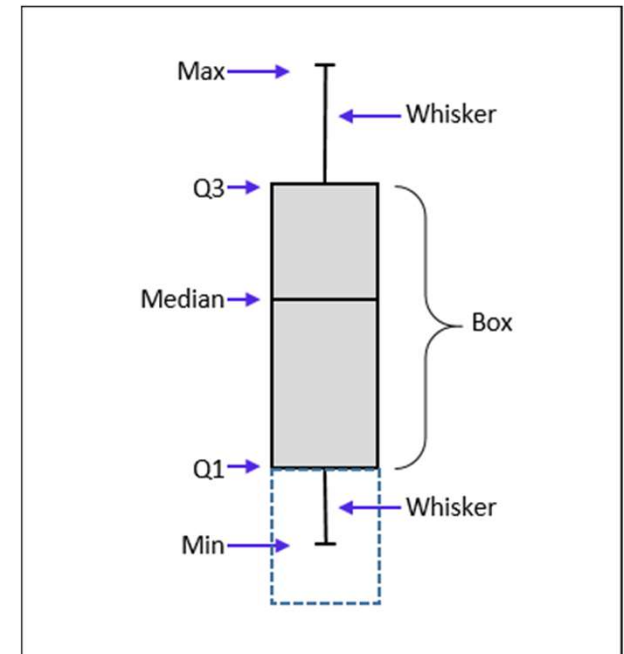
# Graphic Displays of Basic Statistical Descriptions

---

- ❑ **Boxplot:** five-number summary
- ❑ **Histogram:** values and frequencies
- ❑ **Scatter plot:** data plotted as points

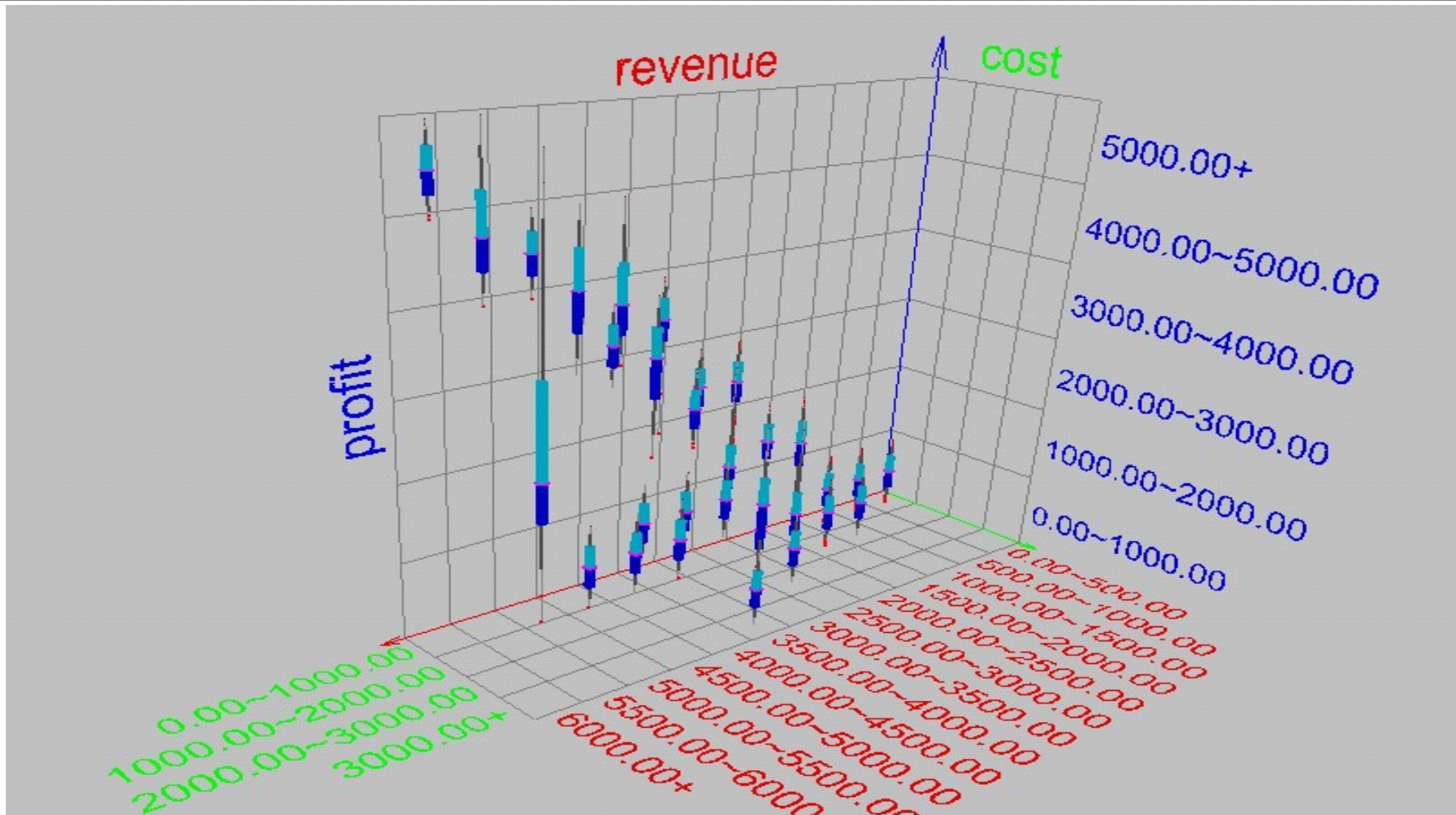
# Measuring the Dispersion of Data: Quartiles & Boxplots

