Chapter 2. Data Preprocessing: Data Description

Meng Jiang

Data Science

Inferring College CS Student's Early Career Salary based on College's **Attributes/Features**

	A	B	C	D
1	http://csrankings.org/			
2	Rank	Institution	Count	Faculty
3	1	Carnegie Mellon University	18.5	150
4	2	Massachusetts Institute of Technology	12.2	82
5	3	Stanford University	10.9	54
6	3	University of California - Berkeley	10.9	81
7	5	Univ. of Illinois at Urbana-Champaign	9.9	84
8	6	Cornell University	8.7	68
9	7	University of Michigan	8.6	63
10	8	University of Washington	8.3	56
11	9	University of California - San Diego	6.9	54
12	10	Georgia Institute of Technology	6.8	75
13	11	University of Wisconsin - Madison	5.9	47
14	12	Columbia University	5.8	47
15	13	University of Pennsylvania	5.6	46
16	14	University of Southern California	5.5	49
17	15	Princeton University	5.3	51
18	16	University of Texas at Austin	5.2	42
19	16	University of Maryland -		
20	18	University of California -		
21	19	Northeastern University		
22	19	Purdue University	4.8	51
23	21	University of Massachusetts Amherst	4.7	50
24	22	New York University	4.5	47
25	23	Harvard University	4.2	29
26	23	University of California - Irvine	4.2	54
27	25	Rutgers University	3.9	43
28	26	University of Cali		
29	27	University of Uta		
30	27	Pennsylvania State University	3.4	31
31	29	Stony Brook University	3.3	41
32	30	University of California - Davis	3.2	29

126 institutions

http://csrankings.org/

	A	B	C
1	ihttps://www.payscale.com/college-salary-report/best-schools-by-majors/computer-science		
2	School	Name	Early Career Pay
3	1	Stanford University	$101,000
4	2	University of Pennsylvania	$90,500
5	3	Dartmouth College	$94,700
6	4	Princeton University	$93,400
7	5	University of California - Berkeley	$97,000
8	6	Yale University	$98,000
9	7	Columbia University	$86,400
10	8	Cornell University - Ithaca, NY	$86,500
11	9	Carnegie Mellon University (CMU)	$92,200
12	10	Duke University	$80,500
13	11	University of California - San Diego (UCSD)	$84,500
14	12	Harvard University	$85,300
15	13	University of Washington (UW) - Main Campus	$79,600
16	14	Massachusetts Institute of Technology (MIT)	$94,100
17	15 (tie)	Brown University	$84,800
18	15 (tie)	Lehigh University	
19	17	University of California - Santa Barbara (UCSB)	
20	18	University of California - Santa Cruz (UCSC)	$77,400
21	19	Rice University	$81,100
22	20	New York University (NYU)	$78,200
23	21	University of California - Irvine (UCI)	$74,100
24	22	Stevens Institute of Technology	$78,800
25	23 (tie)	California Polytec	0
26	23 (tie)	San Jose State Un	0
27	25	University of Virgi	0
28	26	University of Calif	0
29	27	Tufts University	0
30	28	Boston College	$75,600

466 institutions

https://www.payscale.com/college-salary-report/best-schools-by-majors/computer-science

Other College's Attributes/Features

2017 First Year Applicant Pool

19,566 APPLICATIONS

3,700 ADMITTED

2,050 ENROLLED

2017 Enrolled Students Profile

43% RANKED IN THE TOP 2% OF THEIR CLASS

89% RANKED IN THE TOP 10% OF THEIR CLASS

1390-1530 SAT (MID 50%)

33-35 ACT (MID 50%)

51% MALE

49% FEMALE

24% ALUMNI CHILDREN

6% INTERNATIONAL/OUTSIDE OF THE U.S.

7% FIRST GENERATION

81% CATHOLIC

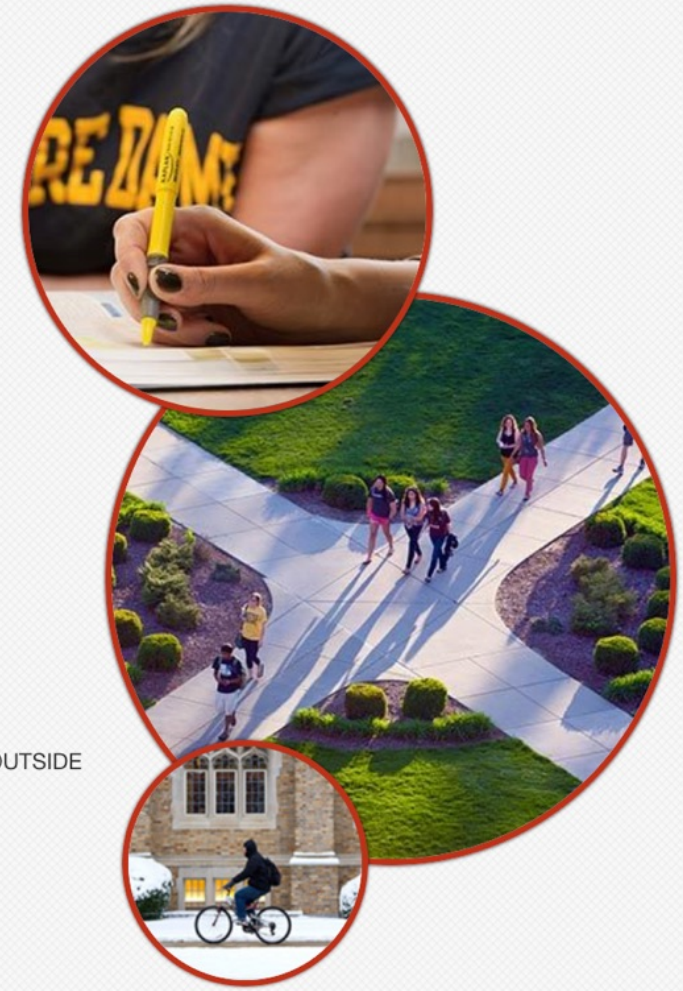
16% SIBLINGS ATTEND OR GRADUATED FROM ND

26% U.S. STUDENTS OF COLOR

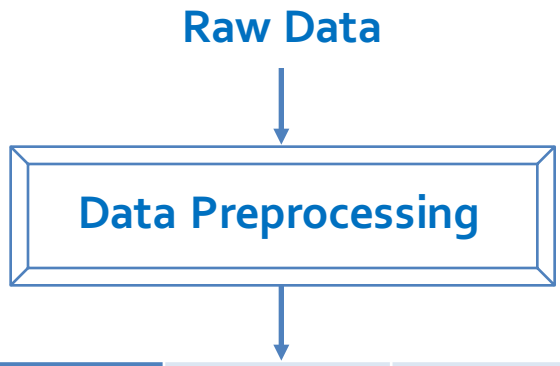
25% HEADED A STUDENT ORGANIZATION

42% CAPTAINS OF A VARSITY SPORT

47% INVOLVED IN MUSIC, DRAMA, FINE ARTS, OR DANCE



Data Preprocessing before Data Mining



	Ranking Score	Faculty Number	Color	Private?	…	Label
…						
…						
CMU	18.5	150	Red	Yes		$92,200
…						
Notre Dame			Gold/ Blue	Yes		?
…						?

