Chapter 2. Getting to Know Your Data: Data Visualization

Meng Jiang

Data Science

Bonzie Colson

From Data to Knowledge

Thank back to last lecture:
Data Resolution

##	Player	gp-gs	min	avg	fg-fga	fg%	3fg-fga	3fg%	ft-fta	ft%	off	def	tot	avg	pf	dq	a	to	blk	stl	pts	avg
35	Colson,Bonzie	36-36	1156	32.1	236-449	.526	26-60	.433	141-180	.783	104	258	362	10.1	81	0	56	44	50	40	639	17.8
03	Beachem,VJ	36-36	1230	34.2	187-443	.422	87-241	.361	61-73	.836	23	123	146	4.1	46	0	31	40	38	33	522	14.5
05	Farrell,Matt	36-36	1238	34.4	172-384	.448	81-193	.420	81-102	.794	9	63	72	2.0	71	0	196	91	5	51	506	14.1
32	Vasturia,Steve	36-36	1244	34.6	162-374	.433	58-162	.358	91-100	.910	25	116	141	3.9	75	1	119	57	4	42	473	13.1
00	Pflueger,Rex	35-11	750	21.4	59-133	.444	27-68	.397	19-29	.655	18	78	96	2.7	60	0	53	24	12	31	164	4.7
02	Gibbs,TJ	36-1	539	15.0	51-136	.375	17-53	.321	49-59	.831	12	41	53	1.5	49	0	62	28	2	26	168	4.7
04	Ryan,Matt	36-0	286	7.9	43-99	.434	36-83	.434	9-10	.900	6	26	32	0.9	29	0	14	13	1	6	131	3.6
23	Geben,Martinas	34-23	421	12.4	42-65	.646	0-0	.000	23-30	.767	42	73	115	3.4	66	3	25	22	11	13	107	3.1
01	Torres,Austin	36-1	261	7.3	21-38	.553	0-0	.000	6-18	.333	23	31	54	1.5	46	0	7	9	8	9	48	1.3
33	Mooney,John	12-0	46	3.8	5-8	.625	2-4	.500	2-2	1.000	6	13	19	1.6	5	0	2	1	1	1	14	1.2
12	Burns,Elijah	11-0	44	4.0	1-4	.250	0-1	.000	7-8	.875	5	5	10	0.9	5	0	1	2	1	2	9	0.8
34	Mazza,Patrick	4-0	4	1.0	1-2	.500	0-0	.000	0-0	.000	0	1	1	0.3	0	0	0	1	1	0	2	0.5
21	Gregory,Matt	5-0	6	1.2	0-4	.000	0-4	.000	0-0	.000	0	0	0	0.0	0	0	0	0	0	0	0	0.0

From Data to Knowledge

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

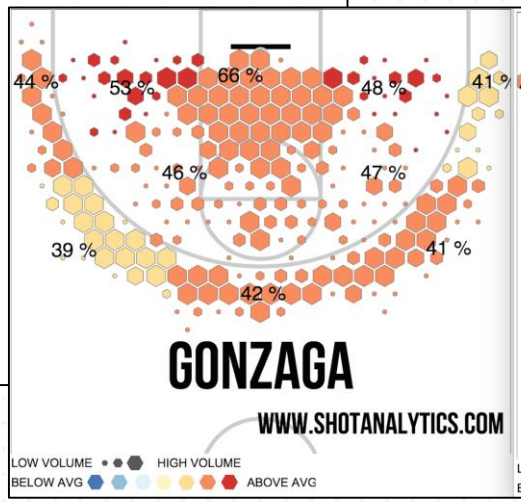
##	Player	gp-gs	min	avg	fg-fga	fg%	3fg-fga	3fg%	ft-fta	ft%	off	def	tot	avg	pf	dq	a	to	blk	stl	pts	avg
35	Colson,Bonzie	36-36	1156	32.1	236-449	.526	26-60	.433	141-180	.783	104	258	362	10.1	81	0	56	44	50	40	639	17.8
03	Beachem,VJ	36-36	1230	34.2	187-443	.422	87-241	.361	61-73	.836	23	123	146	4.1	46	0	31	40	38	33	522	14.5
05	Farrell,Matt	36-36	1238	34.4	172-384	.448	81-193	.420	81-102	.794	9	63	72	2.0	71	0	196	91	5	51	506	14.1
32	Vasturia,Steve	36-36	1244	34.6	162-374	.433	58-162	.358	91-100	.910	25	116	141	3.9	75	1	119	57	4	42	473	13.1
00	Pflueger,Rex	35-11	750	21.4	59-133	.444	27-68	.397	19-29	.655	18	78	96	2.7	60	0	53	24	12	31	164	4.7
02	Gibbs,TJ	36-1	539	15.0	51-136	.375	17-53	.321	49-59	.831	12	41	53	1.5	49	0	62	28	2	26	168	4.7
04	Ryan,Matt	36-0	286	7.9	43-99	.434	36-83	.434	9-10	.900	6	26	32	0.9	29	0	14	13	1	6	131	3.6
23	Geben,Martinas	34-23	421	12.4	42-65	.646	0-0	.000	23-30	.767	42	73	115	3.4	66	3	25	22	11	13	107	3.1
01	Torres,Austin	36-1	261	7.3	21-38	.553	0-0	.000	6-18	.333	23	31	54	1.5	46	0	7	9	8	9	48	1.3
33	Mooney,John	12-0	46	3.8	5-8	.625	2-4	.500	2-2	1.000	6	13	19	1.6	5	0	2	1	1	1	14	1.2
12	Burns,Elijah	11-0	44	4.0	1-4	.250	0-1	.000	7-8	.875	5	5	10	0.9	5	0	1	2	1	2	9	0.8
34	Mazza,Patrick	4-0	4	1.0	1-2	.500	0-0	.000	0-0	.000	0	1	1	0.3	0	0	0	1	1	0	2	0.5
21	Gregory,Matt	5-0	6	1.2	0-4	.000	0-4	.000	0-0	.000	0	0	0	0.0	0	0	0	0	0	0	0	0.0

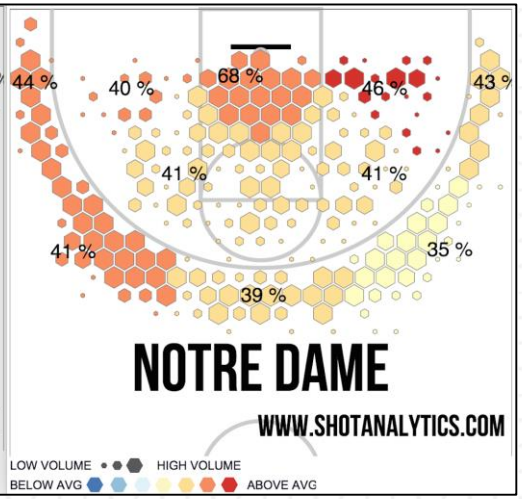
From Data to Knowledge

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							
01	Austin Torres.....		0-1	0-0	0-0	1	0	1							
02	TJ Gibbs.....		0-1	0-0	2-2	0	2	2							
04	Matt Ryan.....		2-3	0-0	2-2	0	2	2							
23	Martinas Geben.....		1-1	0-0	0-0	1	0	1							
	TEAM.....					2	1	3							
	Totals.....		21-52	4-11	14-21	9	27	36							

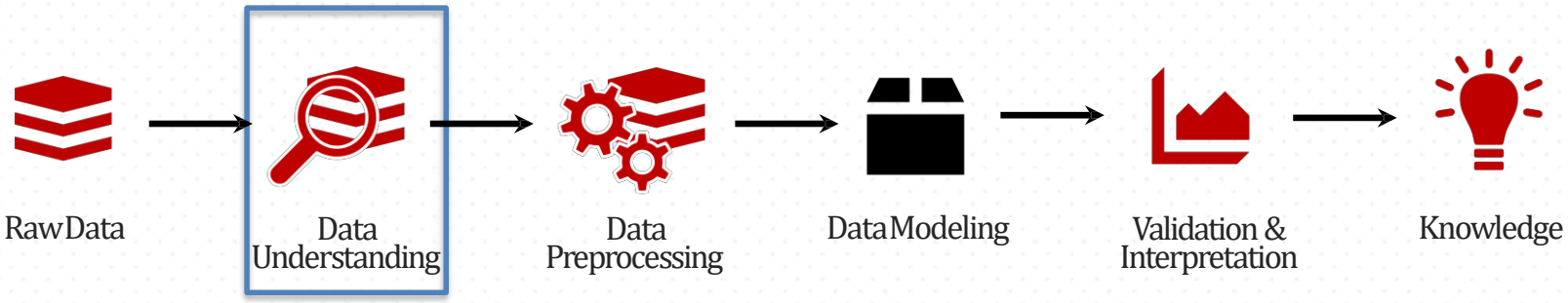
TOTAL FG% 1st Half: 14-30 46.7% 2nd Half: 7-22 31.8%
3-Pt. FG% 1st Half: 2-5 40.0% 2nd Half: 6-6 33.3%
F Throw % 1st Half: 6-8 75.0% 2nd Half: 8-13 61.5%





