Data Science
(lecture)

- Simona Vasilache -
University of Tsukuba

“Data Science” course at UT

- Mandatory course for all freshmen
- 10 weeks, double-period class (75 min x 2)
- Each class: two parts - lecture and exercises
- English version (for undergraduate English programs)

Course overview and keywords

Week	Overview	Keywords
1	Overview of Data Science	course overview, video lecture (“Introduction of Data Science”)
2	Data Type and Collection (1)	data types and scale levels, data collection and survey methods, basic usage of Excel
3	Data Type and Collection (2)	basic process of data science, data preprocessing, data reusability; Excel (statistics)
4	Data Type and Collection (3)	data management, separation of document structure from visual representation; Excel (data operation)
5	Advanced Data Management	video lecture (“Big Data”)
6	Data Visualization and Qualitative Data Analysis	significance and purpose of visualization, choice of visual representations, visualization of data with Excel
7	One Variable Quantitative Data	understanding one variable quantitative data, histograms, descriptive statistics; statistical tests
8	Two Variable Quantitative Data and Time Series Data	understanding two variable quantitative data, correlation and correlation coefficient, time series data (regression line)
9	Analysis of Actual of Data	analysis of various data, network data, correlation and causality
10	Advanced Data Management	video lecture (“Artificial Intelligence”)

Visualizing Data

Data Science

(lecture)

Goals of the lecture

- Data visualization
  - **Understand significance of visualization and the visualization process**
  - Learn how to choose visual attributes and visual representations (graphs)
  - Understand suitable / unsuitable graphs
  - Learn how to draw basic graphs in Excel
- Analysis of qualitative data
  - Categorize and understand frequency distribution table
  - Understand qualitative data statistics
  - Understand and utilize cross tabulations and odds ratio
  - Understand stacked bar charts

Significance of visualization and Visualization processing model

Famous example: Anscombe's quartet

- Four data sets with almost identical simple descriptive statistics
- Proposed in 1973 by Francis Anscombe (English statistician)
- Purpose: to show importance of visualization and counteract some common beliefs

- "Numerical calculations are exact, but graphs are rough"
- "Performing intricate calculations is virtuous, whereas actually looking at the data is cheating"

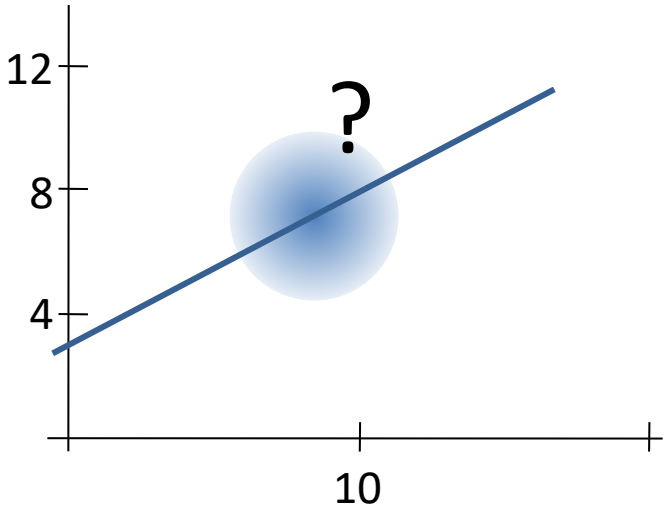
“Anscombe’s Quartet”

I		II		III		IV	
x	y	x	y	x	y	x	y
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
12.0	10.84	12.0	9.13	12.0	8.15	8.0	5.56
7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89

Source: F. J. Anscombe, Graphs in Statistical Analysis, The American Statistician, Vo. 27, No. 1, pp. 17-21, 1973

Features of the four datasets

- Each dataset has 11 pairs
  - mean of x: 9.0
  - variance of x: 11.0
  - mean of y: 7.5
  - variance of y: 4.1
  - correlation coefficient : 0.82