Data Integration

	A	B	C	D
1	http://csrankings.org/			
2	Rank	Institution	Count	Faculty
3	1	Carnegie Mellon University	18.5	150
4	2	Massachusetts Institute of Technology	12.2	82
5	3	Stanford University	10.9	54
6	3	University of California - Berkeley	10.9	81
7	5	Univ. of Illinois at Urbana-Champaign	9.9	84
8	6	Cornell University	8.7	68
9	7	University of Michigan	8.6	63
10	8	University of Washington	8.3	56
11	9	University of California - San Diego	6.9	54
12	10	Georgia Institute of Technology	6.8	75
13	11	University of Wisconsin - Madison	5.9	47
14	12	Columbia University	5.8	47

	A	B	C
1	ihttps://www.payscale.com/college-salary-report/best-schools-by-majors/computer-science		
2	School	Name	Early Career Pay
3	1	Stanford University	$101,000
4	2	University of Pennsylvania	$90,500
5	3	Dartmouth College	$94,700
6	4	Princeton University	$93,400
7	5	University of California - Berkeley	$97,000
8	6	Yale University	$98,000
9	7	Columbia University	$86,400
10	8	Cornell University - Ithaca, NY	$86,500
11	9	Carnegie Mellon University (CMU)	$92,200
12	10	Duke University	$80,500
13	11	University of California - San Diego (UCSD)	$84,500

Data Cleaning

Carnegie Mellon University	18.5	150
Carnegie Mellon University (CMU)		

Cornell University	8.7	68
Cornell University - Ithaca, NY		

115	Tulane University	1.1	4
115	Naval Postgraduate School	1.1	4
115	Northern Arizona University	1.1	5
115	University of Miami	1.1	4
115	Queen's University	1.1	1
115	OHSU	1.1	2
125	Auburn University	1	2
125	University of Vermont	1	3

Data Reduction

	F1	F2	…	F300
College 1				
…				
College 400				

	F1	…	**F10**
C1			
…			
C80			

Data Understanding before Data Preprocessing

- What are the data objects?
- What are the attributes/features?
- What are the attribute types?

attributes/features

data objects

	Ranking Score	Faculty Number	Color	Private?	…
…					
…					
CMU	18.5	150	Dark red	Yes	
…					

Questions You May Have before Data Preprocessing

- Is this college's **#faculty** incorrect/missing?
- Is it possible to match **college names** across data tables?

- What is the average **ranking score**? Which is the highest? Which is the lowest? Which college is the median one?
- What is the average **#faculty**? Which is the largest? Which is the smallest? Which college is the median one?
- Is there correlation between **ranking score** and **#faculty**? If they are strongly correlated, then…
- Is there correlation between **#faculty** and **salary label**? If they are strongly correlated, then…

Before Data Mining

(Chapter 2-3)

- Data Understanding
  - Data Description
  - Data Visualization
- Data Preprocessing
  - Data Cleaning and Integration
  - Data Reduction

Today: Data Description

- Understand what is data object and attribute/feature;
- **Understand different attribute types;**
- **Understand different data set types;**
- Describe basic statistical descriptions
  - Describe and calculate central tendency
    - **Mean, Median, Mode, Frequency, Percentiles**
    - Population and sample
  - Describe and calculate outlier-ness
    - **Variance, Standard Deviation, Z-score**
    - Biased/Unbiased sample variance
    - Z-score normalization vs min-max normalization

Data Object

- A **data object** represents an entity. Data sets are made up of **data objects**, also called *samples*, *examples*, *instances*, *data points*, *objects*, *tuples.*
- Examples:
  - Sales database: customers, store items, sales.
  - Medical database: patients, treatments.
  - University database: students, professors, courses.
- Data objects are described by **attributes.**

Attribute and Attribute Type

- **Attribute** (or feature, variable)
  - A data field, representing a characteristic or feature of a data object
- **Attribute type**
  - Nominal (e.g., "red", "blue")
  - Binary (e.g., {true, false})
  - Ordinal (e.g., {freshman, sophomore, junior, senior}, {small, medium, large}, {S, M, L, XL, XXL})
  - Numeric (e.g., #faculty)

	Ranking Score	Faculty Number	Color	Private?	…
…					
…					
CMU	18.5	150	Red	Yes	
…					