##	Player	gp-gs	min	avg	fg-fga	fg%	3fg-fga	3fg%	ft-fta	ft%	off	def	tot	avg	pf	dq	a	to	blk	stl	pts	avg
35	Colson,Bonzie	36-36	1156	32.1	236-449	.526	26-60	.433	141-180	.783	104	258	362	10.1	81	0	56	44	50	40	639	17.8
03	Beachem,VJ	36-36	1230	34.2	187-443	.422	87-241	.361	61-73	.836	23	123	146	4.1	46	0	31	40	38	33	522	14.5
05	Farrell,Matt	36-36	1238	34.4	172-384	.448	81-193	.420	81-102	.794	9	63	72	2.0	71	0	196	91	5	51	506	14.1
32	Vasturia,Steve	36-36	1244	34.6	162-374	.433	58-162	.358	91-100	.910	25	116	141	3.9	75	1	119	57	4	42	473	13.1
00	Pflueger,Rex	35-11	750	21.4	59-133	.444	27-68	.397	19-29	.655	18	78	96	2.7	60	0	53	24	12	31	164	4.7
02	Gibbs,TJ	36-1	539	15.0	51-136	.375	17-53	.321	49-59	.831	12	41	53	1.5	49	0	62	28	2	26	168	4.7
04	Ryan,Matt	36-0	286	7.9	43-99	.434	36-83	.434	9-10	.900	6	26	32	0.9	29	0	14	13	1	6	131	3.6
23	Geben,Martinas	34-23	421	12.4	42-65	.646	0-0	.000	23-30	.767	42	73	115	3.4	66	3	25	22	11	13	107	3.1
01	Torres,Austin	36-1	261	7.3	21-38	.553	0-0	.000	6-18	.333	23	31	54	1.5	46	0	7	9	8	9	48	1.3
33	Mooney,John	12-0	46	3.8	5-8	.625	2-4	.500	2-2	1.000	6	13	19	1.6	5	0	2	1	1	1	14	1.2
12	Burns,Elijah	11-0	44	4.0	1-4	.250	0-1	.000	7-8	.875	5	5	10	0.9	5	0	1	2	1	2	9	0.8
34	Mazza,Patrick	4-0	4	1.0	1-2	.500	0-0	.000	0-0	.000	0	1	1	0.3	0	0	0	1	1	0	2	0.5
21	Gregory,Matt	5-0	6	1.2	0-4	.000	0-4	.000	0-0	.000	0	0	0	0.0	0	0	0	0	0	0	0	0.0

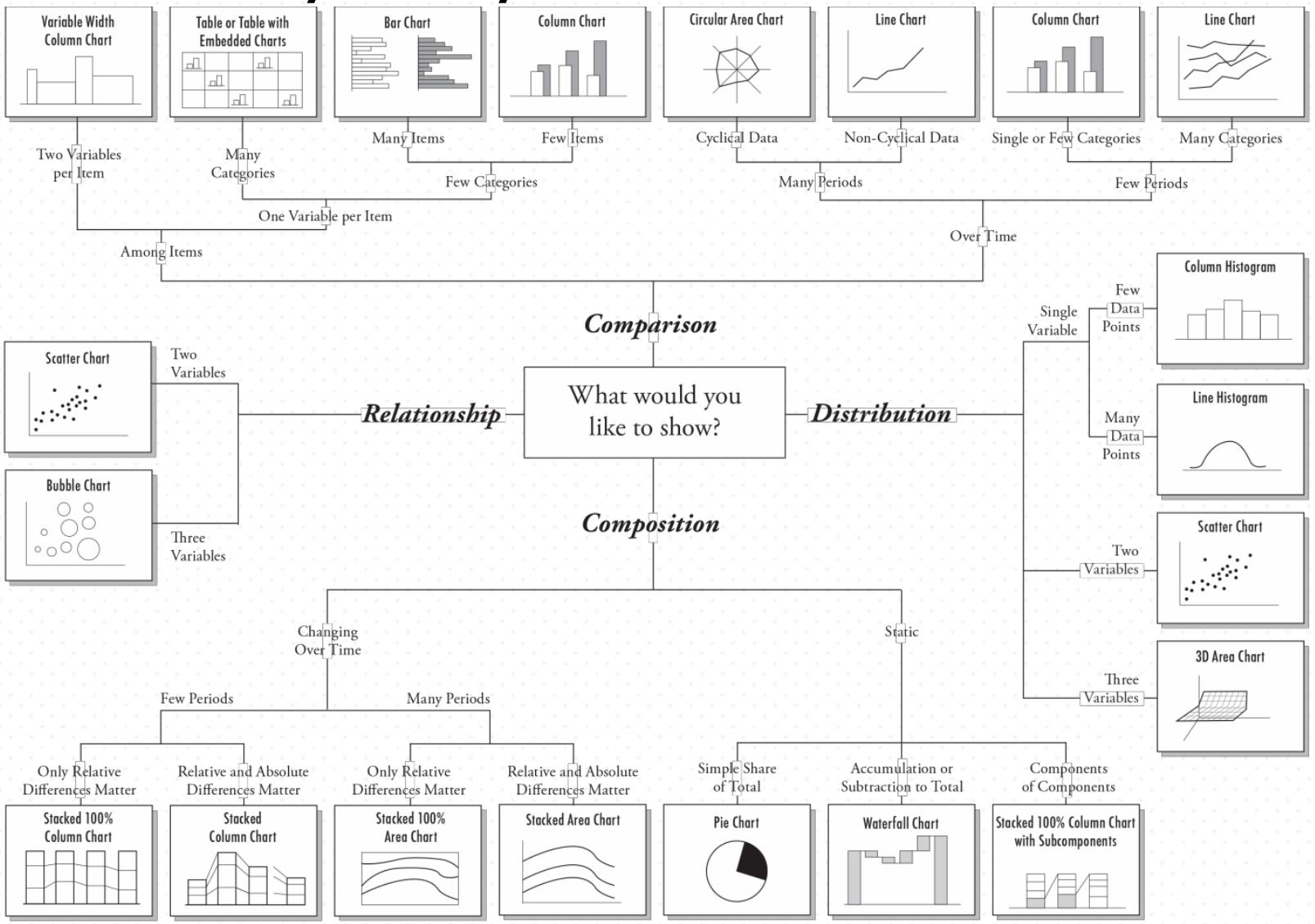
Chapter 2. Getting to Know Your Data



Chapter 2. Getting to Know Your Data

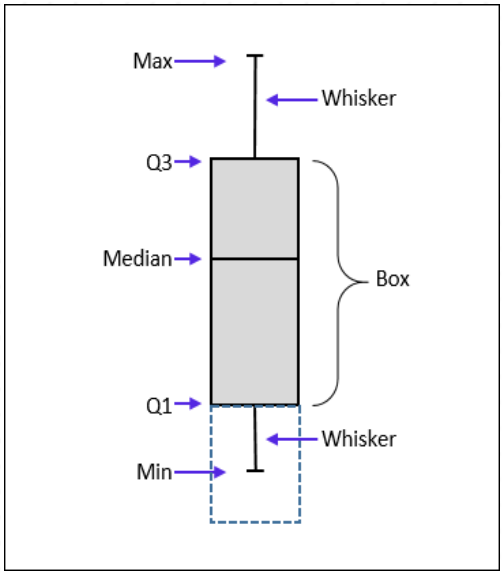
- Data Objects and Attribute Types
- Basic Statistical Descriptions
- **Data Visualization**
- Measuring Data Similarity and Dissimilarity

Many Ways To Visualize Data



Measuring the Dispersion of Data: Quartiles & Boxplots

- **Quartiles**: Q_1 (25th percentile), Q_3 (75th percentile)
- **Inter-quartile range**: $IQR = Q_3 - Q_1$
- **Five number summary**: min, Q_1, median, Q_3, max
- **Boxplot**: Data is represented with a box
  - Q_1, Q_3, IQR: The ends of the box are at the first and third quartiles, i.e., the height of the box is IQR
  - Median (Q_2) is marked by a line within the box
  - Whiskers: Two lines outside the box extended to Minimum and Maximum



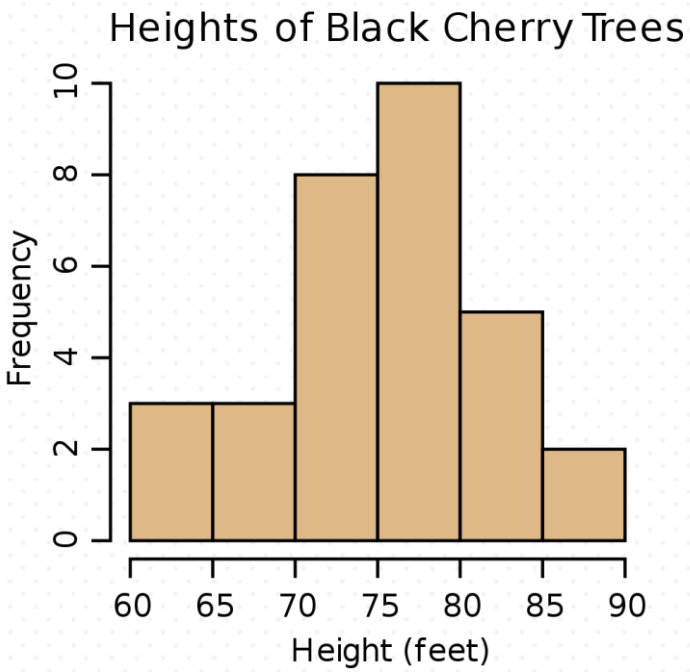
Box Plots Address Things Like

- Is a feature significant?
- Does the location differ between subgroups?
- Does the variation differ between subgroups?
- Are there outliers in the data?