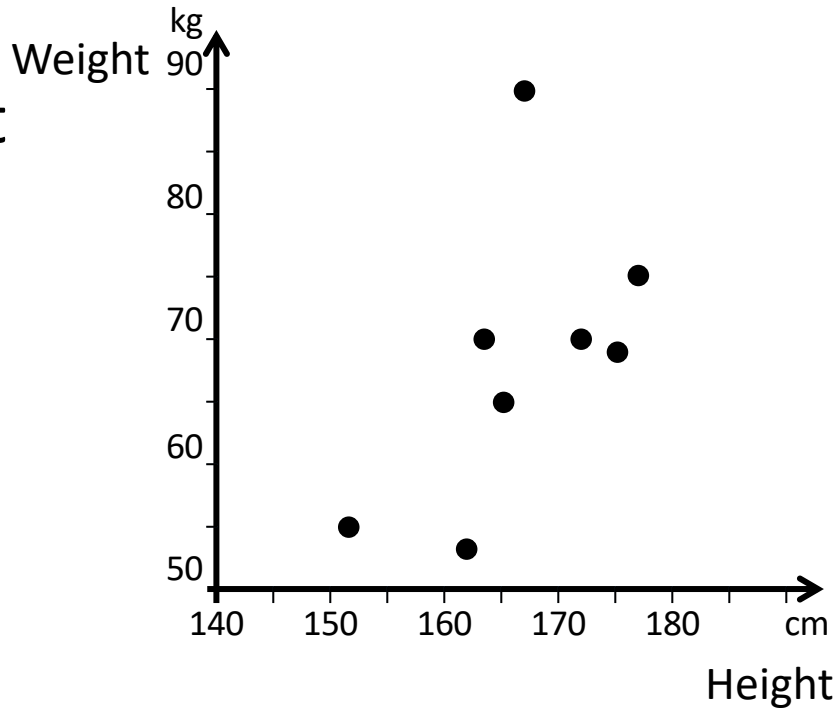


- Mean, variance, correlation coefficient are the same for all datasets
- Q: *Are all four datasets the same? What is the difference between them?*

Features of the four datasets

Scatter plot visualization

- Example: visualization of quantitative data for two variables - height and weight

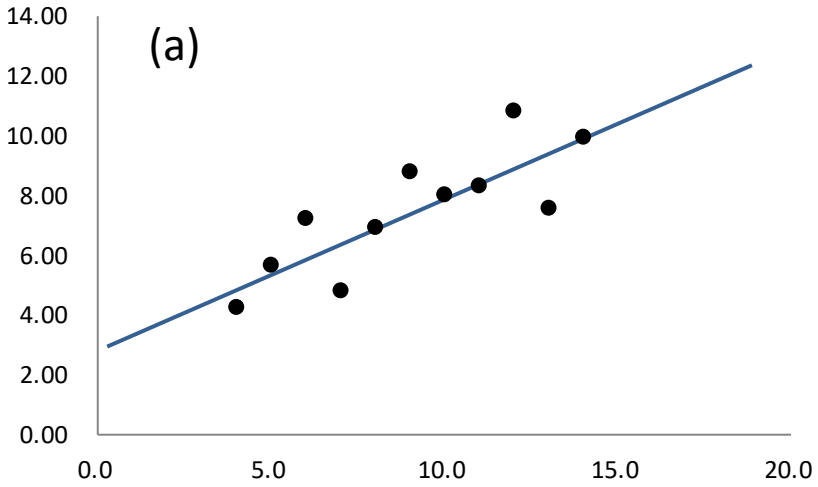


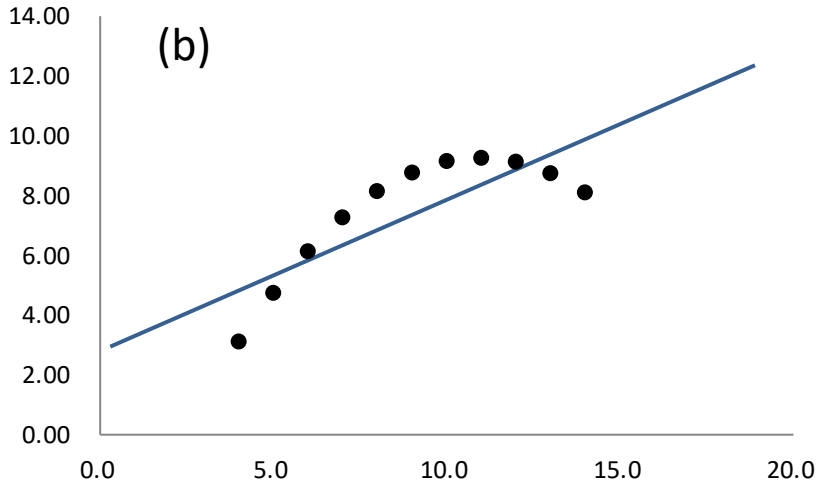
Target Data	**Quantitative data (two variables):** Height, Weight

