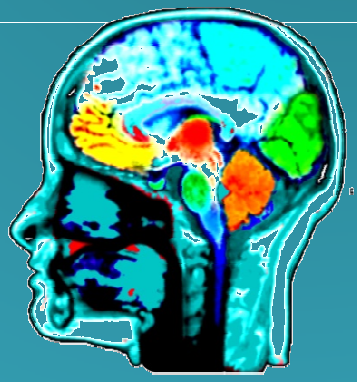
Introduction To Data Mining

Isfahan University of Technology (IUT)

Bahman 1401



Getting to Know Your Data

Dr. Hamidreza Hakim
hamid.hakim.u@gmail.com

Content

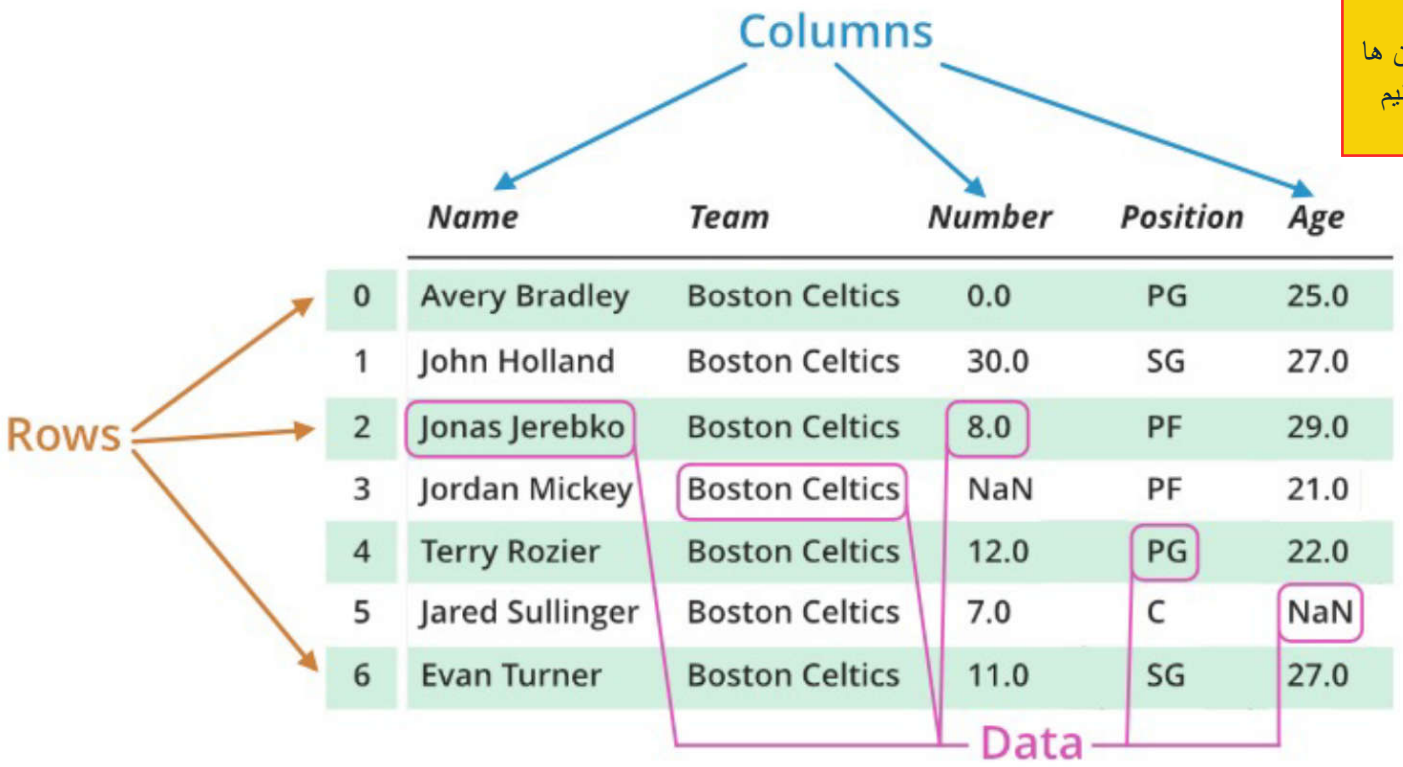
- Attributes and Objects
- Types of Data
- Basic Statistical Descriptions of Data
- Data Visualization
- Similarity and Dissimilarity Measures

ATTRIBUTES AND OBJECTS

What is Data?

- Collection of *data objects* and their *attributes*

ما یک شی را با ویژگی هاش میشناسیم
ویژگی ها را به عنوان ستون ها و اتریبیوت ها توصیف میکنیم



	Name	Team	Number	Position	Age
0	Avery Bradley	Boston Celtics	0.0	PG	25.0
1	John Holland	Boston Celtics	30.0	SG	27.0
2	Jonas Jerebko	Boston Celtics	8.0	PF	29.0
3	Jordan Mickey	Boston Celtics	NaN	PF	21.0
4	Terry Rozier	Boston Celtics	12.0	PG	22.0
5	Jared Sullinger	Boston Celtics	7.0	C	NaN
6	Evan Turner	Boston Celtics	11.0	SG	27.0

What is Data?

یه سری مشخصه هستند که ابجکت را برامون توصیف میکنند

- An ***attribute*** is a property or characteristic of an object
  - Examples: eye color of a person, temperature, etc.
  - Attribute is also known as variable, field, characteristic, dimension, or feature
- A collection of attributes describe an ***object***
  - Object is also known as record, point, case, sample, entity, or instance

Attributes

Objects

Tid	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

Attribute Values

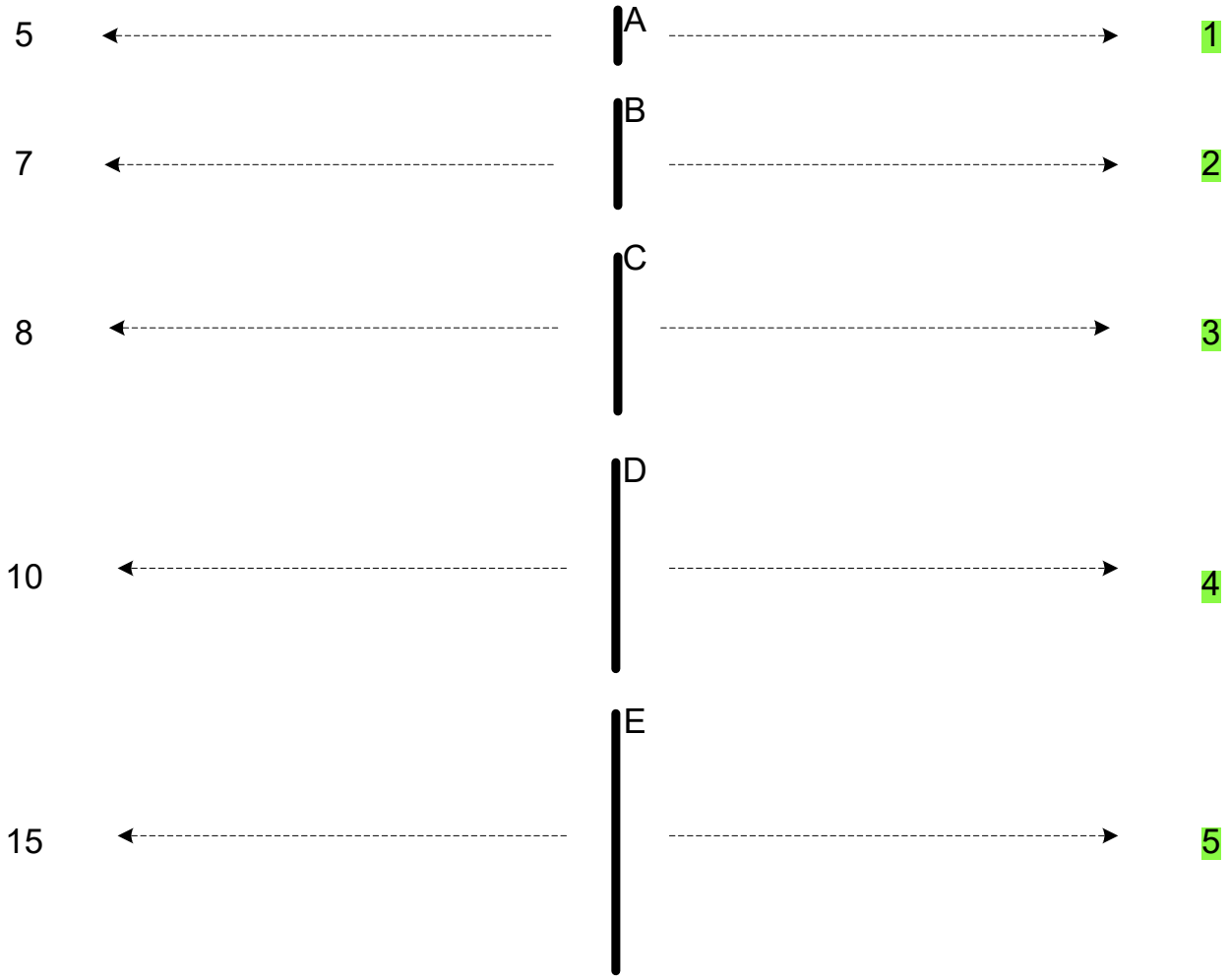
- ***Attribute values*** are numbers or symbols assigned to an attribute for a particular object

- Distinction between attributes and attribute values
  - Same attribute can be mapped to different attribute values
    - Example: height can be measured in feet or meters
  - Different attributes can be mapped to the same set of values
    - Example: Attribute values for ID and age are integers
  - But properties of attribute can be different than the properties of the values used to represent the attribute

Measurement of Length

ایا فاصله ها و مقیاس هارا داریم حفظ میکنیم؟ ایا اصلا خوبه که مقیاس ها حفظ شوند؟
چطوری با اندازه ها باید برخورد کنیم

- The way you measure an attribute may not match the attributes properties.



This scale preserves only the ordering property of length.