HISTOGRAMS

Graph display of tabulated frequencies, shown as bars

- Usually shows the distribution of values of a single variable of objects in each bin.
- The height of each bar indicates the number of objects.
- The shape depends on the number of bins.

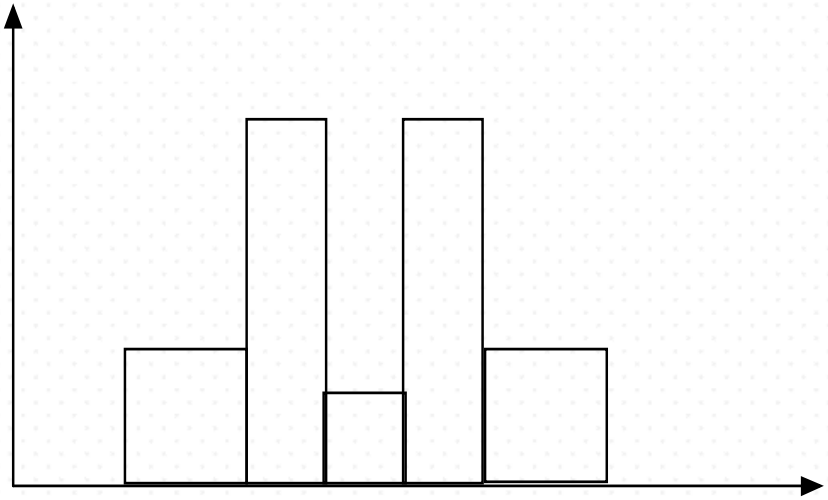


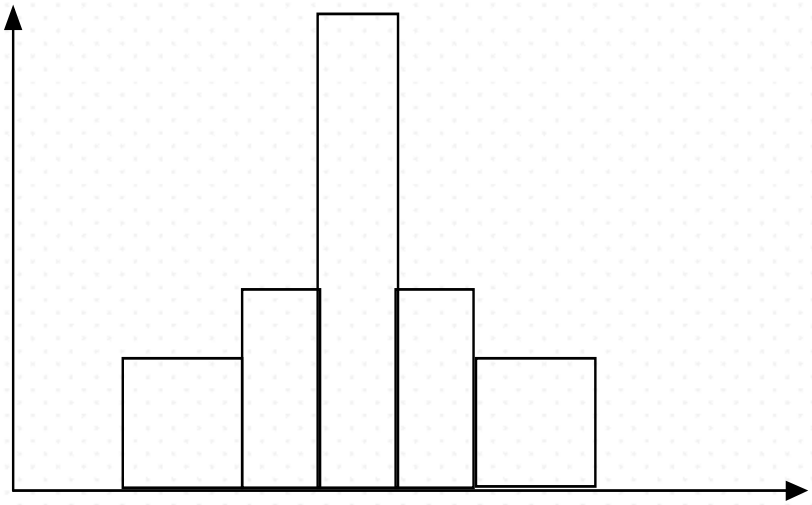
WHAT CAN HISTOGRAMS ADDRESS?

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

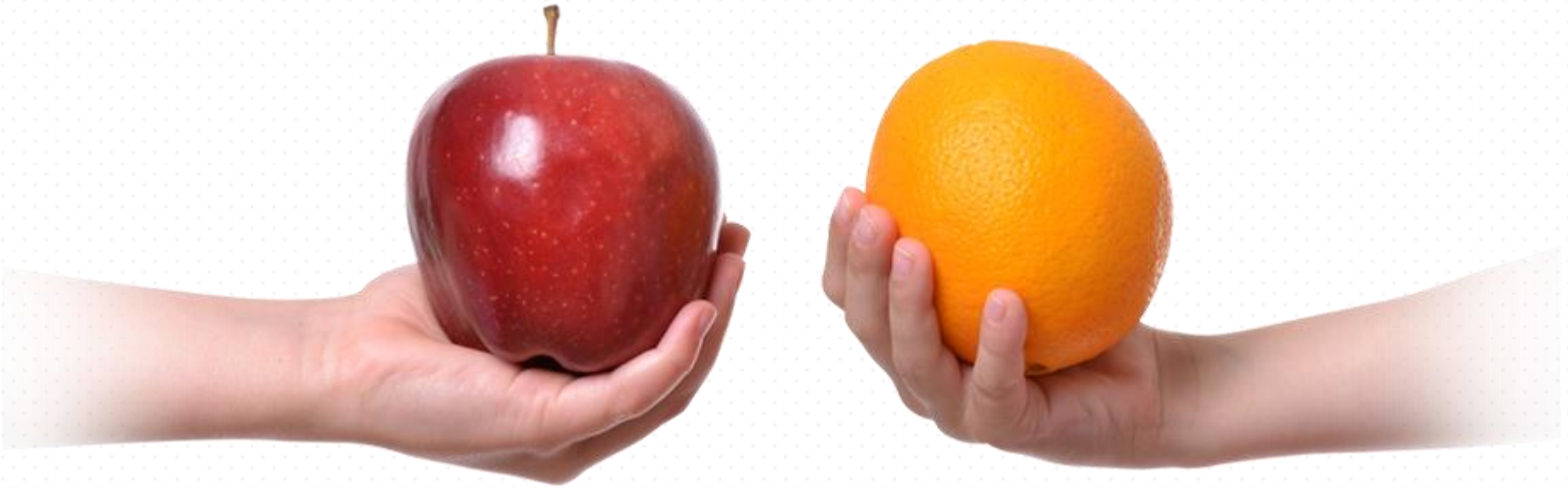
Histograms Often Tell More than Boxplots

- The two histograms may have the same boxplot
  - The same values for: min, Q1, median, Q3, max
- But they have rather different data distributions



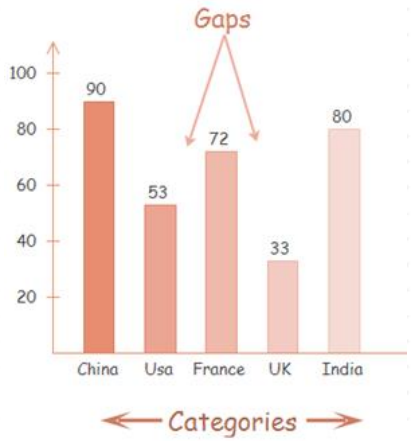


Histogram vs Bar Chart

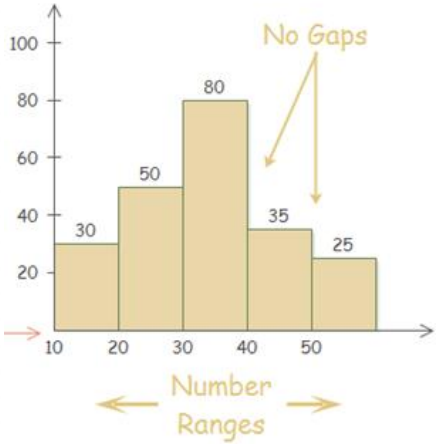


Histogram vs Bar Chart

- Between **histograms** and **bar charts**
  - **Histograms** are used to **show distributions of variables** while **bar charts** are used to **compare variables**
  - **Histograms** plot **binned quantitative data** while **bar charts** plot **categorical data**
  - Bars can be reordered in **bar charts** but not in **histograms**



Bar Chart

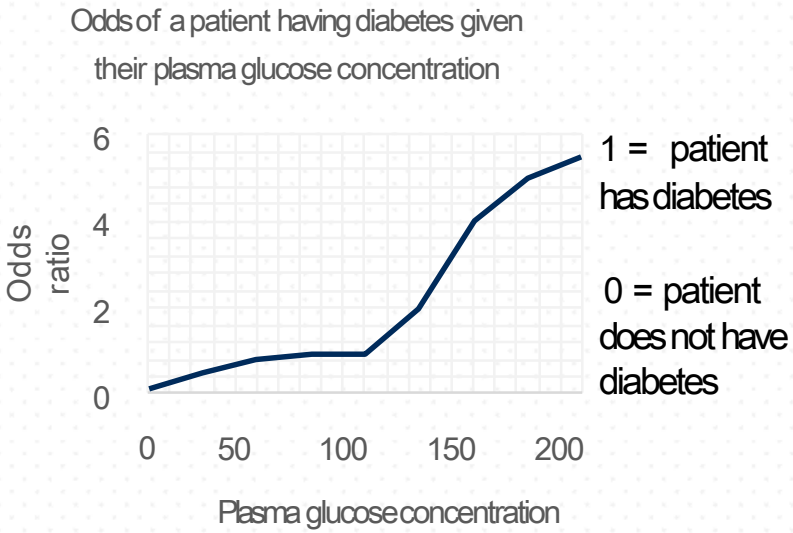


Histogram

ODDS PLOTS

The ratio of one class to another as a function of feature values.

$$OddRatio(x_i, y) = \sum_{j \in k} \frac{p(x_{ij} \mid y = 1)}{p(x_{ij} \mid y = 0)}$$