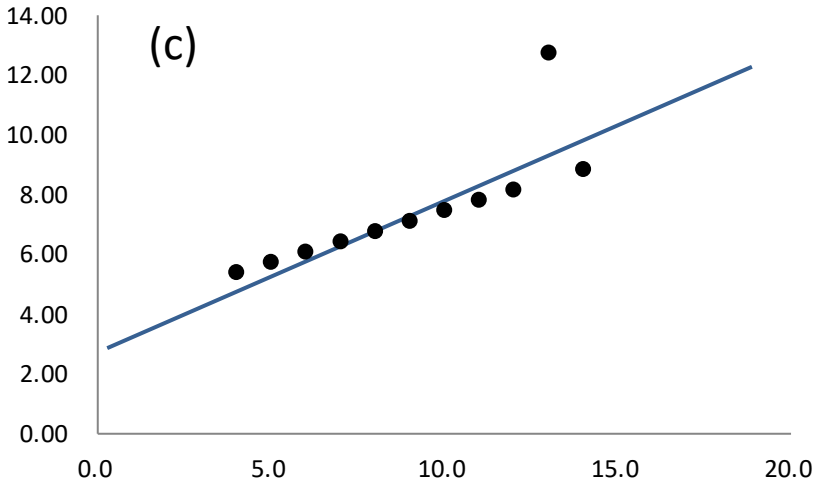
- Record ⇒ Dot
- Variable 1 (Height) ⇒ Horizontal coordinates
- Variable 2 (Weight) ⇒ Vertical coordinates

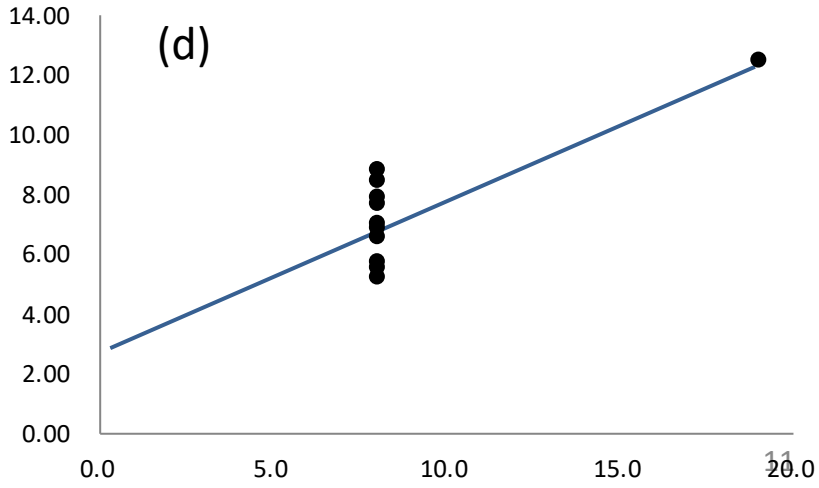
Quiz: Find corresponding scatter plot for each of the 4 Anscombe's data sets

(data given in excel file on manaba)