This scale preserves the ordering and additvity properties of length.

Types of Attributes

- There are different types of attributes
  - Nominal:
    - categories, states, or "names of things"
    - Hair_color = {auburn, black, blond, brown, grey, red, white}
    - marital status, occupation, ID numbers, zip codes
    - Examples: ID numbers, eye color, zip codes
  - Ordinal:
    - Values have a meaningful order (ranking) but magnitude between successive values is not known
    - Size = {small, medium, large}, grades, army rankings
    - Examples: rankings (e.g., taste of potato chips on a scale from 1-10), grades, height {tall, medium, short}

اسمی
دسته ای

اتریبیوت هایی که دسته دسته و
حالت حالت هستند مثل برچسب ها

یک توالی و سیکونسی بین مقادیر وجود
داره
اندازه کتاب : کوچک، متوسط، بزرگ

مقادیر دارای نظم (رتبه بندی) معنیداری هستند اما
مقدار بین مقادیر متوالی مشخص نیست

Types of Attributes(Example)

- There are different types of attributes
  - Nominal
  - Ordinal
  - Interval

متغیر های فاصله ای

    - Measured on a scale of equal-sized units q Values have order
    - Examples: calendar dates, temperatures in Celsius or Fahrenheit.

منفی هم میتوانند بشوند.

  - Ratio

اونایی که یه پایه وبیس صفر دارند حتما ratio هستند یعنی منفی ندارند مثل تعداد افراد در اتاق

منفی برای طول معنا نداره پس میشه ریشیو

    - Inherent zero-point
    - Examples: temperature in Kelvin, length, counts, elapsed time (e.g., time to run a race)

مثلا دما که صفر هم توش هست اگه بگیم دما منفی دو درجه است، دوبرابرش چی میشه؟ دوبرابر بیشتر را چطوری حساب کنیم؟ اگه بگیم دوبرابر گرمتر یا دوبرابر سردتر مشخص میشه به سمت چی باید ببریم عدد را

Question

- Q1: Is student ID a nominal, ordinal, or interval-scaled data?
- Q2: What about eye color? Or color in the color spectrum of physics?

Question

- Q1: Is student ID a nominal, ordinal, or interval-scaled data?

Nominal

- Q2: What about eye color? Or color in the color spectrum of physics? q

Eye color: Nominal (similar to hair color)

Color spectrum of physics: Interval (RGB space supports +/-)

Properties of Attribute Values

- The type of an attribute depends on which of the following properties/operations it possesses:
  - Distinctness: = ≠
  - Order: < >
  - Differences are meaningful + -
  - Ratios are meaningful * /

نسبت

  - Nominal attribute: distinctness

ترتیبی

  - Ordinal attribute: distinctness & order
  - Interval attribute: distinctness, order & meaningful differences
  - Ratio attribute: all 4 properties/operations

Difference Between Ratio and Interval

- Is it physically meaningful to say that a temperature of 10 ° is twice that of 5° on
  - the Celsius scale?
  - the Fahrenheit scale?
  - the Kelvin scale?

چون کلوین یه بیس صفر داره پس ریشیو است یعنی مقدار منفی نداره پس عملیات های ضرب و تقسیم را ساپورت میکنه پس فقط برای کلوین میشه بگیم ۱۰ درجه ی کلوین دوبرابر درجه ی کلوین است. برای سلسیوس و فارنهایت نمیشه گفت چون ممکنه منفی باشه درجه اش و نمیتوانیم بگیم ۲برابر سردتره یا گرمتر؟

- Consider measuring the height above average
  - If Bill's height is three inches above average and Bob's height is six inches above average, then would we say that Bob is twice as tall as Bill?
  - Is this situation analogous to that of temperature?

	Attribute Type	Description	Examples	Operations
Categorical Qualitative	Nominal	Nominal attribute values only distinguish. (=, ≠)	zip codes, employee ID numbers, eye color, sex: {*male, female*}	mode, entropy, contingency correlation, $\chi 2$ test
	Ordinal	Ordinal attribute values also order objects. (<, >)	hardness of minerals, {*good, better, best*}, grades, street numbers	median, percentiles, rank correlation, run tests, sign tests
Numeric Quantitative	Interval	For interval attributes, differences between values are meaningful. (+, -)	calendar dates, temperature in Celsius or Fahrenheit	mean, standard deviation, Pearson's correlation, *t* and *F* tests
	Ratio	For ratio variables, both differences and ratios are meaningful. (*, /)	temperature in Kelvin, monetary quantities, counts, age, mass, length, current	geometric mean, harmonic mean, percent variation

This categorization of attributes is due to S. S. Stevens

	Attribute Type	Transformation	Comments
Categorical Qualitative	Nominal	Any permutation of values	If all employee ID numbers were reassigned, would it make any difference?
	Ordinal	An order preserving change of values, i.e., *new_value = f(old_value)* where *f* is a monotonic function	An attribute encompassing the notion of good, better best can be represented equally well by the values {1, 2, 3} or by { 0.5, 1, 10}.
Numeric Quantitative	Interval	*new_value = a * old_value + b* where a and b are constants	Thus, the Fahrenheit and Celsius temperature scales differ in terms of where their zero value is and the size of a unit (degree).
	Ratio	*new_value = a * old_value*	Length can be measured in meters or feet.

اگه همه ی شماره دانشجویی ها عوض بشه هیچ تفاوتی ایجاد میشه چون فرقی نداره شماره دانشجویی من ۱ باشه یا ۲

تغییر و Transform و بدیل کردن اتریبیوت های اردینال باید ترتیب را حفظ کنه یعنی کوچک و بزرگی مهم است ولی ارتباط بین ولیوها مهم نیست.

This categorization of attributes is due to S. S. Stevens

گسسته پیوسته

Discrete and Continuous Attributes

- Discrete Attribute
  - Has only a finite or countably infinite set of values
  - Examples: zip codes, counts, or the set of words in a collection of documents
  - Often represented as integer variables.
  - Note: binary attributes are a special case of discrete attributes
- Continuous Attribute
  - Has real numbers as attribute values
  - Examples: temperature, height, or weight.
  - Practically, real values can only be measured and represented using a finite number of digits.
  - Continuous attributes are typically represented as floating-point variables.

Asymmetric Attributes

- Only presence (a non-zero attribute value) is regarded as important
  - Words present in documents
  - Items present in customer transactions

- If we met a friend in the grocery store would we ever say the following?

  "I see our purchases are very similar since we didn't buy most of the same things."

Critiques of the attribute categorization

- Incomplete
  - Asymmetric binary
  - Cyclical
  - Multivariate
  - Partially ordered
  - Partial membership
  - Relationships between the data
- Real data is approximate and noisy
  - This can complicate recognition of the proper attribute type
  - Treating one attribute type as another may be approximately correct

TYPES OF DATA

Types of data sets

- Record(Tabuler)
  - Data Matrix
  - Document Data
  - Transaction Data
- Graph
  - World Wide Web
  - Molecular Structures
- Ordered
  - Spatial Data
  - Temporal Data
  - Sequential Data
  - Genetic Sequence Data

csv files, database tables each row is independant of another fow
هررکورد مثلا برای یک نفره

network

داده های دارای ترتیب
- داده های فضایی
- داده های زمانی
- داده های متوالی
- داده های توالی ژنتیکی

Record Data

- Data that consists of a collection of records, each of which consists of a fixed set of attributes

Tid	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

Data Matrix

همه ی اتریبیوت ها عددی هستند.

- If data objects have the same fixed set of numeric attributes, then the data objects can be thought of as points in a multi-dimensional space, where each dimension represents a distinct attribute

- Such a data set can be represented by an *m* by *n* matrix, where there are *m* rows, one for each object, and *n* columns, one for each attribute

Projection of x Load	Projection of y load	Distance	Load	Thickness
10.23	5.27	15.22	2.7	1.2
12.65	6.25	16.22	2.2	1.1

Document Data

کاربا داده های متنی
مثلا میخایم بفهمیم
موضوع فایل پی دی افی
که دستمون هست چیه؟
مثلا اگه کلمات مرتبط با
ورزش زیاد تکرار شده
میگیم احتمالا پی دی افه
درباره ورزشه

- Each document becomes a 'term' vector
  - Each term is a component (attribute) of the vector
  - The value of each component is the number of times the corresponding term occurs in the document.

این ترم ها میشن
اتریبیوت های داده ی
داکیومنت ما

	team	coach	play	ball	score	game	win	lost	timeout	season
Document 1	3	0	5	0	2	6	0	2	0	2
Document 2	0	7	0	2	1	0	0	3	0	0
Document 3	0	1	0	0	1	2	2	0	3	0

هرکلمه ای چندبار توی
متن تکرار شده؟

Transaction Data

- A special type of data, where
  - Each transaction involves a set of items.
  - For example, consider a grocery store. The set of products purchased by a customer during one shopping trip constitute a transaction, while the individual products that were purchased are the items.
  - Can represent transaction data as record data

ایدی تراکنش، هررکورد یک تراکنش است.

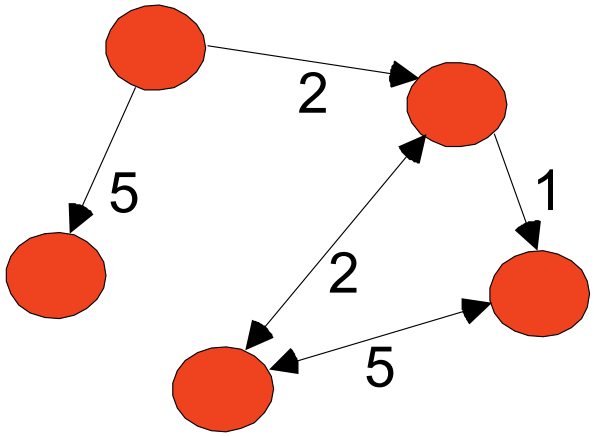
هرکدوم از این محصول هایی که خریداری شدند یک ایتم هستند.