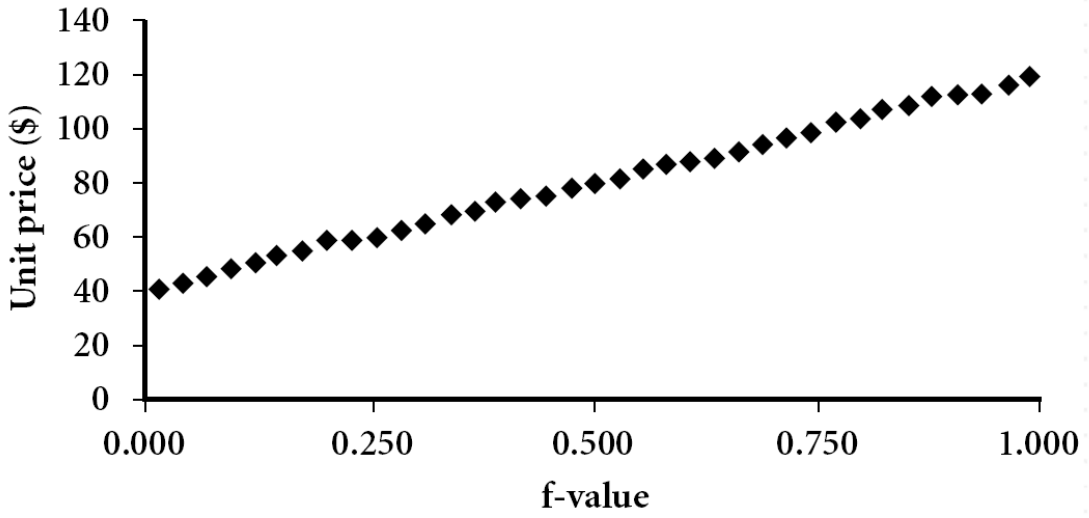


WHAT CAN ODDS PLOTS ADDRESS?

- How do feature values affect the probability of occurrence?
- Is there a threshold for the effect?

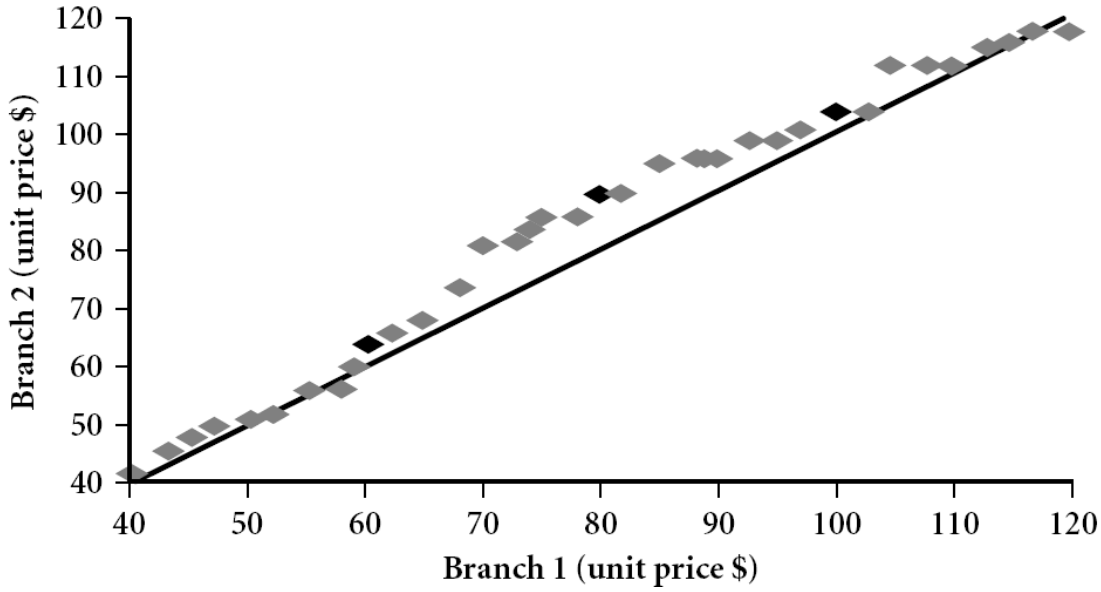
Quantile Plot

- 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



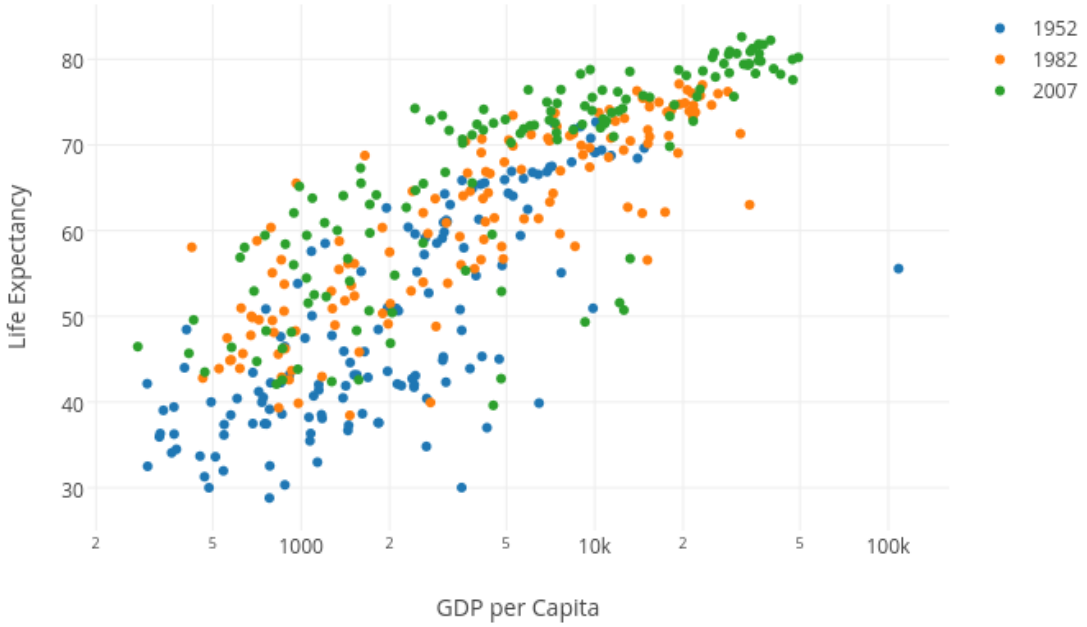
Quantile-Quantile (Q-Q) Plot

- Graphs the quantiles of one univariate distribution against the corresponding quantiles of 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



Scatter plot

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



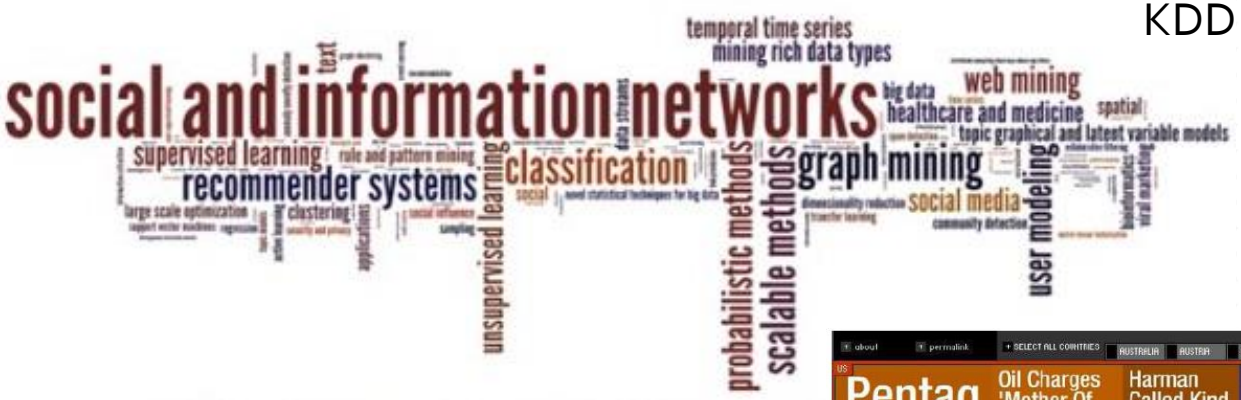
Summary: Graphic Displays of Basic Statistical Descriptions

- **Boxplot**: graphic display of five-number summary
- **Histogram**: x-axis are values, y-axis representative 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