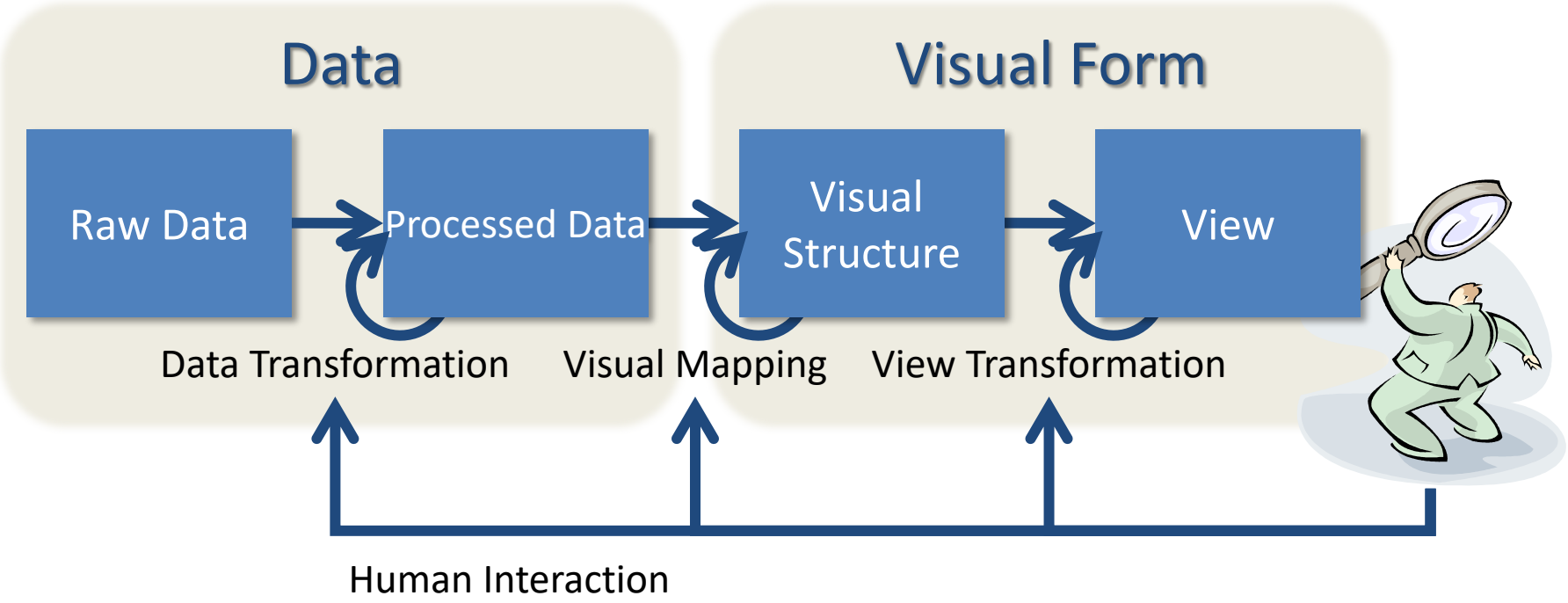






Visualization processing model

- Generalized processing flow in visualization

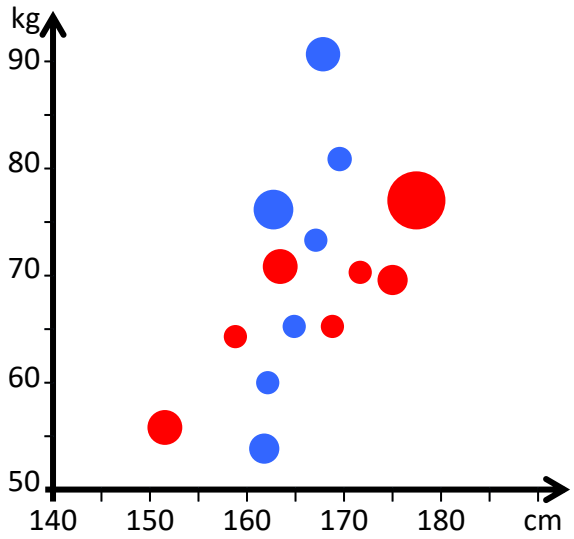


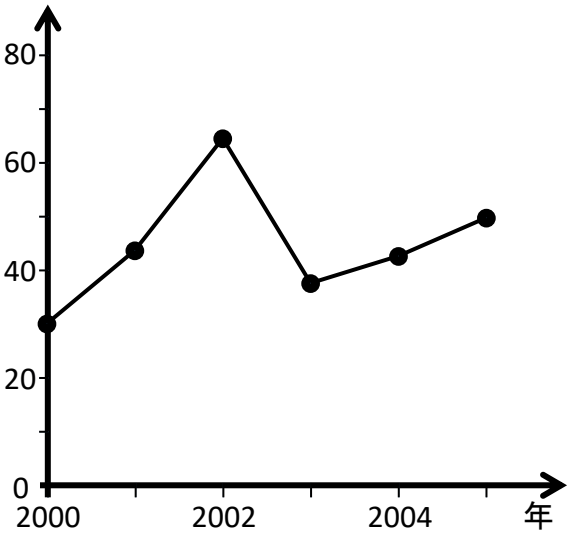
Source:
Figure 1.23 in Stuart K. Card et al., Readings in Information Visualization: Using Vision to Think, Morgan Kaufmann, 1999.