Hair color

Marital status

Occupation

Driver license ID

Zip code

Gender

Medical test

Course grade

Army ranking

Age

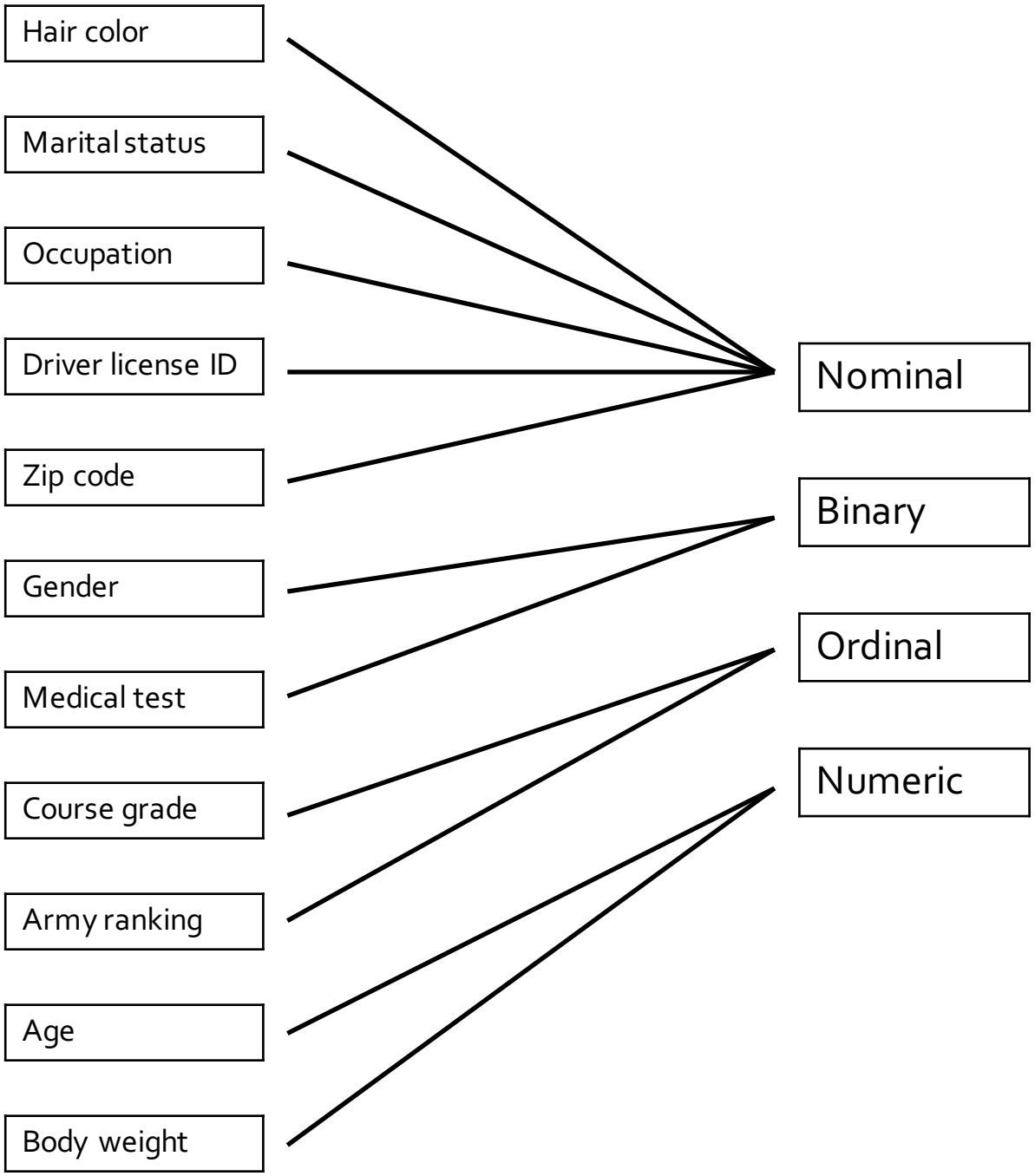
Body weight

Nominal

Binary

Ordinal

Numeric


Hair color
Marital status
Occupation
Driver license ID
Zip code
Gender
Medical test
Course grade
Army ranking
Age
Body weight
Nominal
Binary
Ordinal
Numeric

Types of Data Sets: (1) Record Data

- Relational records in relational tables: highly structured
  - Transaction data
  - Document data: Term-frequency matrix of text documents
  - …

HOME TEAM: Notre Dame 26-9

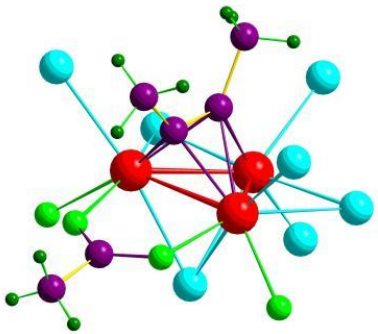
##	Player Name		TOT-FG FG-FGA	3-PT FG-FGA	FT-FTA	REBOUNDS OF	DE	TOT	PF	TP	A	TO	BLK	S	MIN
03	VJ Beachem	f	1-9	0-3	0-0	0	6	6	1	2	3	0	0	1	37
35	Bonzie Colson	f	6-13	0-1	6-10	2	5	7	2	18	2	0	2	1	31
00	Rex Pflueger	g	2-3	0-0	0-0	0	2	2	2	4	0	1	0	0	28
05	Matt Farrell	g	6-9	3-5	1-3	0	4	4	2	16	4	3	0	2	36
32	Steve Vasturia	g	3-12	1-2	3-4	3	5	8	0	10	1	0	0	0	37
01	Austin Torres		0-1	0-0	0-0	1	0	1	0	0	0	1	1	0	7
02	TJ Gibbs		0-1	0-0	2-2	0	2	2	1	2	0	0	0	0	13
04	Matt Ryan		2-3	0-0	2-2	0	2	2	0	6	0	0	0	0	9
23	Martinas Geben		1-1	0-0	0-0	1	0	1	1	2	0	1	0	0	2
	TEAM					2	1	3							
	Totals		21-52	4-11	14-21	9	27	36	9	60	10	6	3	4	200

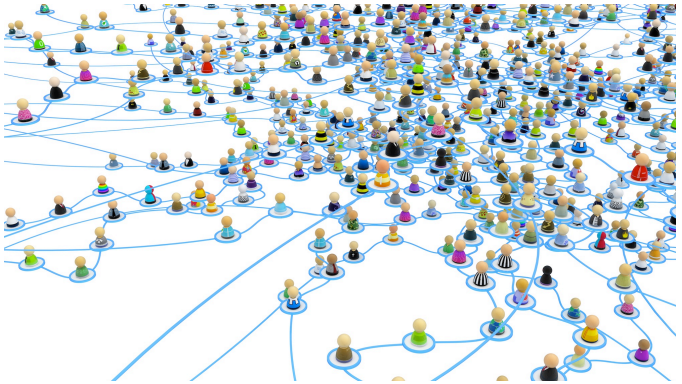
TOTAL FG% 1st Half:	14-30	46.7%	2nd Half:	7-22	31.8%	Game: 40.4% DEADB
3-Pt. FG% 1st Half:	2-5	40.0%	2nd Half:	2-6	33.3%	Game: 36.4% REBS
F Throw % 1st Half:	6-8	75.0%	2nd Half:	8-13	61.5%	Game: 66.7% 3

Types of Data Sets: (2) Graphs and Networks

- Transportation networks
- World Wide Web
- Molecular structures
- Social or information networks

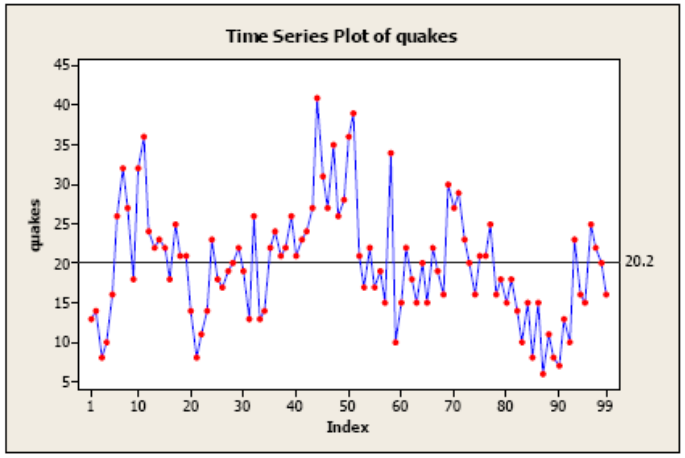
What are the data objects?
What are the attributes?