- ❑ **Quartiles:**  $Q_1$  (25<sup>th</sup> percentile),  $Q_3$  (75<sup>th</sup> percentile)
- ❑ **Inter-quartile range:**  $IQR = Q_3 - Q_1$
- ❑ **Five number summary:** min,  $Q_1$ , median,  $Q_3$ , max
- ❑ **Boxplot:**
  - ❑ **Outliers:** points beyond a specified outlier threshold, plotted individually
  - ❑ **Outlier:** usually, a value higher/lower than  $1.5 \times IQR$





# Visualization of Data Dispersion: 3-D Boxplots

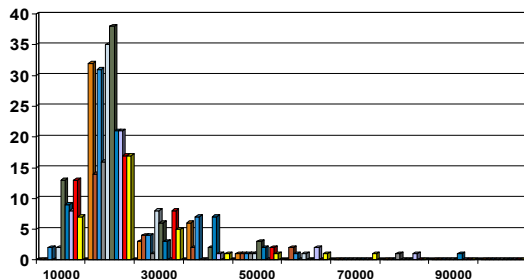




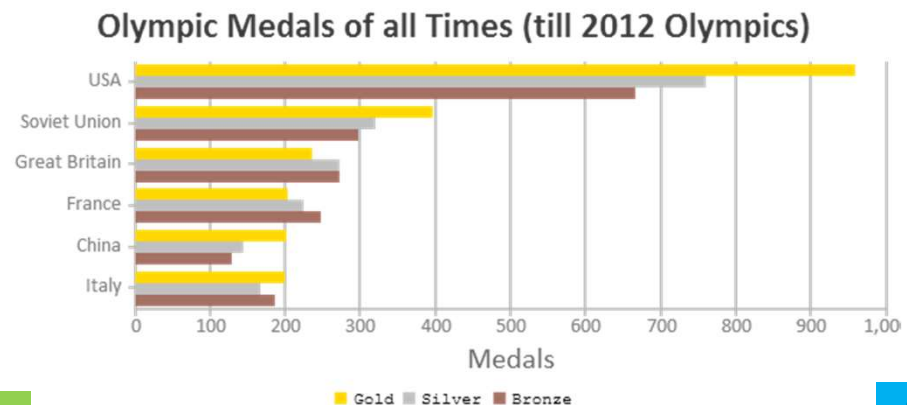
# Histogram Analysis

- Histogram: tabulated frequencies, shown as bars

Histogram	Bar charts
distributions of variables	compare variables
quantitative data	categorical data
Value: area of the bar	Value: height of the bar (a crucial distinction when the categories are not of uniform width )
Order matters	Can be reordered