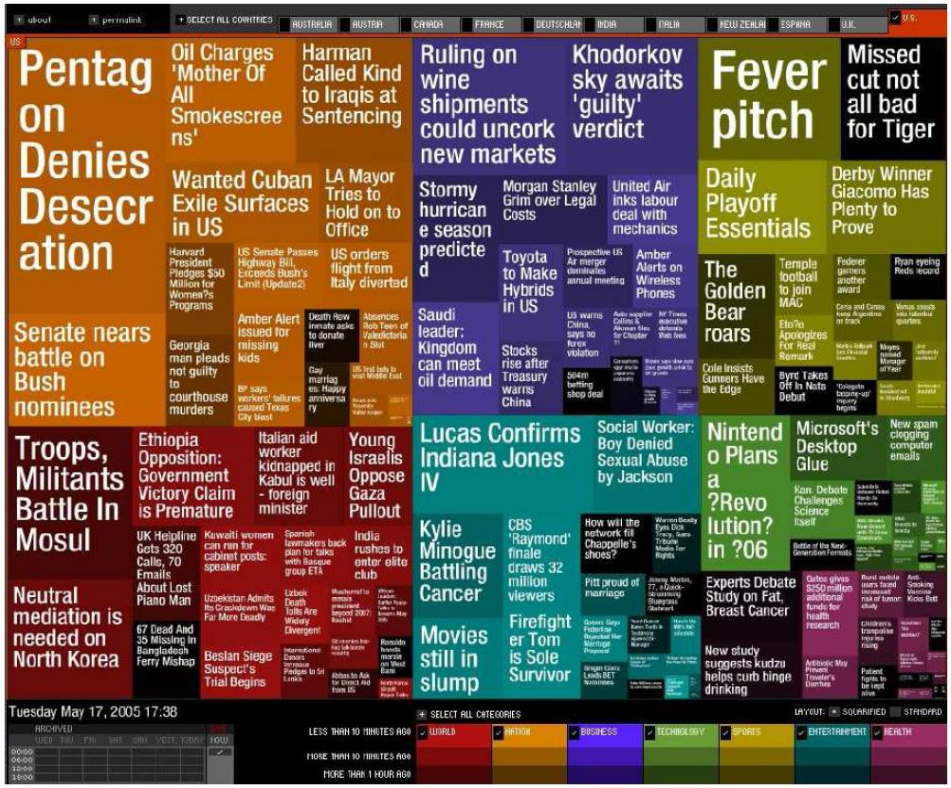
iPython Examples

Other Visualization: Tag Cloud

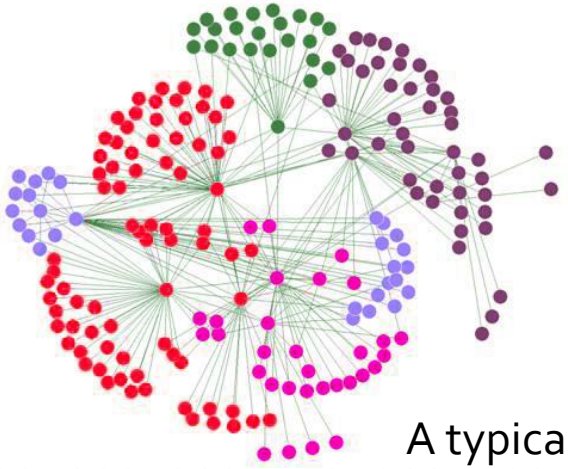
KDD 2013 Research Paper Title



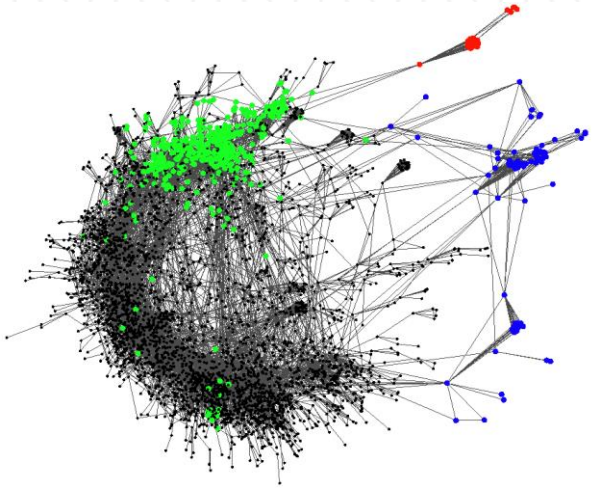
Newsmap: Google News Stories in 2005



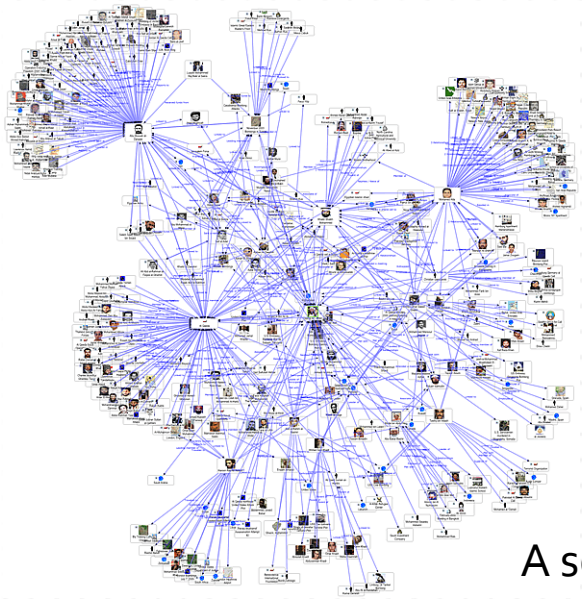
Other Visualization: Networks



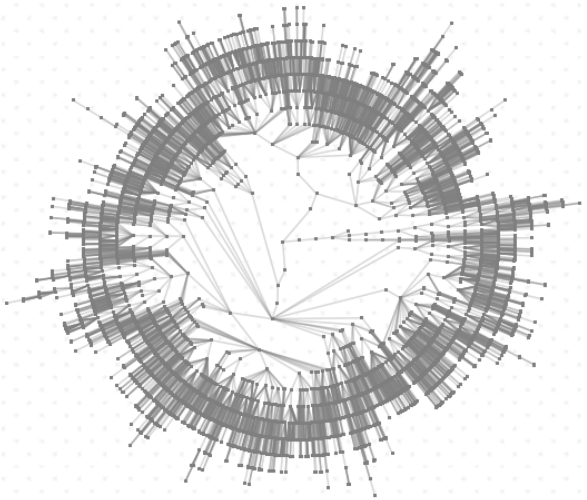