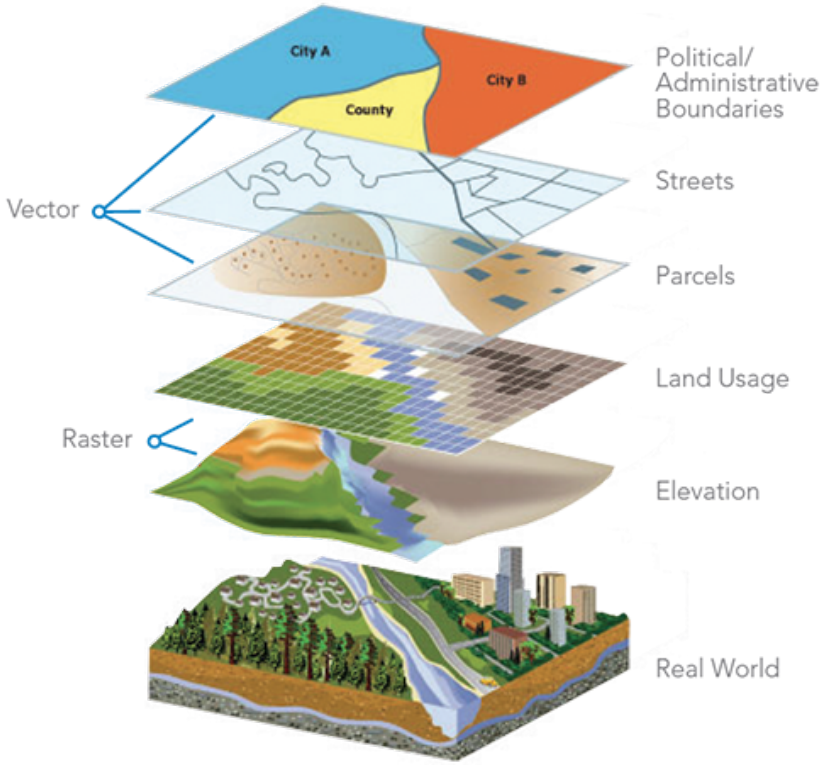
Types of Data Sets: (3) Ordered Data

- Video data: Sequence of images
- Temporal data: Time-series
- Sequential Data: Transaction sequences
- Genetic sequence data



Other Types of Data Sets

- Spatial data
- Image and multimedia data



Basic Statistical Descriptions

To Better Understand the Data

Basic Statistical Descriptions

- Central tendency
  - Mean
  - Median
  - Percentiles
  - Mode
  - Frequency
  - Max
  - Min
  - …
- Outlier-ness
  - Variance
  - Standard deviation
  - Z-score
  - …

Population vs Sample

- Mean: What is the average #faculty of CS colleges in the US?
- However, the US has *N* CS colleges, where *N* >> 126.
- N is the population size.

- *n* = 126 is the sample size.
- Biased sampling!

	A	B	C	D
1	http://csrankings.org/			
2	Rank	Institution	Count	Faculty
3	1	Carnegie Mellon University	18.5	150
4	2	Massachusetts Institute of Technology	12.2	82
5	3	Stanford University	10.9	54
6	3	University of California - Berkeley	10.9	81
7	5	Univ. of Illinois at Urbana-Champaign	9.9	84
8	6	Cornell University	8.7	68
9	7	University of Michigan	8.6	63
10	8	University of Washington	8.3	56
11	9	University of California - San Diego	6.9	54
12	10	Georgia Institute of Technology	6.8	75
13	11	University of Wisconsin - Madison	5.9	47
14	12	Columbia University	5.8	47
15	13	University of Pennsylvania	5.6	46
16	14	University of Southern California	5.5	49
17	15	Princeton University	5.3	51
18	16	University of Texas at Austin	5.2	42
19	16	University of Maryland - College Park	5.2	44
20	18	University of California - Los Angeles	5	37
21	19	Northeastern University	4.8	54
22	19	Purdue University	4.8	51
23	21	University of Massachusetts Amherst	4.7	50
24	22	New York University	4.5	47
25	23	Harvard University	4.2	29
26	23	University of		54
27	25	Rutgers Univ		43
28	26	University of		25
29	27	University of Utah	3.4	39
30	27	Pennsylvania State University	3.4	31
31	29	Stony Brook University	3.3	41
32	30	University of California - Davis	3.2	29

126 institutions

Mean and Median

- Mean:
  - Note: n is sample size and N is population size.

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

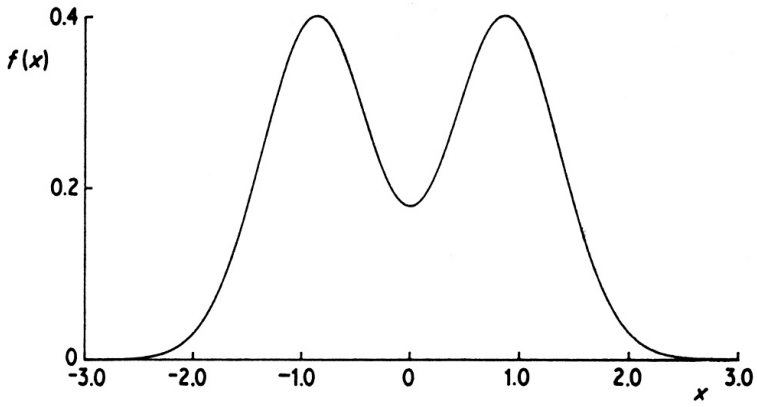
  - Trimmed mean: Chopping extreme values
- Median:
  - Middle value if odd number of values, or average of the middle two values otherwise

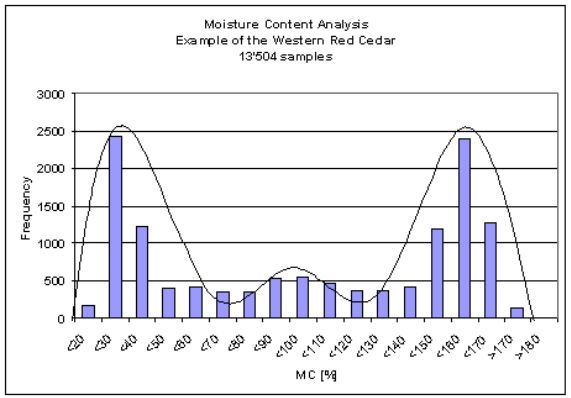
Percentiles

- Given an ordinal or continuous feature x and a number p between 0 and 100, the p-th percentile is a value x_p of such that p% of the observed value of x are less than x_p.
  - For example, the 50th percentile is the value $x_{50\%}$ such that 50% of all values of x are less than $x_{50\%}$.
- Usually,
  - Mean: The 50th percentile
  - Quantiles:
    - Q1: The 25th percentile
    - Q3: The 75th percentile

Frequency and Mode

- Frequency:
  - The percentage of time the value occurs in the dataset
- Mode:
  - Value that occurs most frequently in the data
  - Multi-modal: Bi-modal, tri-modal…
- Typically used with categorical data