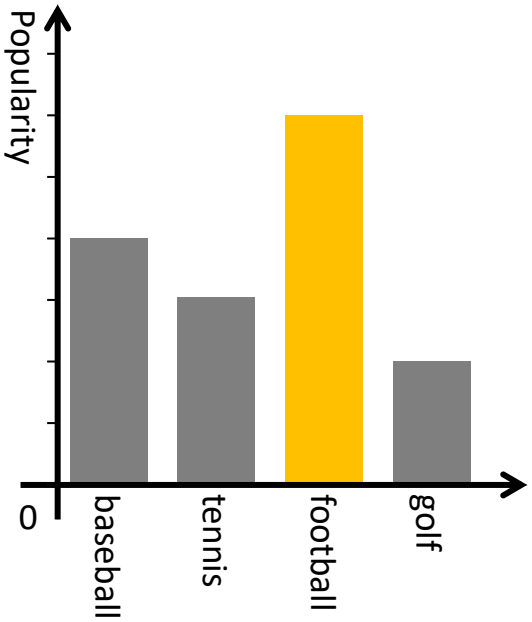
Components of visual representation

VISUAL ATTRIBUTES

Representing data



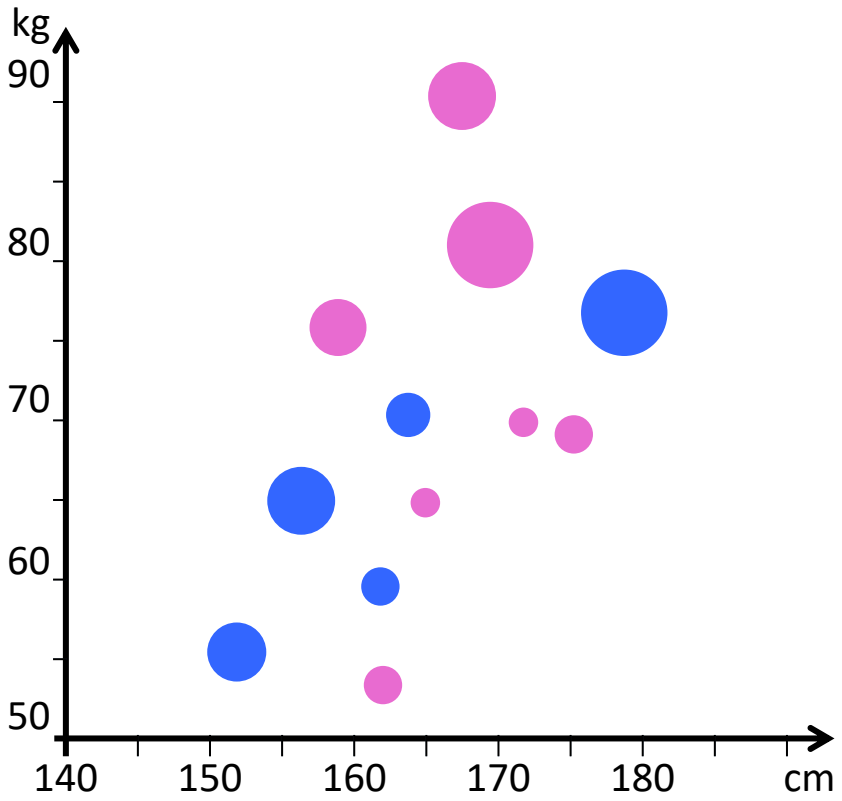




What is used to represent values?

- What is used to represent values in the graph?