A typical network structure



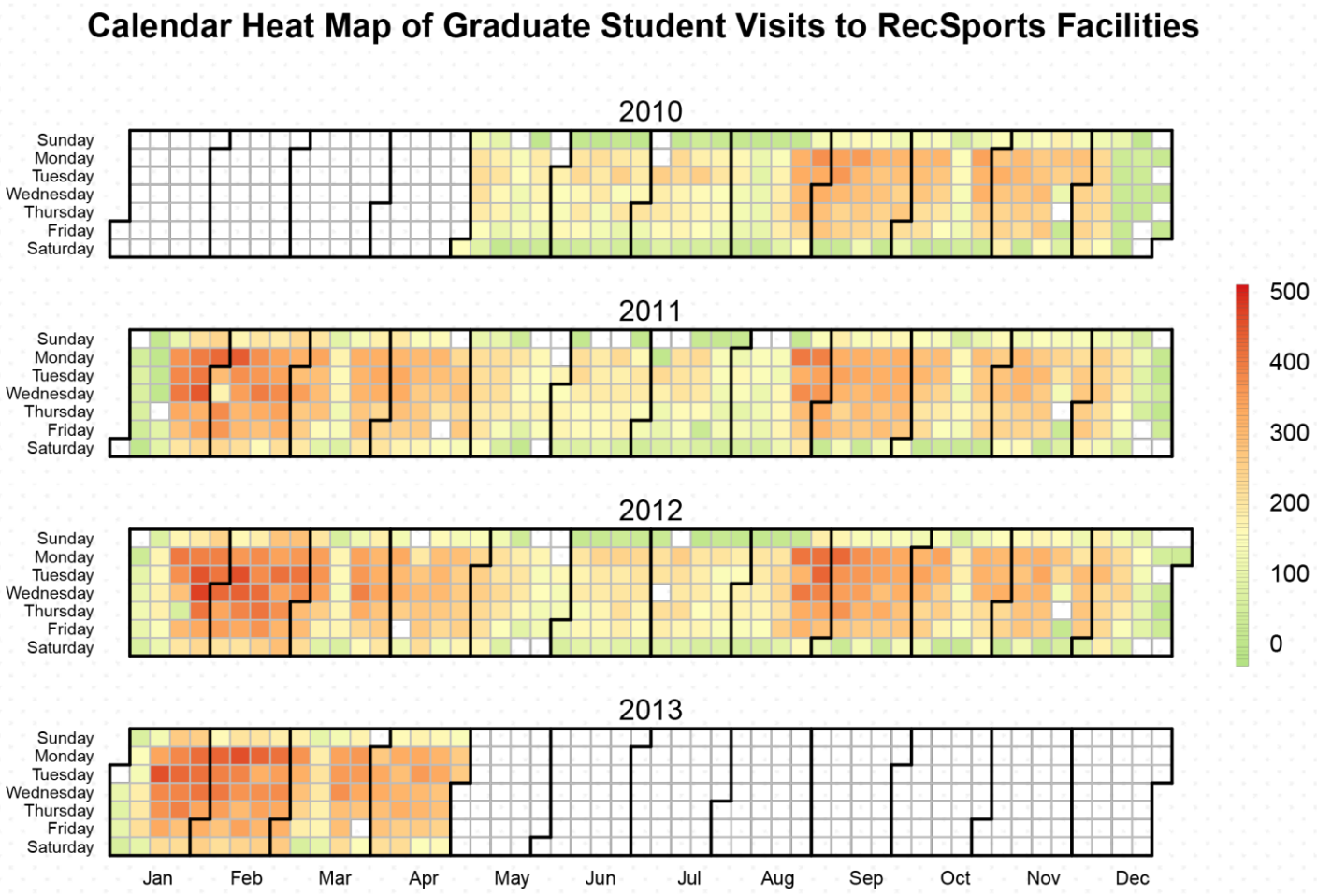
organizing information networks



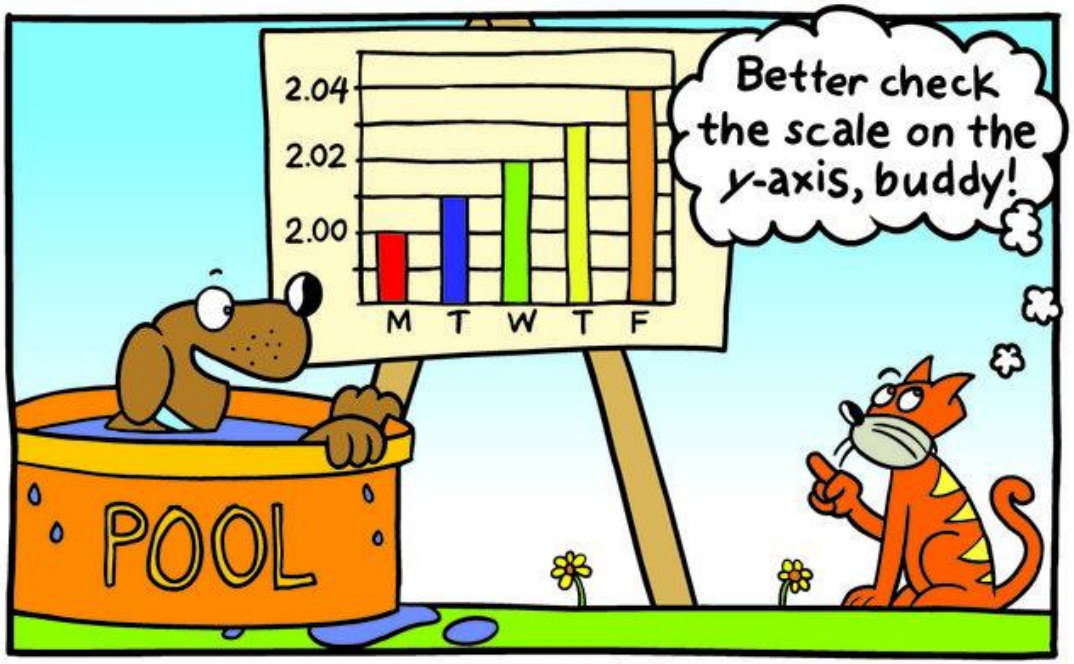
A social network



Other Visualization: Temporal Data

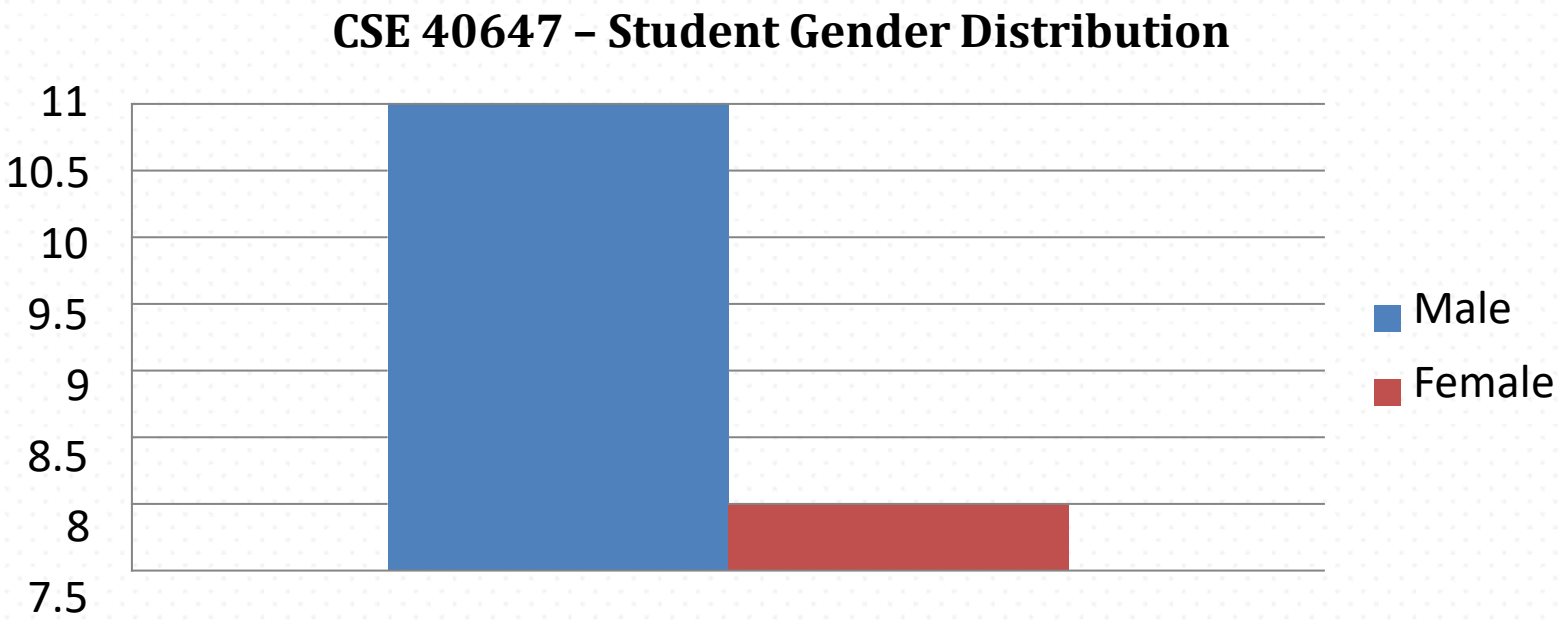


Extra: Graphical Integrity

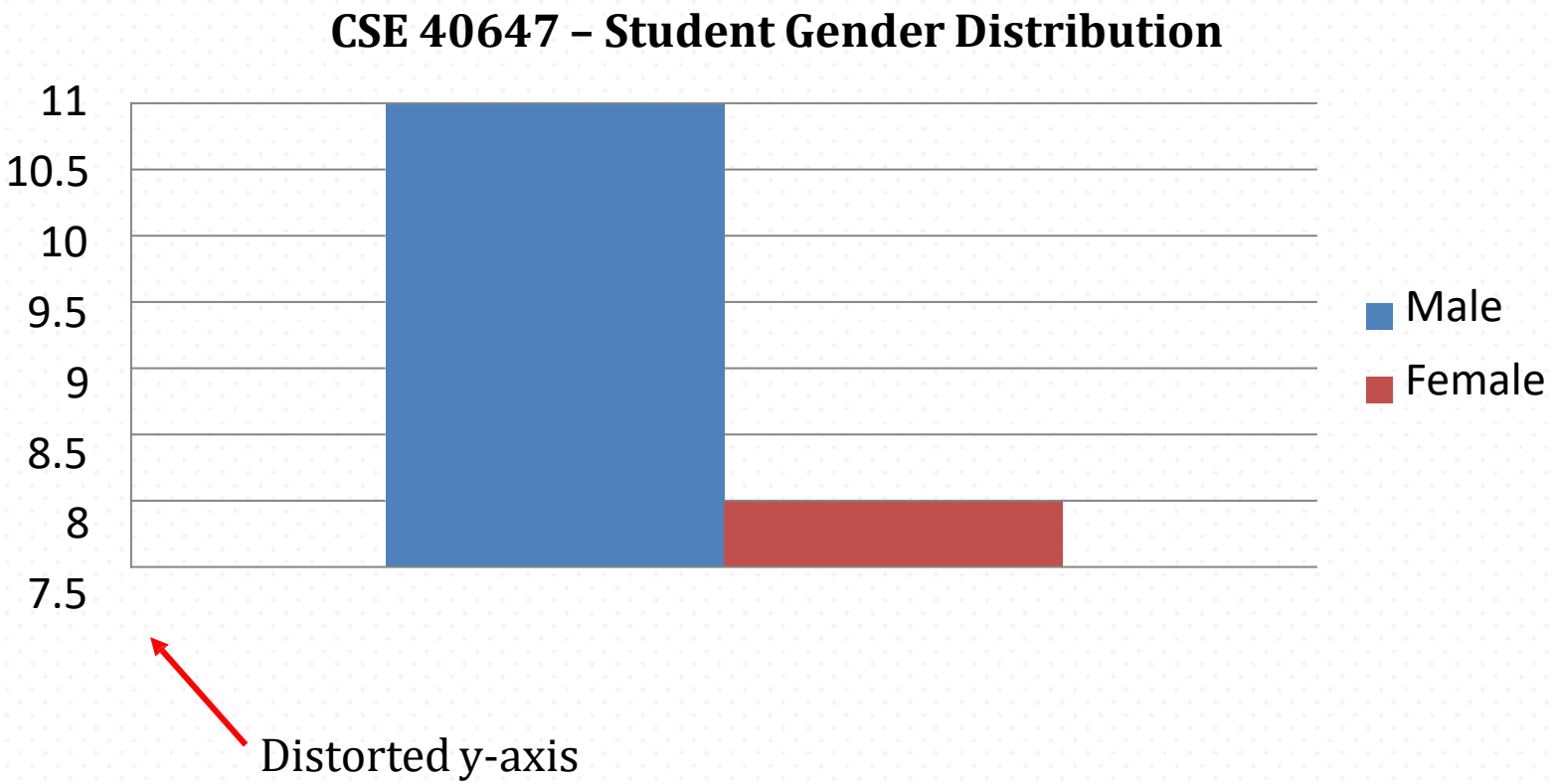


"Wow. The number of minutes I can dog paddle is growing like crazy!"