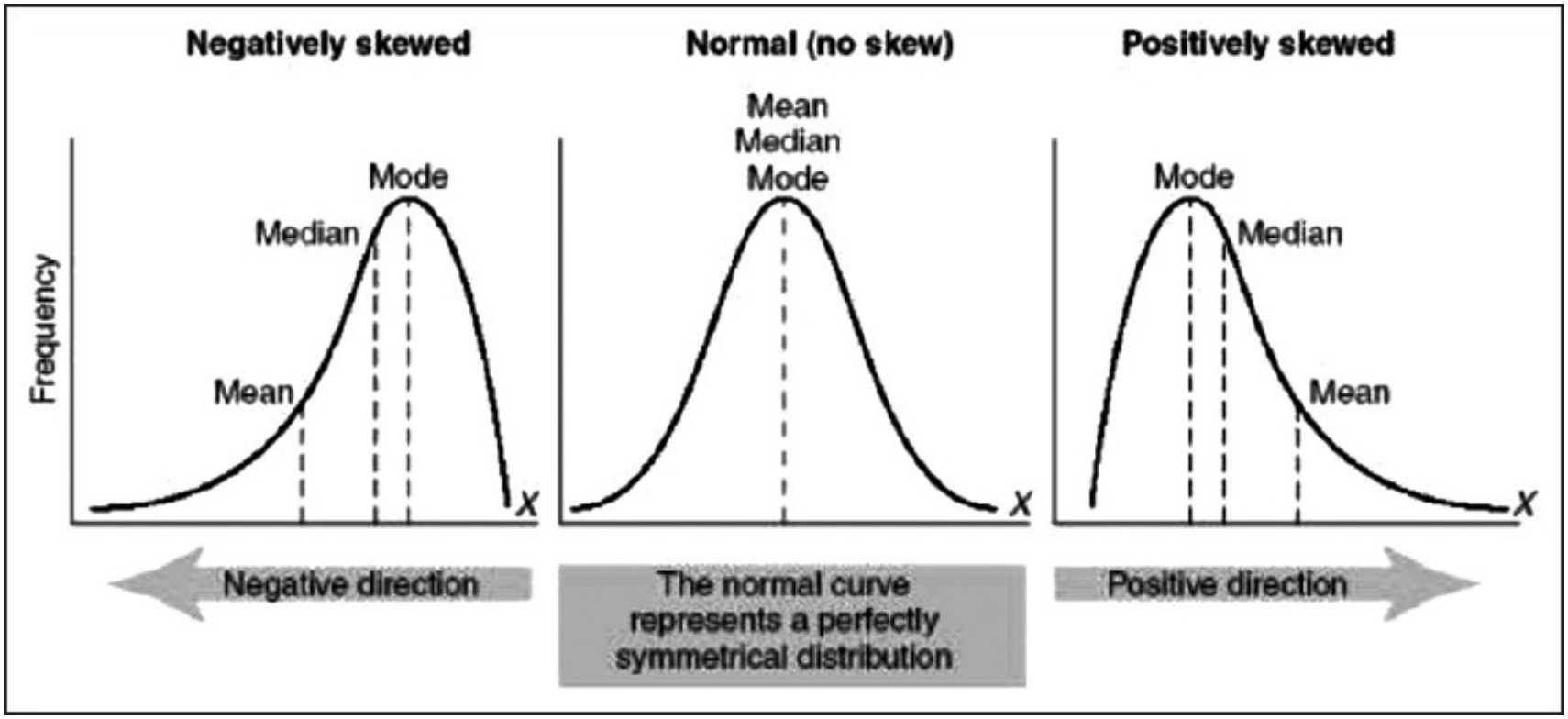


iPython (I)

	A	B	C	D
1	http://csrankings.org/			
2	Rank	Institution	Count	Faculty
3	1	Carnegie Mellon University	18.5	150
4	2	Massachusetts Institute of Technology	12.2	82
5	3	Stanford University	10.9	54
6	3	University of California - Berkeley	10.9	81
7	5	Univ. of Illinois at Urbana-Champaign	9.9	84
8	6	Cornell University	8.7	68
9	7	University of Michigan	8.6	63
10	8	University of Washington	8.3	56
11	9	University of California - San Diego	6.9	54
12	10	Georgia Institute of Technology	6.8	75
13	11	University of Wisconsin - Madison	5.9	47
14	12	Columbia University	5.8	47
15	13	University of Pennsylvania	5.6	46
16	14	University of Southern California	5.5	49
17	15	Princeton University	5.3	51
18	16	University of Texas at Austin	5.2	42
19	16	University of Maryland - College Park	5.2	44
20	18	University of California - Los Angeles	5	37
21	19	Northeastern University	4.8	54
22	19	Purdue University	4.8	51
23	21	University of Massachusetts Amherst	4.7	50
24	22	New York University	4.5	47
25	23	Harvard University	4.2	29
26	23	University of California - Irvine	4.2	54
27	25	Rutgers University	3.9	43
28	26	University of California - Santa Barbara	3.5	25
29	27	University of Utah	3.4	39
30	27	Pennsylvania State University	3.4	31
31	29	Stony Brook University	3.3	41
32	30	University of California - Davis	3.2	29

126 institutions

#Faculty: Positively Skewed



Variance and Standard Deviation

- Variance and standard deviation (sample: s, population: σ):
  - Variance is the expectation of the squared deviation of a random variable from its mean.

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 \qquad \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Why?

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

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

Variance: Difference between Biased and Unbiased

$$\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2 = \frac{1}{n}\sum_{i=1}^{n}\left[(X_i - \mu) + (\mu - \bar{X})\right]^2$$

Biased

$$= \frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2 + \frac{2}{n}\sum_{i=1}^{n}(X_i - \mu)(\mu - \bar{X}) + \frac{1}{n}\sum_{i=1}^{n}(\mu - \bar{X})^2$$

$$= \frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2 + 2(\bar{X} - \mu)(\mu - \bar{X}) + (\mu - \bar{X})^2$$

$$= \frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2 - (\mu - \bar{X})^2$$

Unbiased

$$\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

Bessel's Correction (Wikipedia):

3 alternative proofs of correctness

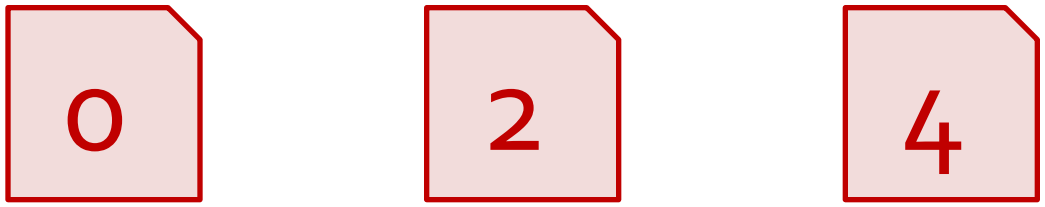
Bias in Population Estimates

Population Mean = 4

- Consider we have samples {2, 5, 11}
  - Mean = 6
  - Median = 5
- Consider we have samples {2, 6, 7}
  - Mean = 5
  - Median = 6

How About Variance?