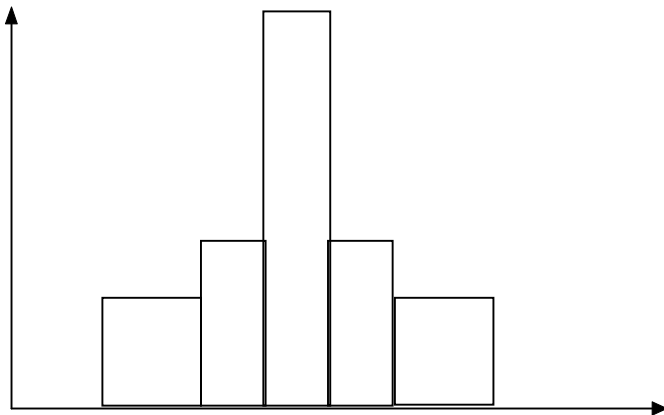
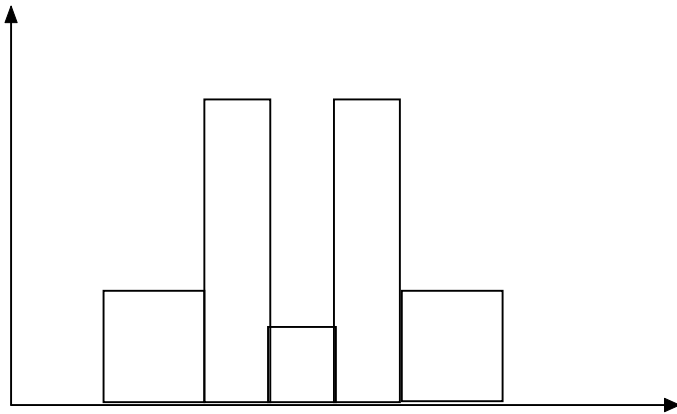
Histogram