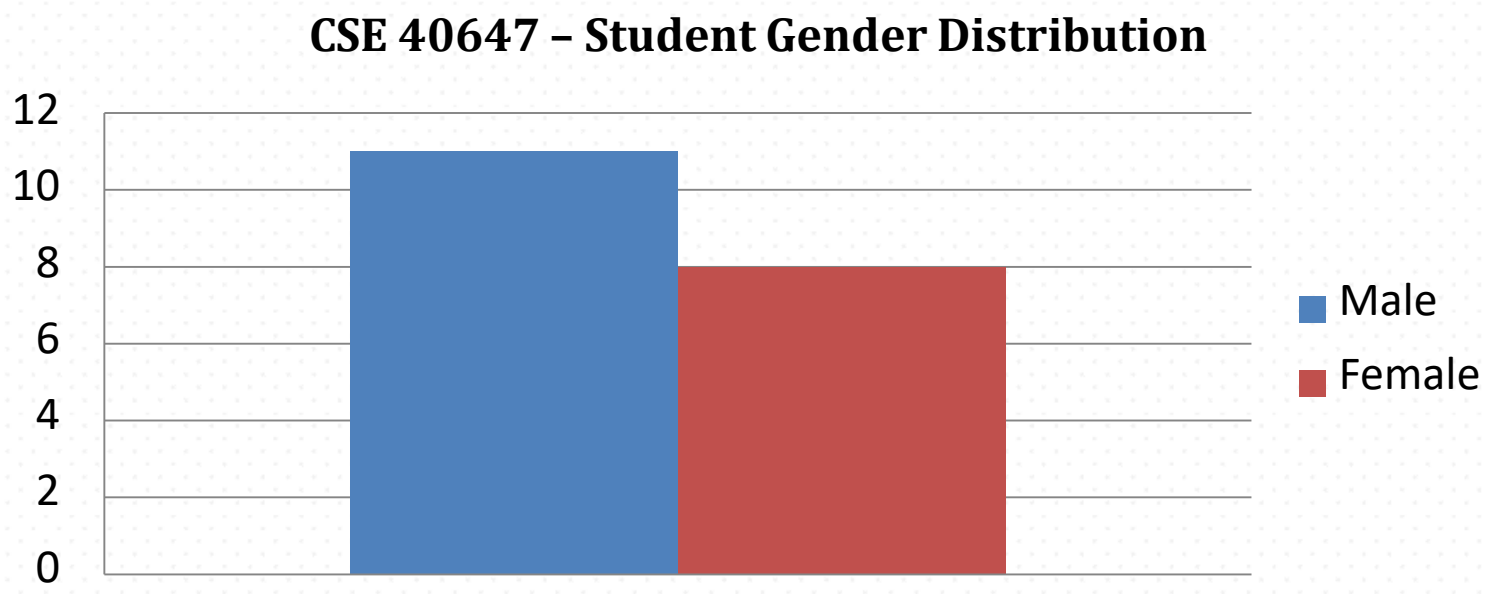
Graphical Integrity



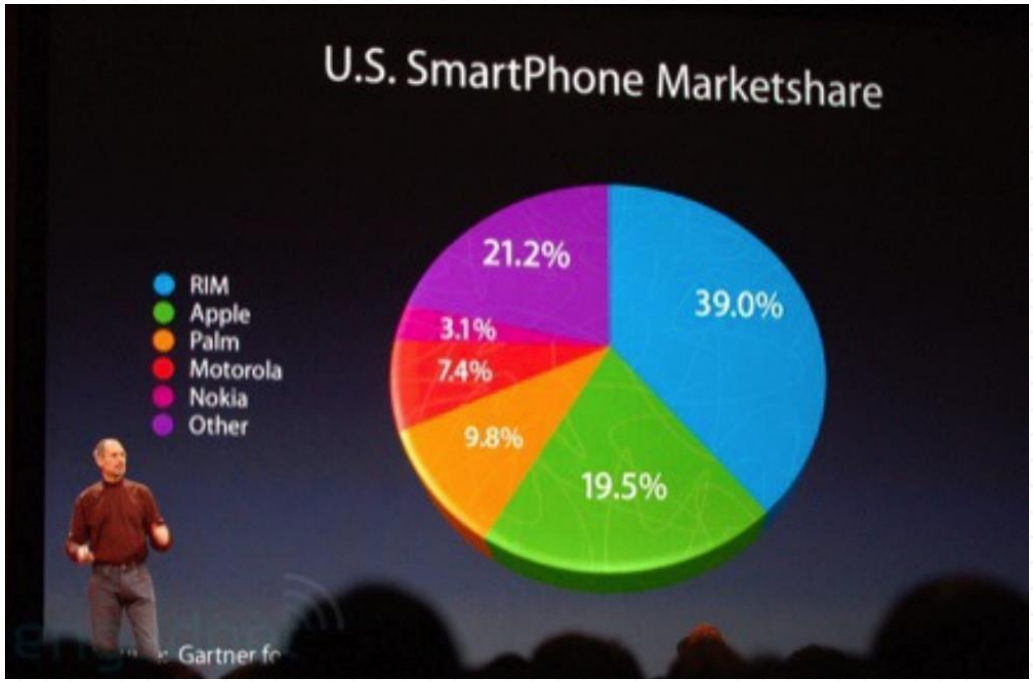
Graphical Integrity



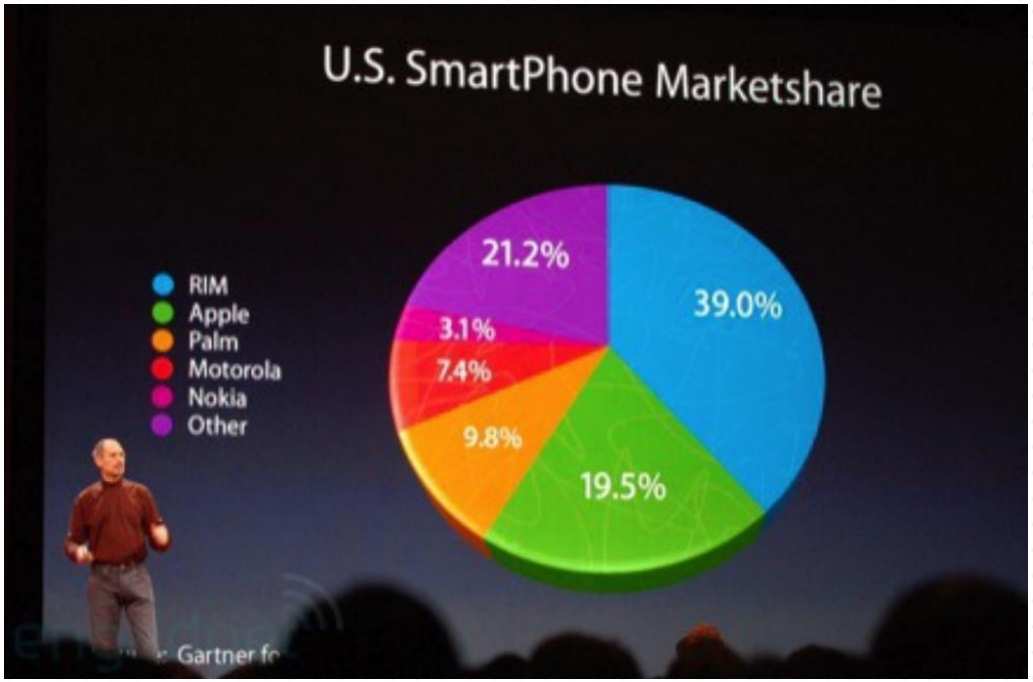
Graphical Integrity

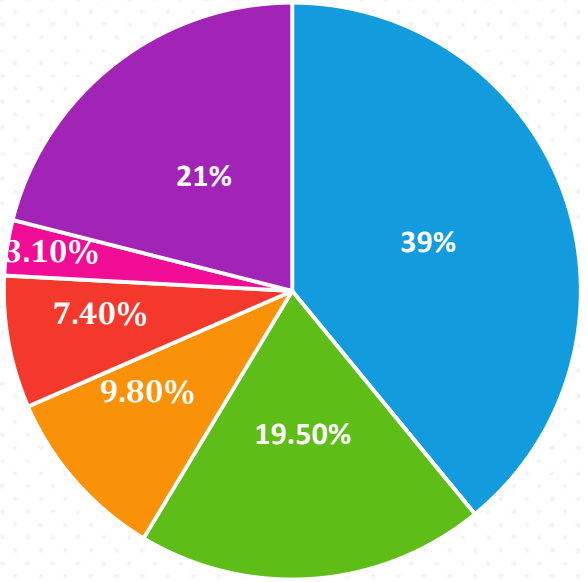


Graphical Integrity



Graphical Integrity





The Lie Factor

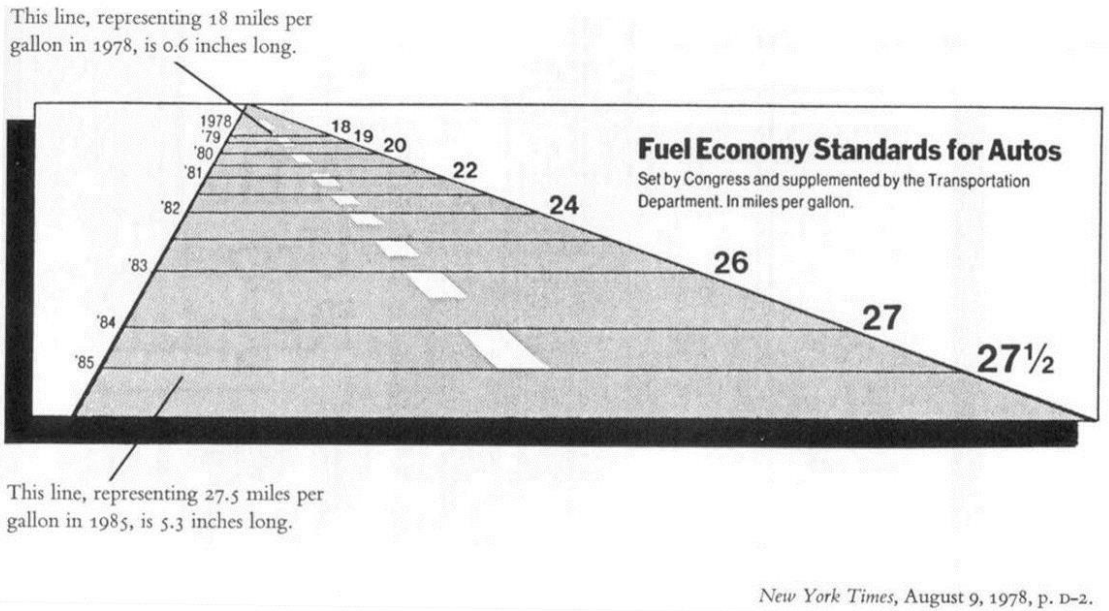
- Coined by Eduard Tufte, the Lie Factor is defined to be a measure of the amount of "distortion" in a graph. That is:

$$\text{Lie Factor} = \frac{\text{size of effect shown in graph}}{\text{size of effect shown in data}}, \text{ where}$$

$$\text{Size of effect} = \frac{|\text{second value - first value}|}{\text{first value}}$$

- If the lie factor is greater than 1, the graph is exaggerating the size of the effect.

The Lie Factor



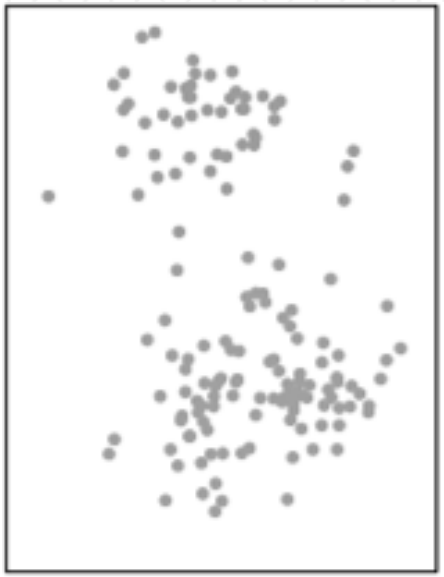
New York Times, August 9, 1978, p. D-2.

$$\text{Lie Factor} = \frac{\frac{5.3 - 0.6}{0.6}}{\frac{27.5 - 18}{18}} = \mathbf{14.8}$$

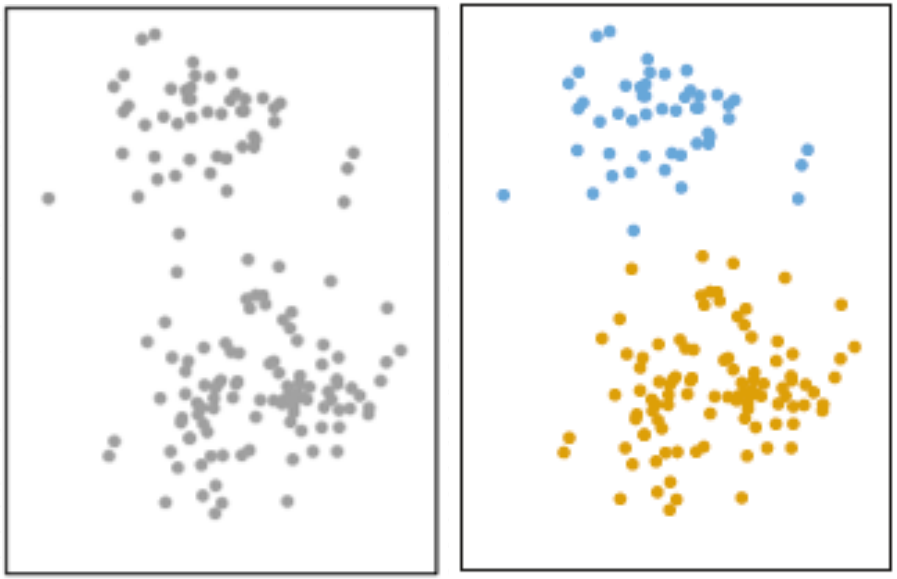
Why Do We Visualize?

0	15600	15600	1200	1200	4800	4800	30	30
0	15600	15600	1200	1200	4800	4800	30	30
15601	15602	15602	4801	4801	4802	4802	15691	15691
15601	15602	15602	4801	4801	4802	4802	15691	15691
1201	4803	4803	1202	1202	4806	4806	1291	1291
1201	4803	4803	1202	1202	4806	4806	1291	1291
4804	4805	4805	4807	4807	4808	4808	5074	5074
4804	4805	4805	4807	4807	4808	4808	5074	5074