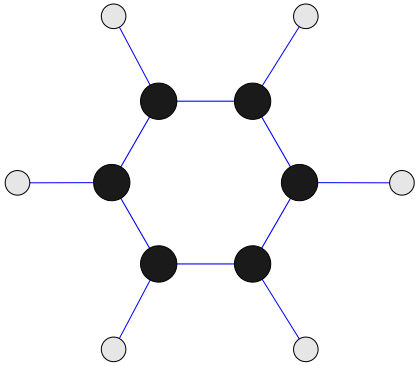
TID	Items
1	Bread, Coke, Milk
2	Beer, Bread
3	Beer, Coke, Diaper, Milk
4	Beer, Bread, Diaper, Milk
5	Coke, Diaper, Milk

Graph Data

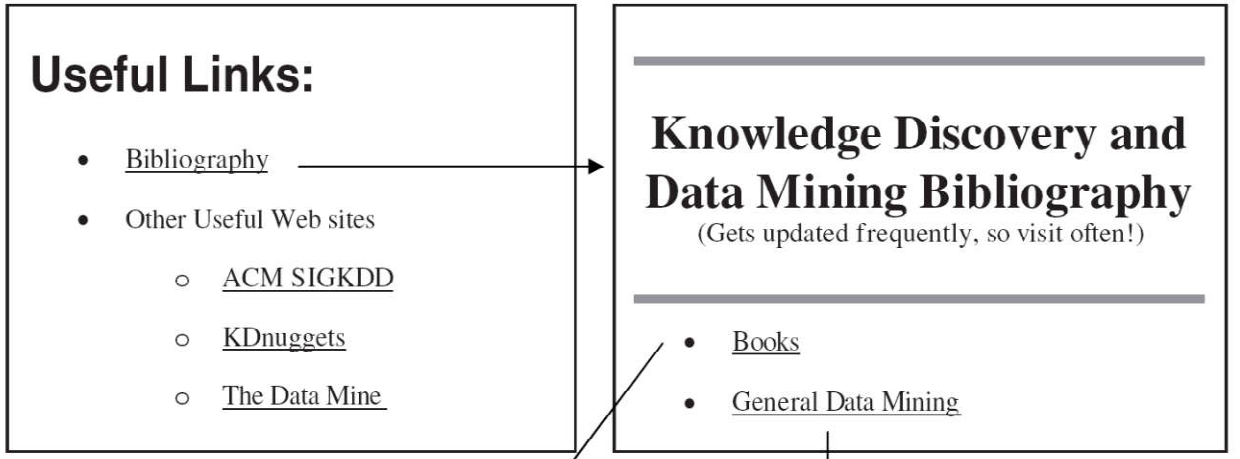
ارتباط بین چندنفر یا
دوستی چندنفر
هرالمان مرتبط با یک
موجودیت
شبکه های اجتماعی
صفحات اینترنتی

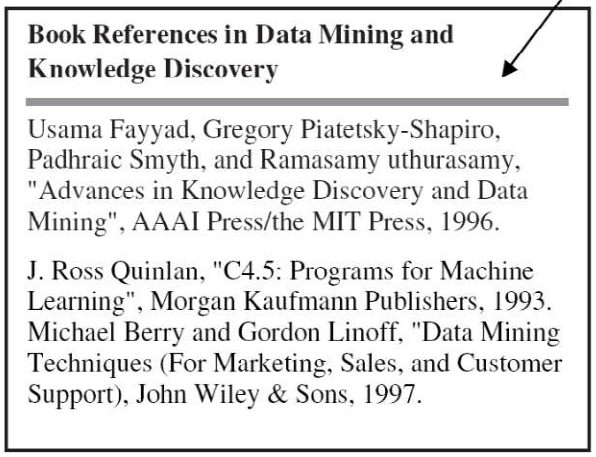
- Examples: Generic graph, a molecule, and webpages

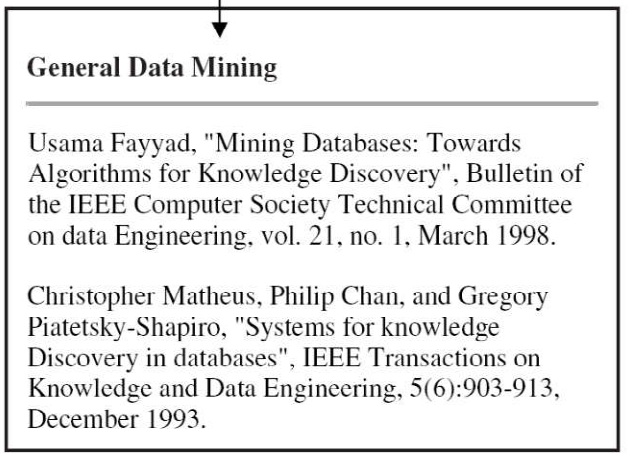




Benzene Molecule: C6H6







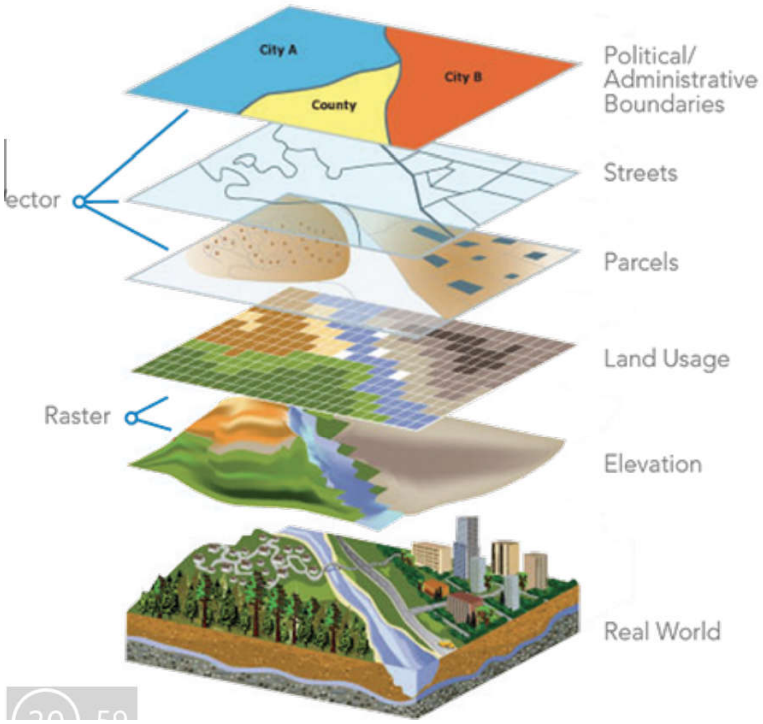
Spatial/Image Data

دیتاهای فضایی
تشخیص چهره یا
شناسایی مکان
ر تصاویر ماهواره ای
با چندین لایه از تصویر
کار داریم

- Spatio-Temporal Data

داده های مکانی-زمانی

❑ Maps

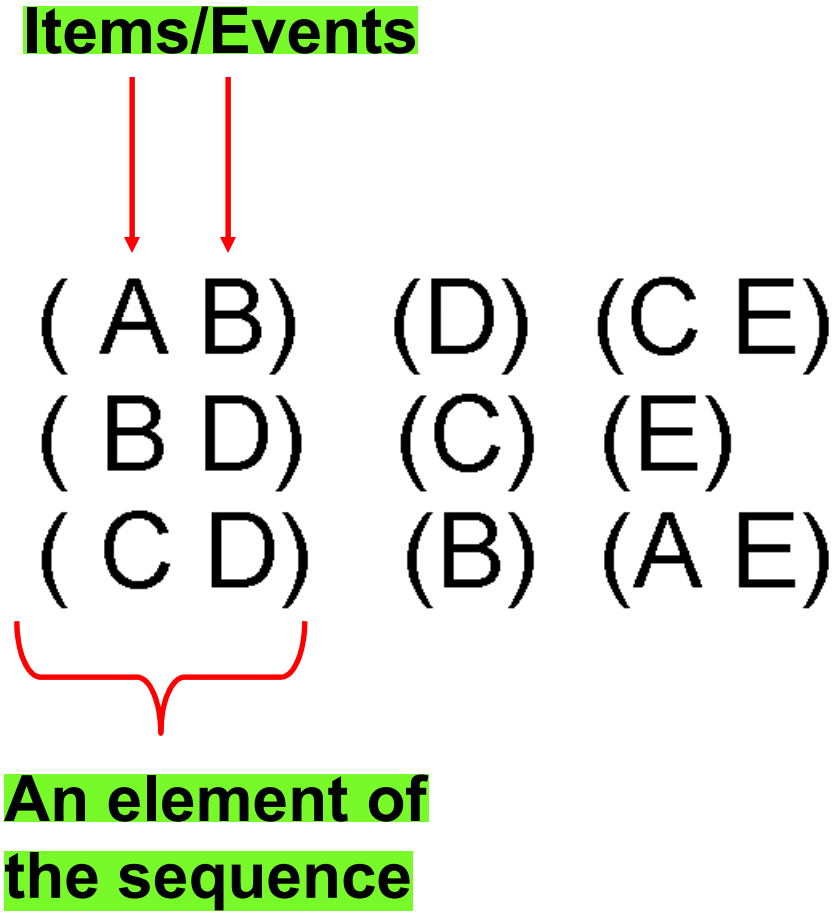


❑ Images

Ordered Data

دیتاهایی که ترتیب دارند برامون مهمه یه فردی که وارد فروشگاه میشه اول چی رو میخره بعد چی رو توالی کالاهای خریداری شده مهمه

- Sequences of transactions



Ordered Data

- Genomic sequence data

توالی ژن ها که براساس ترتیب یه اطلاعاتی به ما میده

```
GGTTCCGCCTTCAGCCCCGCGCC
CGCAGGGCCCGCCCCGCGCCGTC
GAGAAGGGCCCGCCTGGCGGGCG
GGGGGAGGCGGGGCCGCCCGAGC
CCAACCGAGTCCGACCAGGTGCC
CCCTCTGCTCGGCCTAGACCTGA
GCTCATTAGGCGGCAGCGGACAG
GCCAAGTAGAACACGCGAAGCGC
TGGGCTGCCTGCTGCGACCAGGG
```