- Suppose we have 3 cards in a bag: (population)



- Population mean and population variance

$$\mu = \frac{0+2+4}{3} = 2$$

$$\sigma^2 = \frac{(0-2)^2 + (2-2)^2 + (4-2)^2}{3} = \frac{8}{3}$$

Biased Sample Variance

n = 2

$$\bar{x} = \frac{\sum x}{n} \qquad S^2 = \frac{\sum (x - \bar{x})^2}{\cancel{n-1}\ n}$$

(0,0)	$\frac{0+0}{2} = 0$	$\frac{(0-0)^2 + (0-0)^2}{2} = 0$
(0,2)	$\frac{0+2}{2} = 1$	$\frac{(0-1)^2 + (2-1)^2}{2} = 1$
(0,4)	$\frac{0+4}{2} = 2$	$\frac{(0-2)^2 + (4-2)^2}{2} = 4$
(2,0)	$\frac{2+0}{2} = 1$	$\frac{(2-1)^2 + (0-1)^2}{2} = 1$
(2,2)	$\frac{2+2}{2} = 2$	$\frac{(2-2)^2 + (2-2)^2}{2} = 0$
(2,4)	$\frac{2+4}{2} = 3$	$\frac{(2-3)^2 + (4-3)^2}{2} = 1$
(4,0)	$\frac{4+0}{2} = 2$	$\frac{(4-2)^2 + (0-2)^2}{2} = 4$
(4,2)	$\frac{4+2}{2} = 3$	$\frac{(4-3)^2 + (2-3)^2}{2} = 1$
(4,4)	$\frac{4+4}{2} = 4$	$\frac{(4-4)^2 + (4-4)^2}{2} = 0$

Sample mean:

$$\frac{0+1+2+1+2+3+2+3+4}{9} = 2$$

Sample variance:

$$\frac{0+1+4+1+0+1+4+1+0}{9} = \boxed{\frac{4}{3}}$$

Unbiased Sample Variance n = 2

Sample mean:

$$\frac{0+1+2+1+2+3+2+3+4}{9}=2$$

Sample variance:

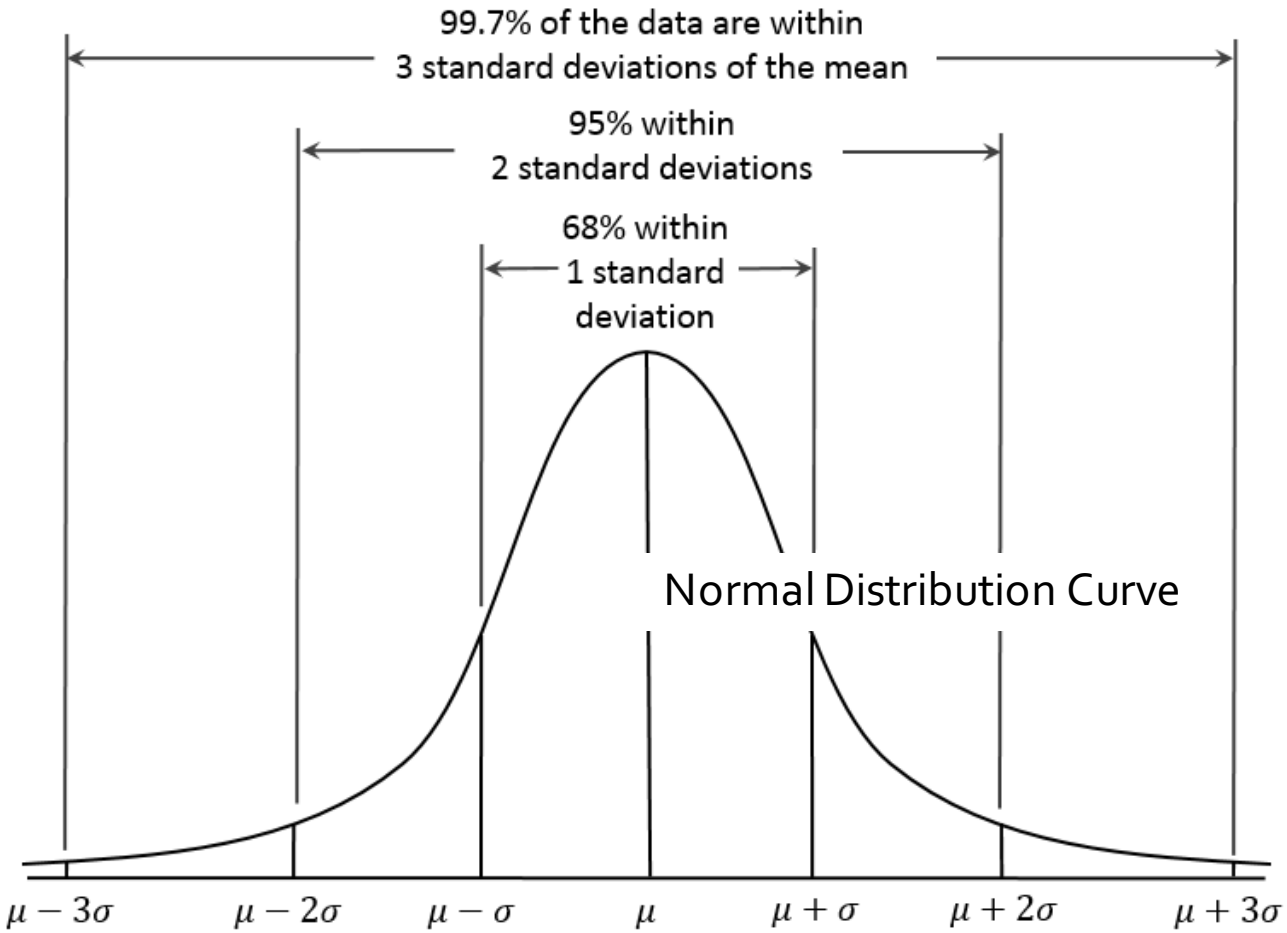
$$\frac{0+2+8+2+0+2+8+2+0}{9}=\boxed{\frac{8}{3}}$$

$$\bar{x}=\frac{\sum x}{n} \qquad S^2=\frac{\sum(x-\bar{x})^2}{\boxed{n-1}}$$

(0,0)	$\frac{0+0}{2}=0$	$\frac{(0-0)^2+(0-0)^2}{2}=0$
(0,2)	$\frac{0+2}{2}=1$	$\frac{(0-1)^2+(2-1)^2}{2}=1$
(0,4)	$\frac{0+4}{2}=2$	$\frac{(0-2)^2+(4-2)^2}{2}=4$
(2,0)	$\frac{2+0}{2}=1$	$\frac{(2-1)^2+(0-1)^2}{2}=1$
(2,2)	$\frac{2+2}{2}=2$	$\frac{(2-2)^2+(2-2)^2}{2}=0$
(2,4)	$\frac{2+4}{2}=3$	$\frac{(2-3)^2+(4-3)^2}{2}=1$
(4,0)	$\frac{4+0}{2}=2$	$\frac{(4-2)^2+(0-2)^2}{2}=4$
(4,2)	$\frac{4+2}{2}=3$	$\frac{(4-3)^2+(2-3)^2}{2}=1$
(4,4)	$\frac{4+4}{2}=4$	$\frac{(4-4)^2+(4-4)^2}{2}=0$