Bar chart

# Histograms Often Tell More than Boxplots

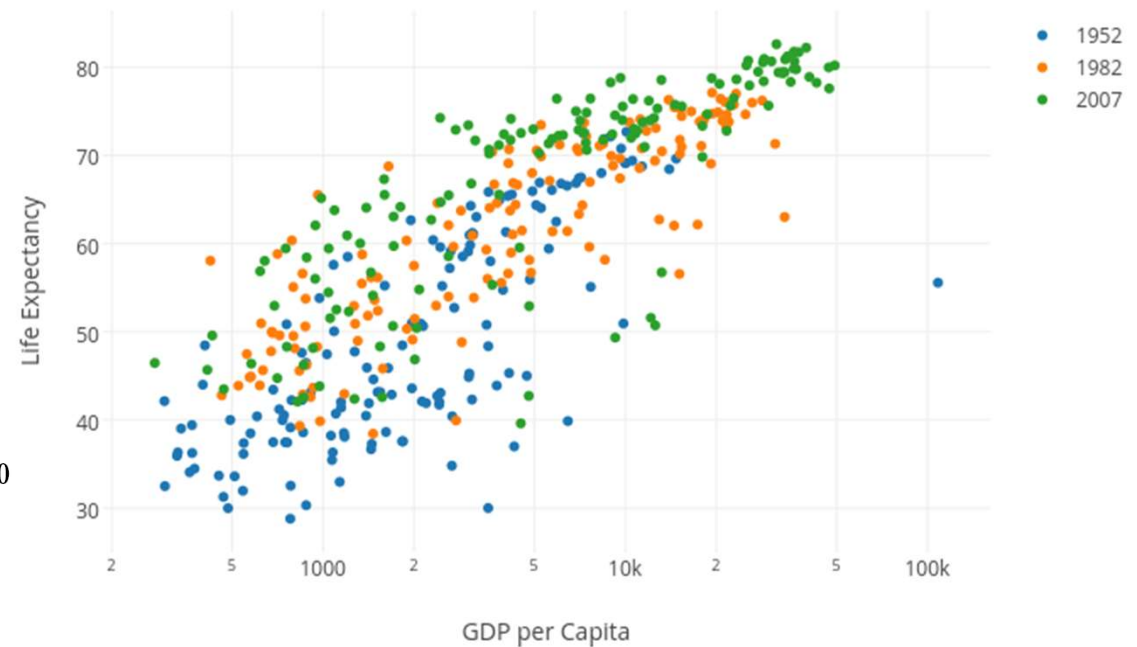
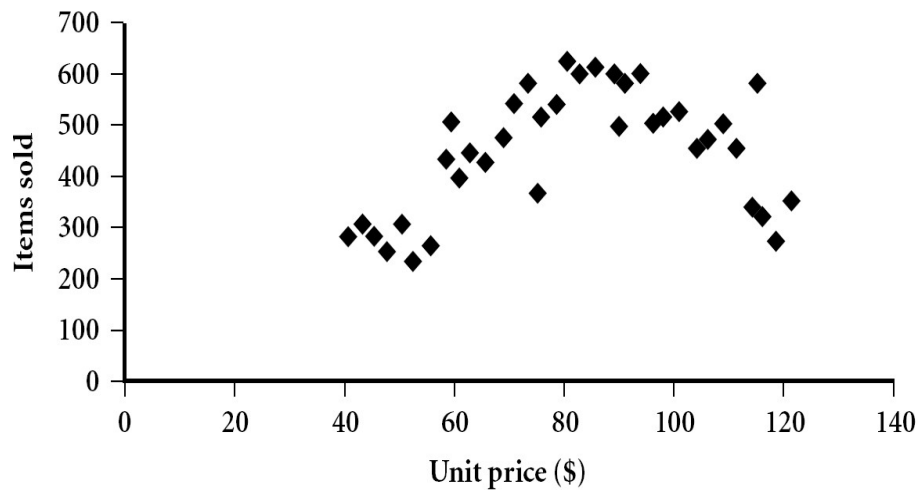
---



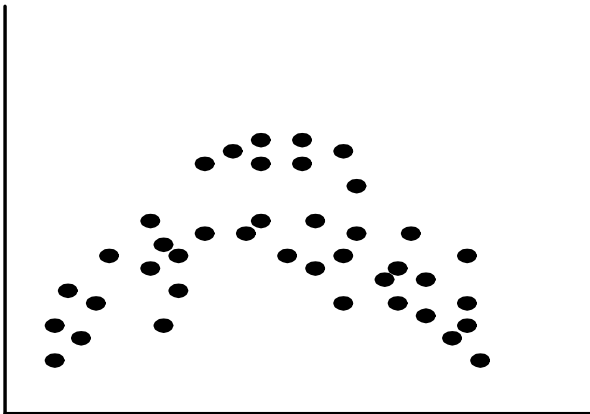
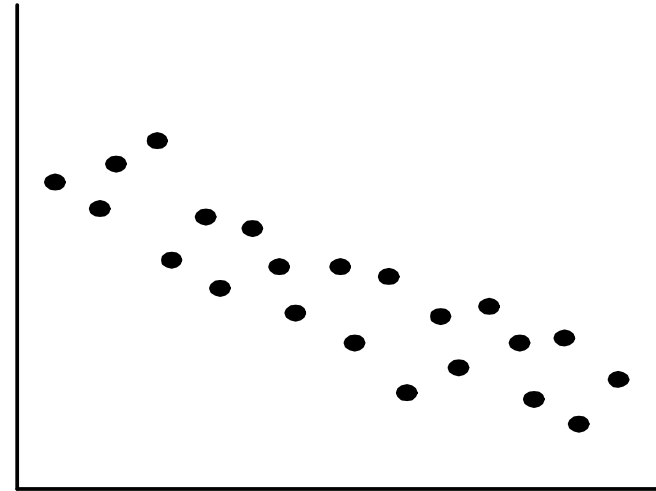
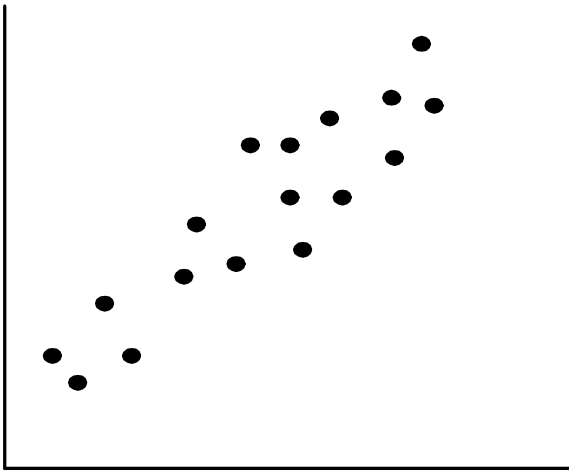
- Same boxplot representation
- The same min, Q1, median, Q3, max
- Different data distributions