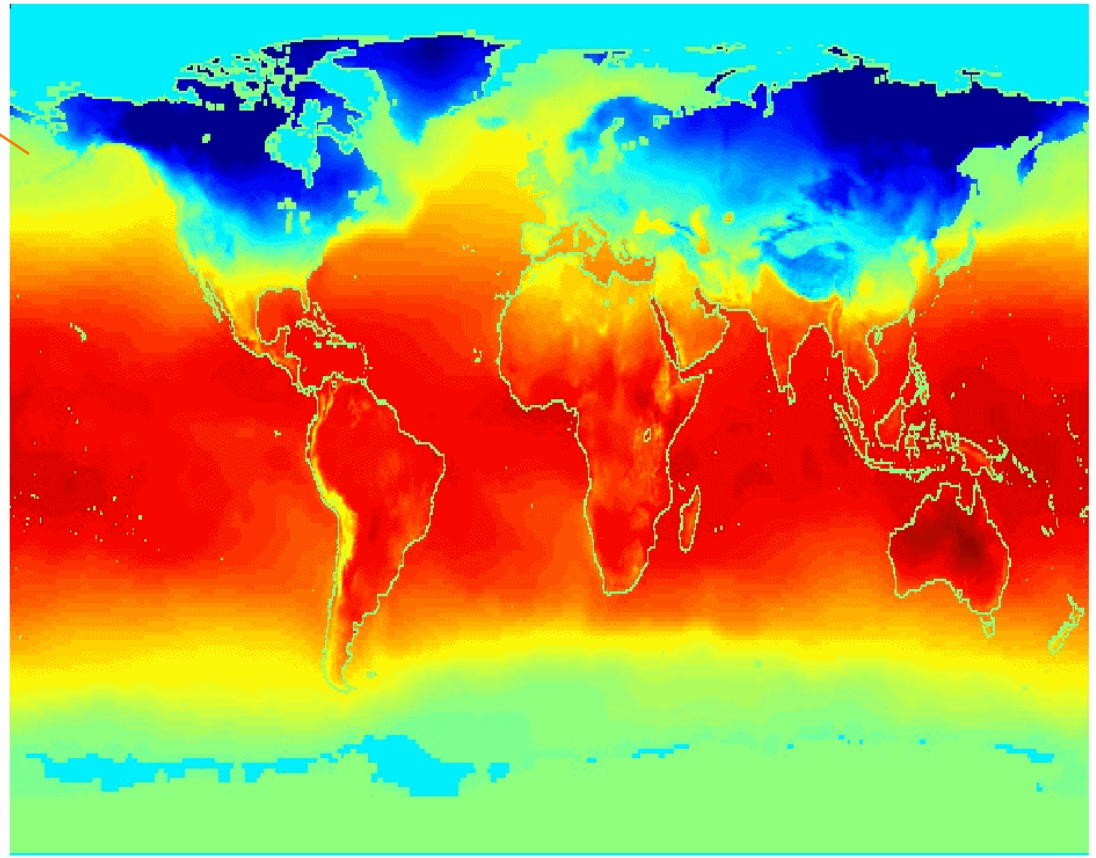
Ordered Data

- Spatio-Temporal Data

تغییرات دما
علاوه بر تغییرات زمانی
که دارند در فضا پراکنده
هم هستند
از اطلاعات همسایه ها
میتونیم استفاده کنیم

ارتباط مکان ها یا
همسایگیشون اهمیت داره

Jan



Average Monthly Temperature of land and ocean

BASIC STATISTICAL DESCRIPTIONS OF DATA

توصیف اماری از دیتا

Basic Statistical Descriptions of Data

تمایل داده ها به چه سمتی است؟ پراکندگی داده ها
چطوریه؟ چطوری توزیع شدن؟
برای اندازه گیری اینها یه سری مقیاس داریم
چارک و واریانس میانه میانگین ...

- Motivation
  - To better understand the data: central tendency, variation and spread
- Data dispersion characteristics
  - median, max, min, quantiles, outliers, variance, etc.

مقادیری که خیلی پراکنده
هستند و از اکثریت
فاصله دارند

- Numerical dimensions correspond to sorted intervals
  - Data dispersion: analyzed with multiple granularities of precision
  - Boxplot or quantile analysis on sorted intervals
- Dispersion analysis on computed measures
  - Folding measures into numerical dimensions
  - Boxplot or quantile analysis on the transformed cube

Measuring the Central Tendency

- Mean (algebraic measure) (sample vs. population):

میانگین

میانگین سمپل یا نمونه

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

سایز سمپل یا نمونه

میانگین جمعیت یا پاپیولیشن

$$\mu = \frac{\sum x}{N}$$

سایز جمعیت

Note: *n* is sample size and *N* is population size.

- Weighted arithmetic mean:

میانگین وزن دار

مثل محاسبه ی معدل کل که ضریب هر درس درش ضرب میشه.

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

جمع وزن دار ایتم ها یعنی وزن هرمتغیری را هم درنظر میگیریم.

جمع وزن ها یا جمع ضرایب هر متغیر

- Trimmed mean: chopping extreme values?(2%)

۲درصد مقادیر خیلی بالا و خیلی پایین را در میانگین درنظر نگیریم

Measuring the Central Tendency

- Median: میانه
  - Middle value if odd number of values, or average of the middle two values otherwise
  - Estimated by interpolation (for *grouped data*):

age	*frequency*
1–5	200
6–15	450
16–20	300
21–50	1500
51–80	700
81–110	44

درون یابی

میانه توی این رنج است.

Approximate median

تخمینی از میانه

Sum before the median interval

Interval width ($L_2 - L_1$)

Low interval limit

$$median = L_1 + \left(\frac{n/2 - (\sum freq)_l}{freq_{median}}\right) width$$

Using Equation (2.3), we have $L_1 = 20$, $N = 3194$, $(\sum freq)_l = 950$, $freq_{median} = 1500$, $width = 30$, $median = 32.94$ years.

200 + 450 + 300+ 1500 + 700 + 44 = 3149

200 + 450 + 300 = 950

50 - 20 = 30

20 + (((3194/2) - 950)/ 1500 *30) = 32.94

Measuring the Central Tendency

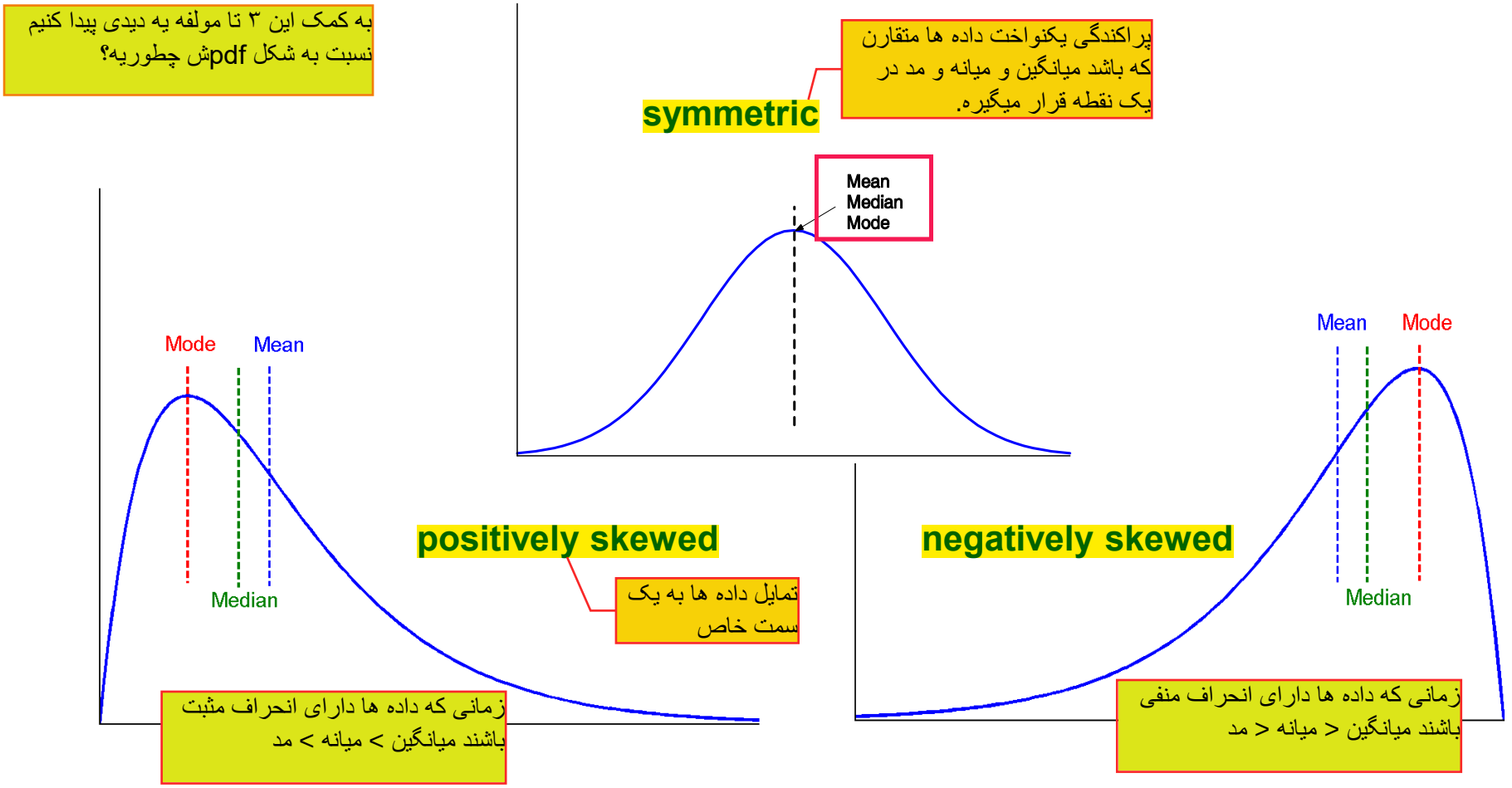
- Mode ← بیشترین تکرار
  - Value that occurs most frequently in the data
  - Unimodal, bimodal, trimodal
  - Empirical formula:

فرمول تجربی ←

$$mean - mode = 3 \times (mean - median)$$

Symmetric vs. Skewed Data

- Median, mean and mode of symmetric, positively and negatively skewed data



پراکندگی

Measuring the Dispersion of Data

- Variance and standard deviation (*sample: s, population: σ*)
  - **Variance**: (algebraic, scalable computation)

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{n}(x_i - \mu)^2 = \frac{1}{N}\sum_{i=1}^{n} x_i^{\,2} - \mu^2$$

واریانس جمعیت

میانگین جمعیت (کل داده ها)

شاخص های پراکندگی
واریانس و انحراف معیار
ز خود واریانس وقتی استفاده میکنیم که
میانگین کل جمعیت رو داشته باشیم

Standard deviation *s (or σ)* is the square root of variance s^2 *(or* σ^2*)*

sample variance
یه نمونه داریم میخایم
راجع به کل جمعیت
تخمین بزنیم

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{1}{n-1}\left[\sum_{i=1}^{n} x_i^{\,2} - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)^2\right]$$

واریانس سمپل یا نمونه
که در مخرجش به جای
n باید n-1بگذاریم.