Z Score

Z-score: The distance between the raw score and the population mean in the unit of the standard deviation

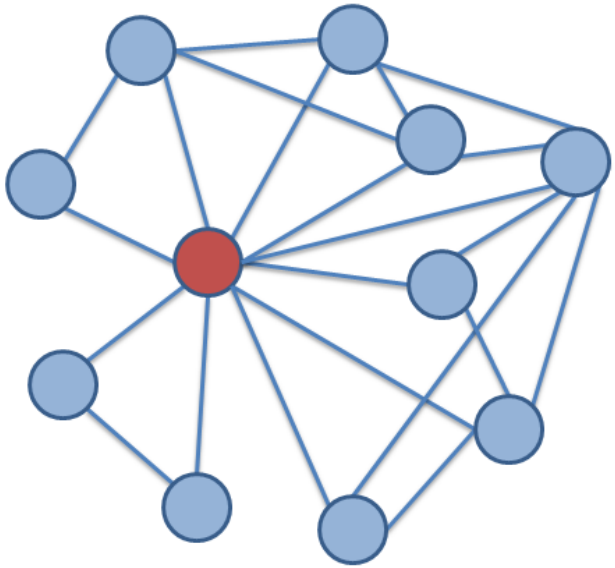


iPython (II)

	A	B	C	D
1	http://csrankings.org/			
2	Rank	Institution	Count	Faculty
3	1	Carnegie Mellon University	18.5	150
4	2	Massachusetts Institute of Technology	12.2	82
5	3	Stanford University	10.9	54
6	3	University of California - Berkeley	10.9	81
7	5	Univ. of Illinois at Urbana-Champaign	9.9	84
8	6	Cornell University	8.7	68
9	7	University of Michigan	8.6	63
10	8	University of Washington	8.3	56
11	9	University of California - San Diego	6.9	54
12	10	Georgia Institute of Technology	6.8	75
13	11	University of Wisconsin - Madison	5.9	47
14	12	Columbia University	5.8	47
15	13	University of Pennsylvania	5.6	46
16	14	University of Southern California	5.5	49
17	15	Princeton University	5.3	51
18	16	University of Texas at Austin	5.2	42
19	16	University of Maryland - College Park	5.2	44
20	18	University of California - Los Angeles	5	37
21	19	Northeastern University	4.8	54
22	19	Purdue University	4.8	51
23	21	University of Massachusetts Amherst	4.7	50
24	22	New York University	4.5	47
25	23	Harvard University	4.2	29
26	23	University of California - Irvine	4.2	54
27	25	Rutgers University	3.9	43
28	26	University of California - Santa Barbara	3.5	25
29	27	University of Utah	3.4	39
30	27	Pennsylvania State University	3.4	31
31	29	Stony Brook University	3.3	41
32	30	University of California - Davis	3.2	29

126 institutions

Descriptions for Network Data

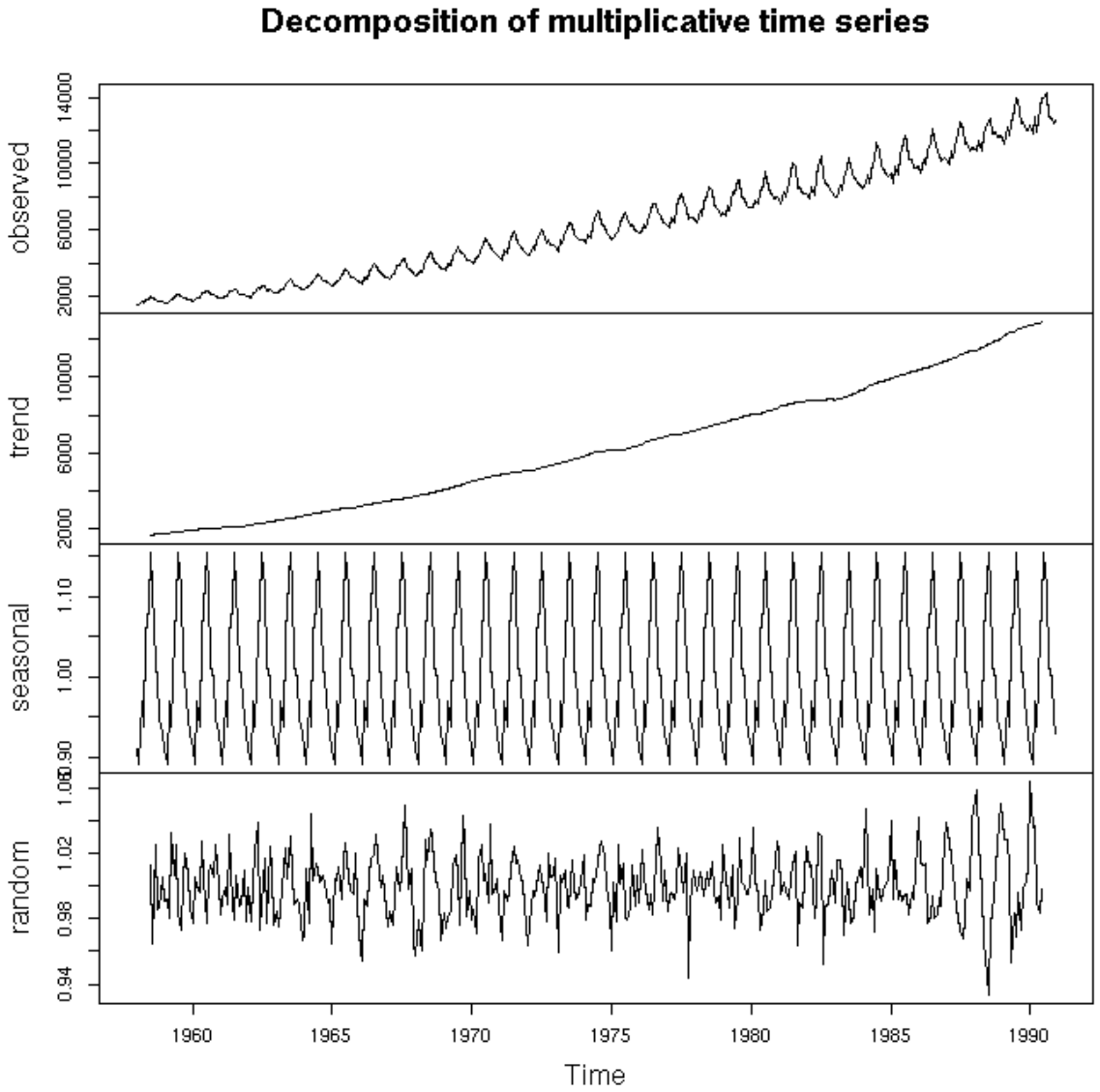


Degree: How many people can this paper reach directly?