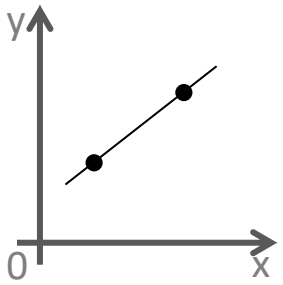
- Hue (blue, magenta)
- Size of circles (area)
- Horizontal positions
- Vertical positions



Visual Attributes

Visual attributes that can represent values

Position



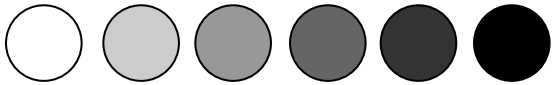
Hue

Texture

Size (area, length, width)

Orientation

Brightness



Shape

Nota Bene

- The graph consists of a combination of basic shapes (e.g. points) and visual attributes (e.g. positions)
- The importance of clarifying the rules cannot be underestimated!
- In order for the visual representation to convey information, the author and the reader must share the same rules!

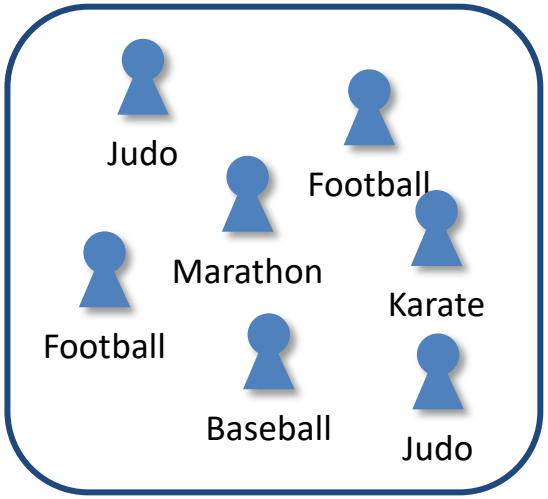
ANALYSIS OF QUALITATIVE DATA

(one qualitative variable)

- Categorization, frequency distribution table
- Bar chart, stacked bar chart
- Statistics of qualitative data

Categorization (1)

- Categorizing records by a qualitative attribute
  - Count records with the same value when only interested in the number of records (frequency)
    - E.g. Number of people who like a certain kind of sport



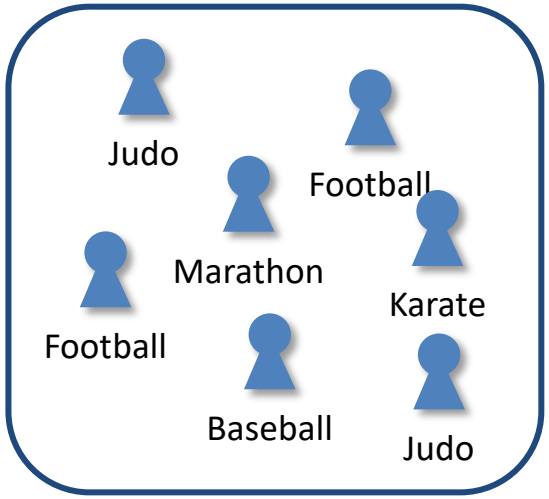
Categorization

Frequency table

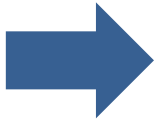
Class	Frequency
Baseball	1
Football	2
Marathon	1
Judo	2
Karate	1

Categorization (2)

- Categorizing records by a qualitative attribute
  - May be grouped into general categories
    - E.g. Baseball, Football → Ball game



Categorization



Frequency table

Class	Frequency
Ball game	3
Athletic	1
Martial arts	3

Frequency and relative frequency (ratio)

- Frequency
  - Number of occurrences of each observed value
- Relative frequency (ratio)
  - Percentage of number of occurrences (frequency) of each observed value when the entire data (no. of cases) is used as denominator
- Cumulative relative frequency
  - Sum of relative frequencies up to the observed value (cumulative sum)

Category	Frequency	Relative frequency	Cumulative relative frequency
Ball game	3	=3/7(≒0.43)	≒0.43
Athletic	1	=1/7(≒0.14)	≒0.43+0.14=0.57
Martial arts	3	=3/7(≒0.43)	≒0.57+0.43 = 1

Remaining points…

- Data visualization
  - Understand significance of visualization and the visualization process
  - Learn how to choose visual attributes and visual representations (graphs)
  - Understand suitable / unsuitable graphs
  - Learn how to draw basic graphs in Excel
- Analysis of qualitative data
  - Categorize and understand frequency distribution table
  - Understand qualitative data statistics
  - Understand and utilize cross tabulations and odds ratio
  - Understand stacked bar charts

Thank you for listening!