میانگین تخمینی از داده ها

Measuring the Dispersion of Data

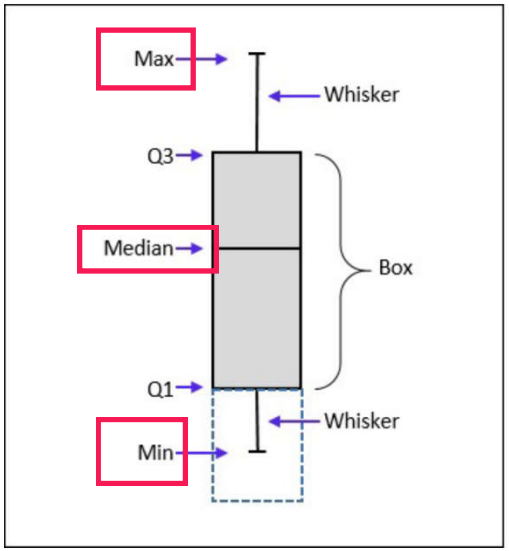
- Quartiles, outliers and boxplots
  - **Quartiles**: Q_1 (25^{th} percentile), Q_3 (75^{th} percentile)
  - **Inter-quartile range**: $IQR = Q_3 - Q_1$
  - **Five number summary**: min, Q_1, median, Q_3, max
  - **Boxplot**: ends of the box are the quartiles; median is marked; add whiskers, and plot outliers individually
  - Whiskers: two lines outside the box extended to Minimum and Maximum
  - **Outlier**: usually, a value higher/lower than 1.

چارک اول و دوم و..

هرچی فاصله این دوتا زیاد باشه یعنی داده هامون پراکنده تره

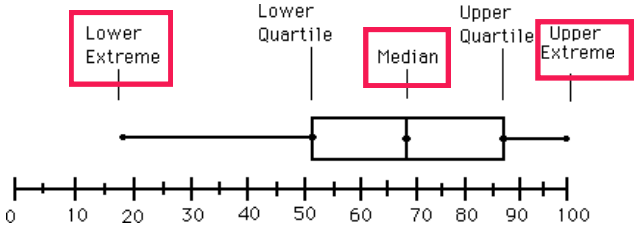
رویکرد ۵ شماره ای

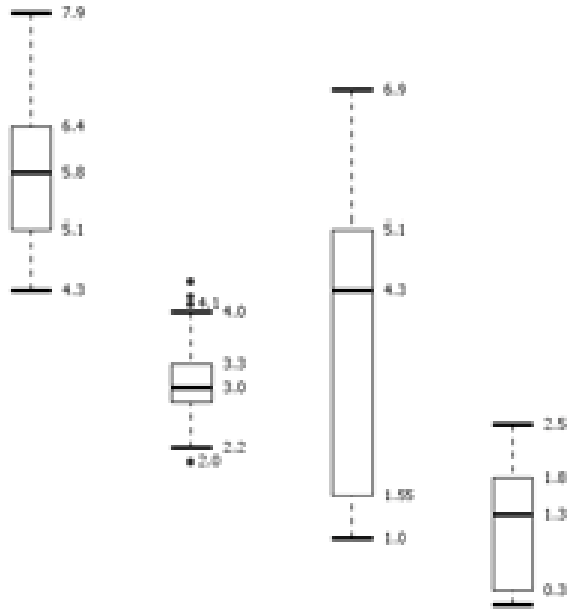
مقادیری که یک و نیم برابر فاصله ی IQR هستند



Boxplot Analysis

- **Five-number summary** of a distribution
  - Minimum, Q1, Median, Q3, Maximum
- **Boxplot**
  - Data is represented with a box
  - The ends of the box are at the first and third quartiles, i.e., the height of the box is IQR
  - The median is marked by a line within the box
  - Whiskers: two lines outside the box extended to Minimum and Maximum
  - Outliers: points beyond a specified outlier threshold, plotted individually





Suppose that the data for analysis includes the attribute *age*. The *age* values for the data tuples are (in increasing order) 13, 15, 16, 16, 19, 20, 20, 21, 22, 22, 25, 25, 25, 25, 30, 33, 33, 35, 35, 35, 35, 36, 40, 45, 46, 52, 70.

(a) What is the *mean* of the data? What is the *median*?

(b) What is the *mode* of the data? Comment on the data's modality (i.e., bimodal, trimodal, etc.).

(c) What is the *midrange* of the data?

(d) Can you find (roughly) the first quartile ($Q1$) and the third quartile ($Q3$) of the data?

(e) Give the *five-number summary* of the data.

(f) Show a *boxplot* of the data.

(g) How is a *quantile-quantile plot* different from a *quantile plot*?

Answer:

(a) What is the *mean* of the data? What is the *median*?

The (arithmetic) mean of the data is: $\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i = 809/27 = 30$ (Equation 2.1). The median (middle value of the ordered set, as the number of values in the set is odd) of the data is: 25.

(b) What is the *mode* of the data? Comment on the data's modality (i.e., bimodal, trimodal, etc.).

This data set has two values that occur with the same highest frequency and is, therefore, bimodal. The modes (values occurring with the greatest frequency) of the data are 25 and 35.

(c) What is the *midrange* of the data?

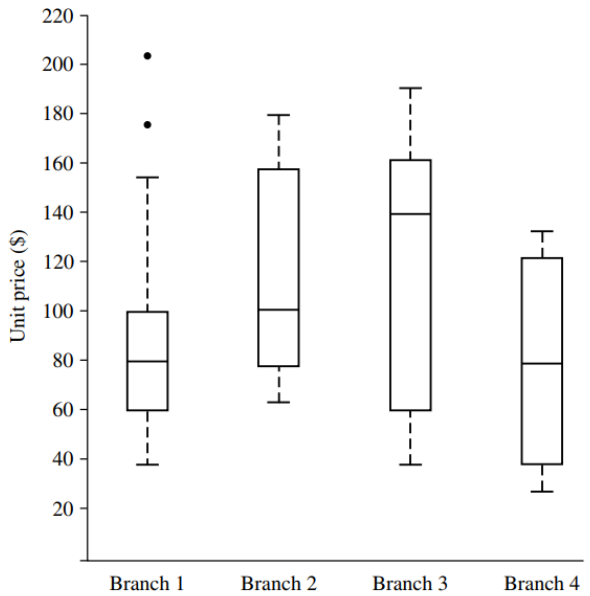
The midrange (average of the largest and smallest values in the data set) of the data is: $(70+13)/2 = 41.5$

(d) Can you find (roughly) the first quartile ($Q1$) and the third quartile ($Q3$) of the data?

The first quartile (corresponding to the 25th percentile) of the data is: 20. The third quartile (corresponding to the 75th percentile) of the data is: 35.

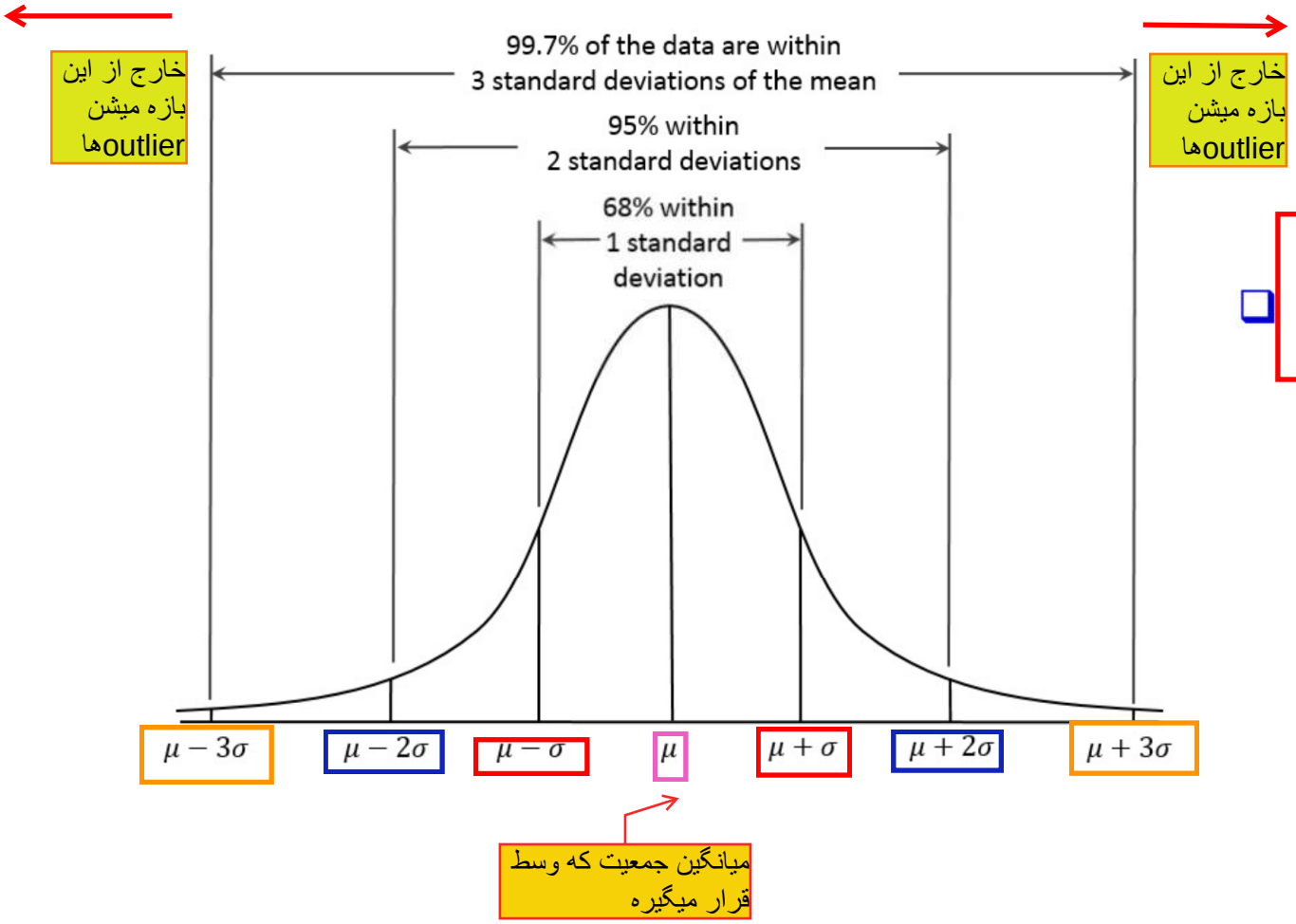
(e) Give the *five-number summary* of the data.