Betweenness: How likely is this person to be the most direct route between two people in the network?

Closeness: How fast can this person reach everyone in the network?

Descriptions for Time Series Data



Data Transformation: Normalization

- Min-max normalization
- Z-score normalization
- Decimal scaling normalization

Min-Max Normalization

Transform the data from measured units to a new interval from new_min_F to new_max_F for feature F:

$$v' = \frac{v - min_F}{max_F - min_F}(new\_max_F - new\_min_F) + new\_min_F$$

where v is the current value of feature F.

Suppose that the minimum and maximum values for the feature income are \$120,000 and \$98,000, respectively. We would like to map income to the range [0.0,1.0]. By min-max normalization, a value of \$73,600 for income is transformed to:

$$\frac{73{,}600 - 12{,}000}{98{,}000 - 12{,}000}(1.0 - 0.0) + 0 = 0.716$$

Z-Score Normalization

Transform the data by converting the values to a common scale with an average of zero and a standard deviation of one. A value, v, of A is normalized to v' by computing:

$$v' = \frac{v - F}{\sigma_F}$$

where F and σ_F are the mean and standard deviation of feature F, respectively.

Z-Score Normalization

- The normalized value of X_i is calculated as:

$$Z_i = \frac{X_i - \bar{X}}{s}$$

$$\mathbf{y} = \begin{bmatrix} 35 \\ 36 \\ 46 \\ 68 \\ 70 \end{bmatrix}$$

$$\begin{aligned} s &= \sqrt{\frac{(35-51)^2 + (36-51)^2 + (46-51)^2 + (68-51)^2 + (70-51)^2}{5-1}} \\ &= \frac{1}{2}\sqrt{(-16)^2 + -15^2 + (-5)^2 + 17^2 + 19^2} \\ &= 17\,. \end{aligned}$$

$$\mathbf{z} = \begin{bmatrix} \frac{35-51}{17} \\ \frac{36-51}{17} \\ \frac{46-51}{17} \\ \frac{68-51}{17} \\ \frac{70-51}{17} \end{bmatrix} = \begin{bmatrix} -\frac{16}{17} \\ -\frac{15}{17} \\ -\frac{5}{17} \\ \frac{17}{17} \\ \frac{19}{17} \end{bmatrix} = \begin{bmatrix} -0.9412 \\ -0.8824 \\ -0.2941 \\ 1.0000 \\ 1.1176 \end{bmatrix}$$

vs. Min-Max Normalization:

[0, 1/35, 11/35, 33/35, 1] = [0, 0.0286, 0.3143, 0.9429, 1.0]

Decimal Scaling Normalization

Transform the data by moving the decimal points of values of feature F. The number of decimal points moved depends on the maximum absolute value of F. A value v of F is normalized to v' by computing :

$$v' = \frac{v}{10^j},$$

where j is the smallest integer such that $Max(|v'|) < 1$.

iPython (III)

	A	B	C	D
1	http://csrankings.org/			
2	Rank	Institution	Count	Faculty
3	1	Carnegie Mellon University	18.5	150
4	2	Massachusetts Institute of Technology	12.2	82
5	3	Stanford University	10.9	54
6	3	University of California - Berkeley	10.9	81
7	5	Univ. of Illinois at Urbana-Champaign	9.9	84
8	6	Cornell University	8.7	68
9	7	University of Michigan	8.6	63
10	8	University of Washington	8.3	56
11	9	University of California - San Diego	6.9	54
12	10	Georgia Institute of Technology	6.8	75
13	11	University of Wisconsin - Madison	5.9	47
14	12	Columbia University	5.8	47
15	13	University of Pennsylvania	5.6	46
16	14	University of Southern California	5.5	49
17	15	Princeton University	5.3	51
18	16	University of Texas at Austin	5.2	42
19	16	University of Maryland - College Park	5.2	44
20	18	University of California - Los Angeles	5	37
21	19	Northeastern University	4.8	54
22	19	Purdue University	4.8	51
23	21	University of Massachusetts Amherst	4.7	50
24	22	New York University	4.5	47
25	23	Harvard University	4.2	29
26	23	University of California - Irvine	4.2	54
27	25	Rutgers University	3.9	43
28	26	University of California - Santa Barbara	3.5	25
29	27	University of Utah	3.4	39
30	27	Pennsylvania State University	3.4	31
31	29	Stony Brook University	3.3	41
32	30	University of California - Davis	3.2	29

126 institutions

Summary: Data Description

- Understand what is data object and attribute/feature;
- **Understand different attribute types;**
- **Understand different data set types;**
- Describe basic statistical descriptions
  - Describe and calculate central tendency
    - **Mean, Median, Mode, Frequency, Percentiles**
    - Population and sample
  - Describe and calculate outlier-ness
    - **Variance, Standard Deviation, Z-score**
    - Biased/Unbiased sample variance
    - Z-score normalization vs min-max normalization

References

- W. Cleveland, Visualizing Data, Hobart Press, 1993
- T. Dasu and T. Johnson. Exploratory Data Mining and Data Cleaning. John Wiley, 2003
- U. Fayyad, G. Grinstein, and A. Wierse. Information Visualization in Data Mining and Knowledge Discovery, Morgan Kaufmann, 2001
- L. Kaufman and P. J. Rousseeuw. Finding Groups in Data: an Introduction to Cluster Analysis. John Wiley & Sons, 1990.
- H. V. Jagadish et al., Special Issue on Data Reduction Techniques. Bulletin of the Tech. Committee on Data Eng., 20(4), Dec. 1997
- D. A. Keim. Information visualization and visual data mining, IEEE trans. on Visualization and Computer Graphics, 8(1), 2002
- D. Pyle. Data Preparation for Data Mining. Morgan Kaufmann, 1999
- S. Santini and R. Jain," Similarity measures", IEEE Trans. on Pattern Analysis and Machine Intelligence, 21(9), 1999
- E. R. Tufte. The Visual Display of Quantitative Information, 2nd ed., Graphics Press, 2001
- C. Yu, et al., Visual data mining of multimedia data for social and behavioral studies, Information Visualization, 8(1), 2009