# Scatter plot

- Provides a first look at **bivariate** data to see clusters of points, outliers, etc.



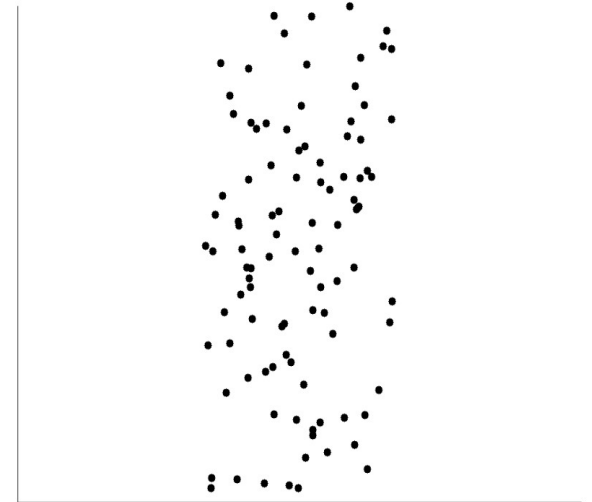
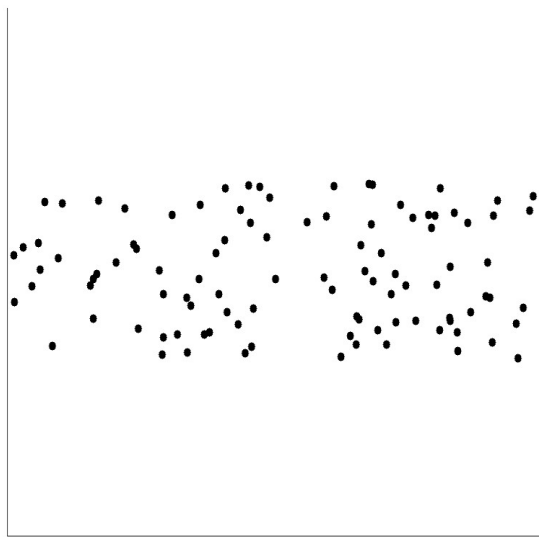
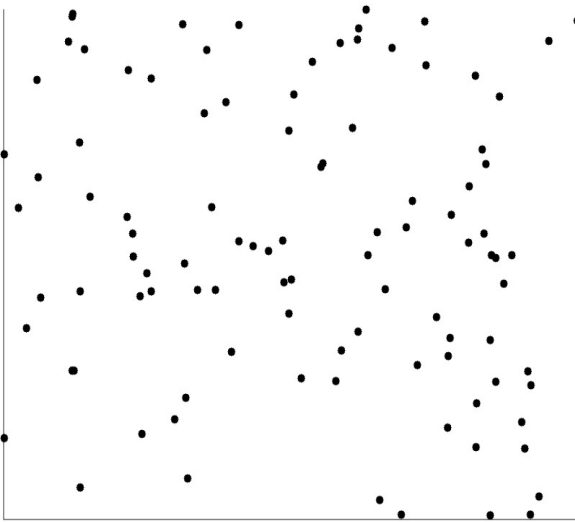
# Positively and Negatively Correlated Data