The five number summary of a distribution consists of the minimum value, first quartile, median value, third quartile, and maximum value. It provides a good summary of the shape of the distribution and for this data is: 13, 20, 25, 35, 70.



Properties of Normal Distribution Curve

ویژگی های توزیع نرمال
توزیع متقارن مثل دما



میانگین جمعیت

- Z-score: $z = \frac{x - \mu}{\sigma}$

انحراف معیار جمعیت

اگه متغیرتصادفیمون نرمال باشه میتونیم با یه زبان بهتری توصیفش کنیم مثلا متغیرتصادفی مون قد ادم ها توی دانشگاه باشه
یه متغیر دیگه هم داریم وزن ادم ها
یه جاهایی میخایم این دوتا رو باهم مقایسه کنیم
میتونیم این دوتا متغیر را نرمال کنیم تا بهتر بشه مقایسه کرد

z-score normalization

The range is $[(old\_min - mean)/stddev, (old\_max - mean)/stddev]$. In general the range for all possible data sets is $(-\infty, +\infty)$.

Example 1:
Suppose a student scores 80 out of 100 on a test, and the mean score for the class is 75 with a standard deviation of 5. To calculate the z-score for this student's score:

z = (80 - 75) / 5
z = 1

Therefore, the student's score is one standard deviation above the mean.

Example 2:
A company has a sales team with an average monthly sales of $10,000 and a standard deviation of $2,000. If one salesperson had sales of $14,000 in a month, what is their z-score?

z = (14,000 - 10,000) / 2,000
z = 2

This means that the salesperson's monthly sales were two standard deviations above the mean.

Example 3:
A researcher wants to know if there is a significant difference in height between men and women. She measures the heights of a random sample of men and women and finds that the mean height for men is 70 inches with a standard deviation of 3 inches, while the mean height for women is 65 inches with a standard deviation of 2 inches. If a man is randomly selected from this sample and his height is recorded as 73 inches:

z = (73 -70) /3
z =1

This means that this man's height is one standard deviation above the mean height for men.

Example 4:
A teacher wants to compare her students' scores on two different tests. The first test has an average score of 80 with a standard deviation of 10, while the second test has an average score of 75 with a standard deviation of 8. If one student scored an 85 on both tests:

For Test #1:
z = (85 -80) /10
z = 0.5

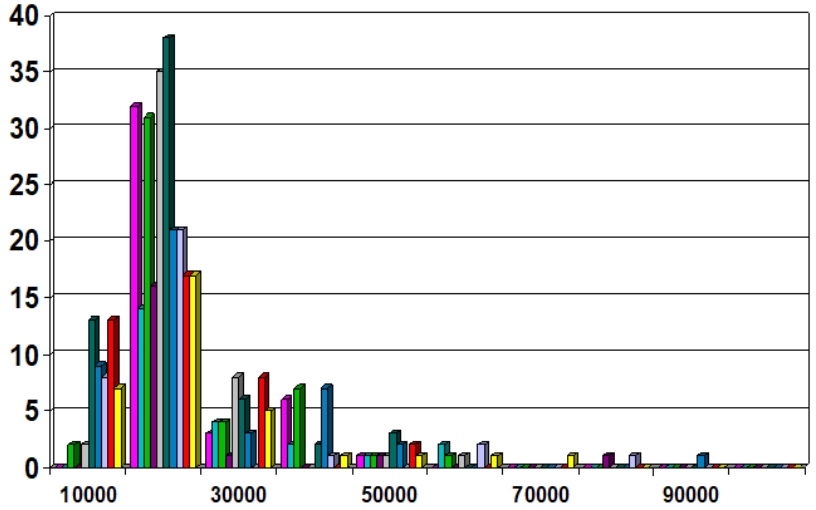
For Test #2:
z = (85 -75) /8
z = 1.25

This means that the student's score on Test #1 was half a standard deviation above the mean, while their score on Test #2 was 1.25 standard deviations above the mean.

یه نموداری که تعداد تکرار مقادیر مختلف یه متغیر رو نشون میده
مثلا متغیرمون از ۱۰ هزاره تا ۹۰ هزار
میایم ۱۰۰ تا ۱۰۰ تا جداشون میکنیم و هربار
توی این بازه ها یه متغیری رخداد یکی اضافه میشه به تعدادشون
درباره ی توزیع داده ها یه توضیحی میده

Histogram Analysis

- Histogram: Graph display of tabulated frequencies, shown as bars
- It shows what proportion of cases fall into each of several categories
- Differs from a bar chart in that it is the *area* of the bar that denotes the value, not the height as in bar charts, a crucial distinction when the categories are not of uniform width
- The categories are usually specified as non-overlapping intervals of some variable. The categories (bars) must be adjacent



تفاوت نمودار هیستوگرام و بار چیه؟

1. Data type: A bar chart is used to represent categorical data, while a histogram is used to represent continuous data.

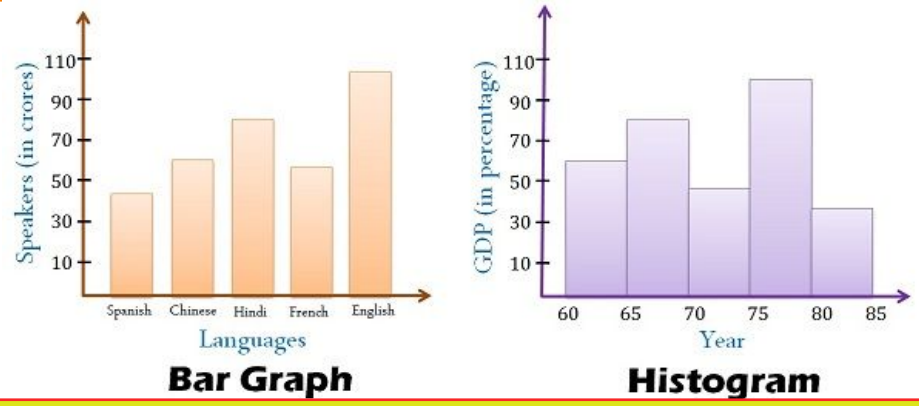
2. X-axis: In a bar chart, the x-axis represents the categories being compared, while in a histogram, the x-axis represents the range of values being measured.

3. Y-axis: In both charts, the y-axis represents the frequency or count of each category or value.

4. Bar width: In a bar chart, there is typically space between each bar, while in a histogram, the bars are usually touching or overlapping to show continuity of data.

5. Interpretation: A bar chart is useful for comparing discrete categories or groups, while a histogram is useful for showing patterns and distributions within continuous data.

Overall, the main difference between a bar chart and a histogram is that a bar chart is used for categorical data with distinct categories, while a histogram is used for continuous data with ranges of values.



Example 1: Odd Number of Data Points
Suppose we have the following dataset: 4, 7, 8, 10, 12, 16, 20. To find Q1 and Q3:

1. Arrange the data in ascending order: 4, 7, 8, 10, 12, 16, 20.
2. Find the median of the dataset. In this case, the median is 10.
3. Split the dataset into two halves: 4, 7, 8 and 12, 16, 20.
4. Find the median of the lower half of the dataset. In this case, the median is 7.
5. Find the median of the upper half of the dataset. In this case, the median is 16.
6. The first quartile (Q1) is the median of the lower half of the dataset, which is 7.
7. The third quartile (Q3) is the median of the upper half of the dataset, which is 16.

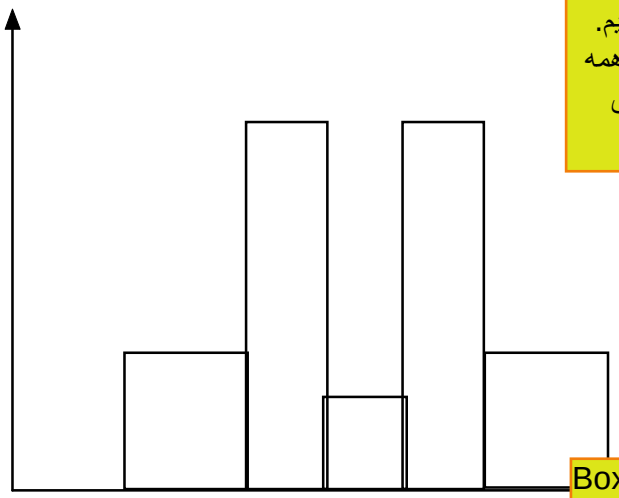
Therefore, Q1 = 7 and Q3 = 16.
Example 2: Even Number of Data Points
Suppose we have the following dataset: 3, 8, 9, 11, 16, 17. To find Q1 and Q3:

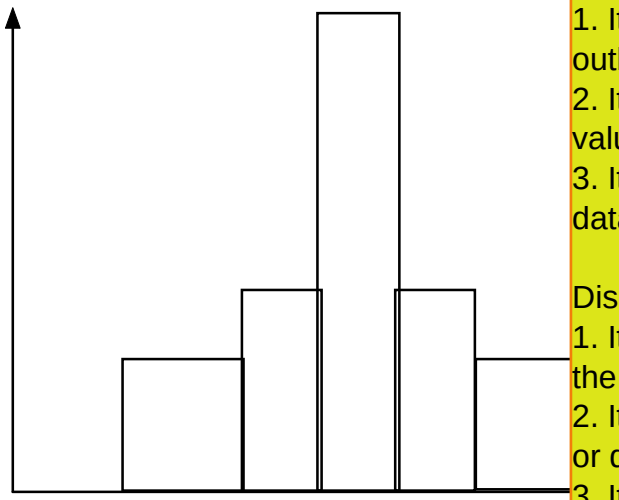
1. Arrange the data in ascending order: 3, 8, 9, 11, 16, 17.
2. Find the median of the dataset. In this case, the median is the average of the two middle values, which is (9 + 11) / 2 = 10.
3. Split the dataset into two halves: 3, 8, 9 and 11, 16, 17.
4. Find the median of the lower half of the dataset. In this case, the median is (8 + 9) / 2 = 8.5.
5. Find the median of the upper half of the dataset. In this case, the median is (16 + 17) / 2 = 16.5.
6. The first quartile (Q1) is the median of the lower half of the dataset, which is 8.5.
7. The third quartile (Q3) is the median of the upper half of the dataset, which is 16.5.

Therefore, Q1 = 8.5 and Q3 = 16.5.

Histograms Often Tell More than Boxplots

توی هیستوگرام راجع به کل طیف مقادیری که متغیر میگیره دید داریم. باکس پلات شکل اول و دوم مثل همه .ولی توزیع داده هاشون خیلی فرق داره





The two histograms shown in the left may have the same boxplot representation

- **The same values for: min, Q1, median, Q3, max**
- **But they have rather different data distributions**

Boxplot Chart:

Advantages:
1. It shows the median, quartiles, and outliers in a dataset.
2. It is robust to outliers and extreme values.
3. It can be used to compare multiple datasets side by side.

Disadvantages:
1. It does not show the exact distribution of the data.
2. It may not be suitable for small datasets or datasets with few observations.
3. It may not be as intuitive as a histogram for some people.

Histogram Chart:

Advantages:
1. It shows the frequency distribution of a continuous variable.
2. It is easy to interpret and understand.
3. It can show the shape, center, and spread of the data.
4. It can be used to identify outliers.

Disadvantages:
1. It can be affected by the choice of bin size.
2. It may not be suitable for small datasets.
3. It does not show individual data points.

Graphic Displays of Basic Statistical Descriptions

- **Boxplot**: graphic display of five-number summary
- **Histogram**: x-axis are values, y-axis repres. frequencies
- **Quantile plot**: each value x_i is paired with f_i indicating that approximately 100 f_i % of data are $\leq x_i$
- **Quantile-quantile (q-q) plot**: graphs the quantiles of one univariant distribution against the corresponding quantiles of another
- **Scatter plot**: each pair of values is a pair of coordinates and plotted as points in the plane

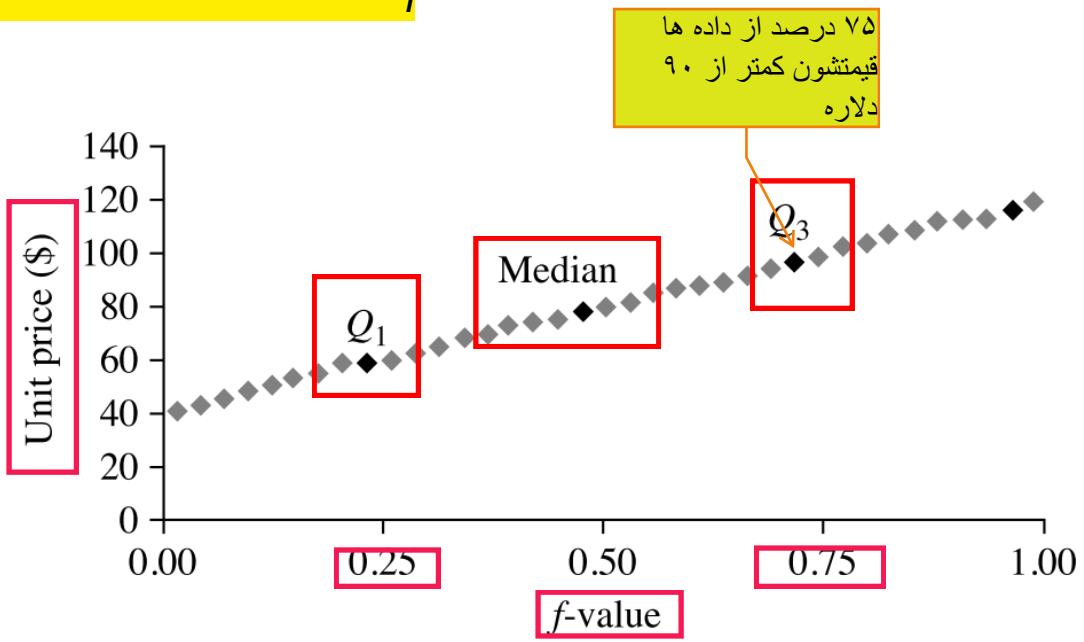
تعداد داده ها هرچی که باشه میبریم به فضای ۱۰۰ درصد و چارک های اول و سومش رو حساب میکنیم مثلا اگه ۱۲۰تا داده داشته باشیم باز هم میبریم به ۱۰۰

Quantile Plot

- Displays all of the data (allowing the user to assess both the overall behavior and unusual occurrences)
- Plots **quantile** information
  - For a data x_i data sorted in increasing order, f_i indicates that approximately 100 f_i% of the data are below or equal to the value x_i

Table 2.1 A Set of Unit Price Data for Items Sold at a Branch of *AllElectronics*

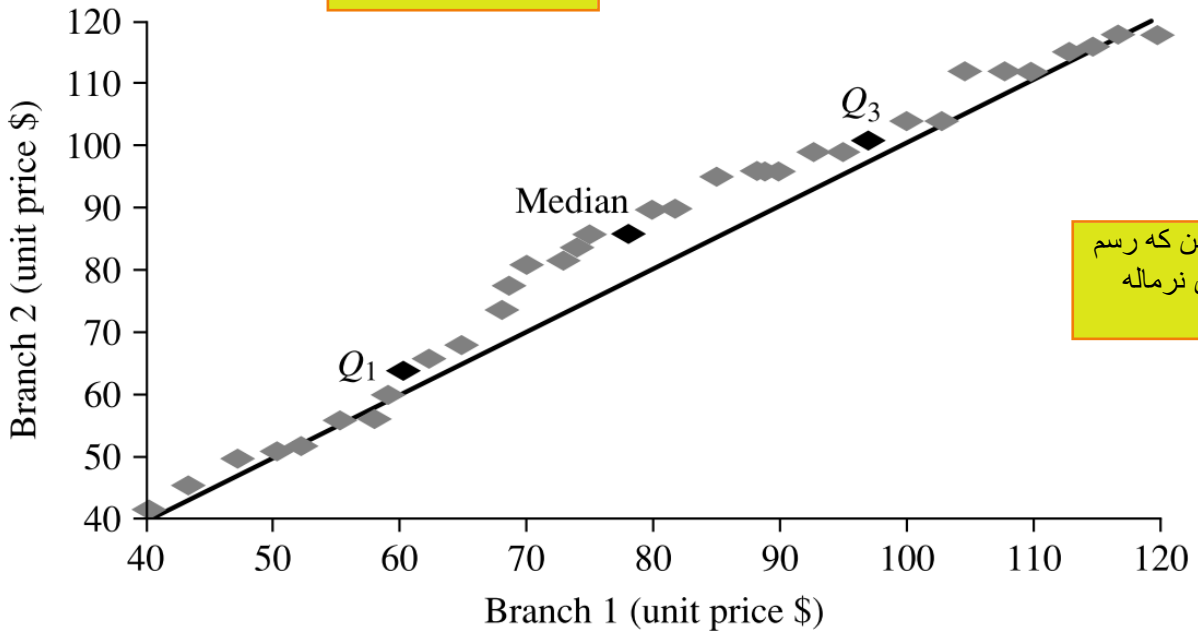
Unit price ($)	*Count of items sold*
40	275
43	300
47	250
—	—
74	360
75	515
78	540
—	—
115	320
117	270
120	350



Quantile-Quantile (Q-Q) Plot

- Graphs the quantiles of one univariate distribution against the corresponding quantiles of another
- View: Is there is a shift in going from one distribution to another?
- Example shows unit price of items sold at Branch 1 vs. Branch 2 for each quantile. Unit prices of items sold at Branch 1 tend to be lower than those at Branch 2.

مقایسه ی دوتا متغیر باهم
قدادم ها در دانشگاه الف
با دانشگاه ب



اگه داده ها منطبق به خطی بشن که رسم
کردیم میگیم توزیع داده هامون نرماله

How is a *quantile-quantile plot* different from a *quantile plot*?

A quantile plot is a graphical method used to show the approximate percentage of values below or equal to the independent variable in a univariate distribution. Thus, it displays quantile information for all the data, where the values measured for the independent variable are plotted against their corresponding quantile.