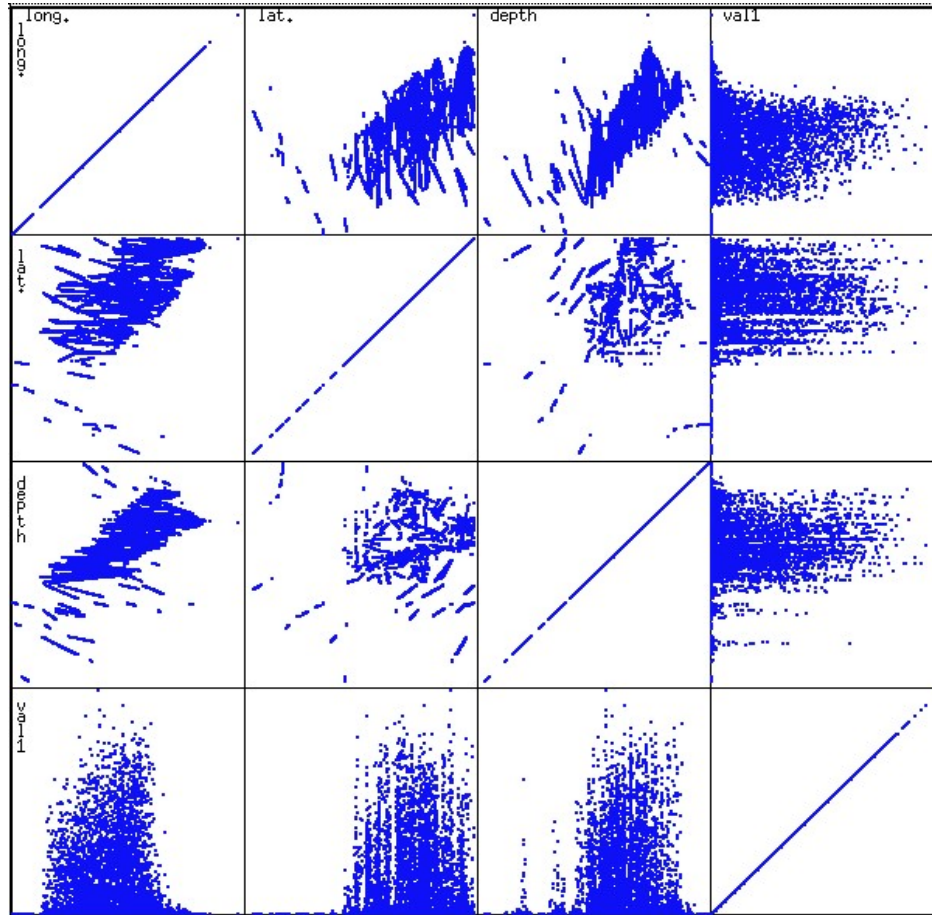
- ❑ The left half fragment is positively correlated
- ❑ The right half is negative correlated

# Uncorrelated Data

---



# Scatterplot Matrices



- ❑ Matrix of scatterplots (x-y-diagrams) of the k-dim. data
- ❑ A total of  $k(k-1)/2$  distinct scatterplots
- ❑ Good for understanding whether two variables are correlated
- ❑ Not as helpful for high-dimensional data