A quantile-quantile plot however, graphs the quantiles of one univariate distribution against the corresponding quantiles of another univariate distribution. Both axes display the range of values measured for their corresponding distribution, and points are plotted that correspond to the quantile values of the two distributions. A line $(y = x)$ can be added to the graph along with points representing where the first, second and third quantiles lie to increase the graph's informational value. Points that lie above such a line indicate a correspondingly higher value for the distribution plotted on the y-axis than for the distribution plotted on the x-axis at the same quantile. The opposite effect is true for points lying below this line.

Suppose a hospital tested the age and body fat data for 18 randomly selected adults with the following result

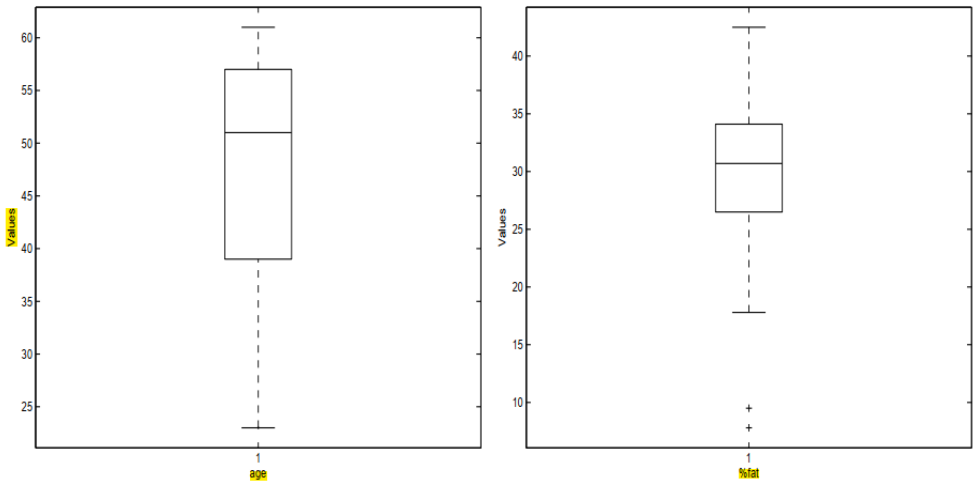
age	23	23	27	27	39	41	47	49	50
%fat	9.5	26.5	7.8	17.8	31.4	25.9	27.4	27.2	31.2
age	52	54	54	56	57	58	58	60	61
%fat	34.6	42.5	28.8	33.4	30.2	34.1	32.9	41.2	35.7

(a) Calculate the mean, median and standard deviation of *age* and *%fat*.

For the variable *age* the mean is 46.44, the median is 51, and the standard deviation is 12.85. For the variable *%fat* the mean is 28.78, the median is 30.7, and the standard deviation is 8.99.

(b) Draw the boxplots for *age* and *%fat*.

See Figure 2.1.



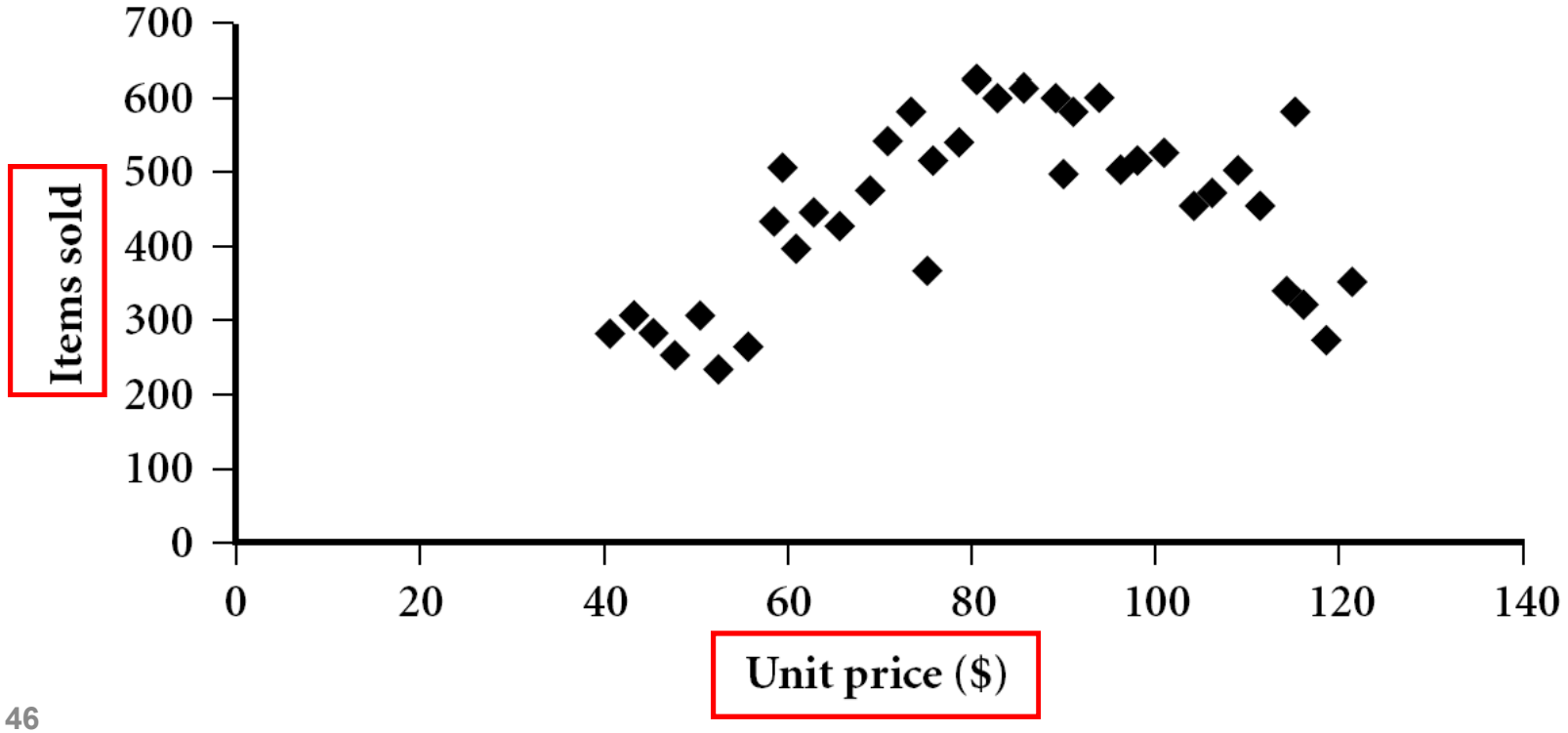
(d) Normalize the two variables based on *z-score normalization*.

age	23	23	27	27	39	41	47	49	50
z-age	-1.83	-1.83	-1.51	-1.51	-0.58	-0.42	0.04	0.20	0.28
%fat	9.5	26.5	7.8	17.8	31.4	25.9	27.4	27.2	31.2
z-%fat	-2.14	-0.25	-2.33	-1.22	0.29	-0.32	-0.15	-0.18	0.27
age	52	54	54	56	57	58	58	60	61
z-age	0.43	0.59	0.59	0.74	0.82	0.90	0.90	1.06	1.13
%fat	34.6	42.5	28.8	33.4	30.2	34.1	32.9	41.2	35.7
z-%fat	0.65	1.53	0.0	0.51	0.16	0.59	0.46	1.38	0.77

Scatter plot

داده های دو متغیره

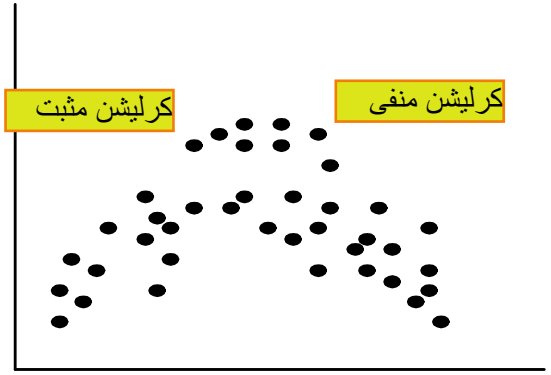
- Provides a first look at bivariate data to see clusters of points, outliers, etc
- Each pair of values is treated as a pair of coordinates and plotted as points in the plane

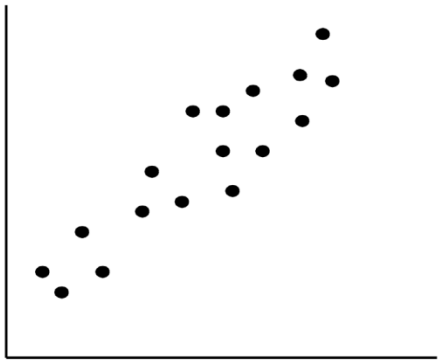


Positively and Negatively Correlated Data

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

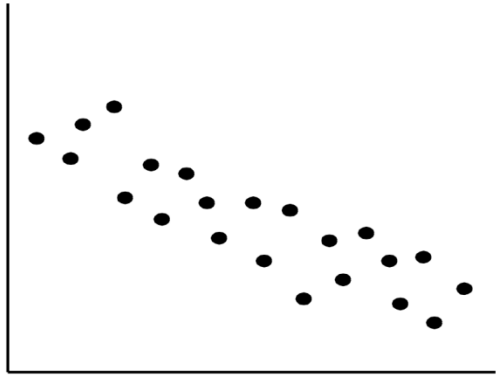
سن و قد تا بیست سالگی رابطه مثبت داره
از بیست سالگی به بعد مثلا ارتباطی وجود
نداره یعنی سن بالا میره ولی قد ثابت میمونه





Positively correlated

شیب خط مثبت است



Negatively correlated

شیب خط منفی است

Uncorrelated Data

یه شکل مستطیلی پیدا میکنه

زمانی که اتریبیوت های داده هامون هیچ ربطی به هم نداشته باشن
مثلا اگه بگیم مقدار اتریبوت اول شد ۱۰ لزوما نمیتونیم بگیم مقدار اتریبیوت دوم هم میشه ۱۰ مثلا ممکنه بشه ۲۰