1	15603	15603	1203	1203	4809	4809	31	31
1	15603	15603	1203	1203	4809	4809	31	31
15604	15605	15605	4810	4810	4811	4811	15694	15694
15604	15605	15605	4810	4810	4811	4811	15694	15694
1204	4812	4812	1205	1205	4815	4815	1294	1294
1204	4812	4812	1205	1205	4815	4815	1294	1294
4813	4814	4814	4816	4816	4817	4817	5083	5083
4813	4814	4814	4816	4816	4817	4817	5083	5083
2	15606	15606	1206	1206	4818	4818	32	32
2	15606	15606	1206	1206	4818	4818	32	32
15607	15608	15608	4819	4819	4820	4820	15697	15697
15607	15608	15608	4819	4819	4820	4820	15697	15697

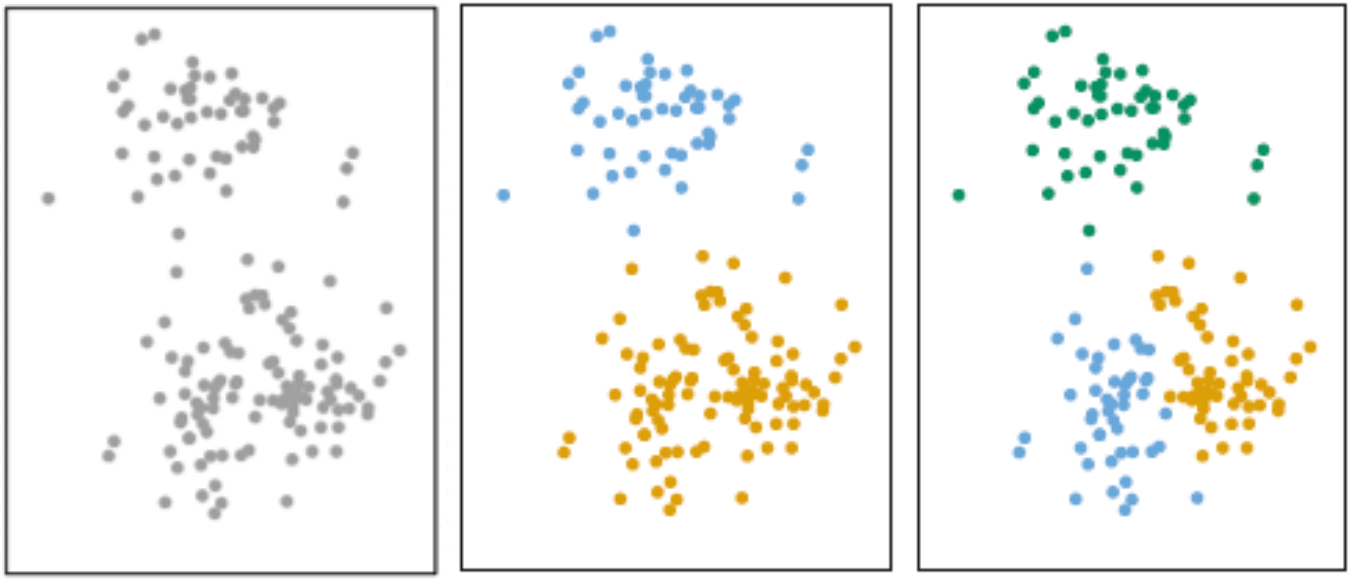
What Do You See?



Is This All?



How About Now?



Chapter 2. Getting to Know Your Data

- Data Objects and Attribute Types
- Basic Statistical Descriptions
- Data Visualization
- **Measuring Data Similarity and Dissimilarity**

Similarity and Dissimilarity

- **Similarity measure** or **similarity function**
  - A real-valued function that quantifies the similarity between two objects
  - Measure how two data objects are alike: The higher value, the more alike
  - Often falls in the range [0,1]: 0: no similarity; 1: completely similar
- **Dissimilarity** (or **distance**) measure
  - Numerical measure of how different two data objects are
  - In some sense, the inverse of similarity: The lower, the more alike
  - Minimum dissimilarity is often 0 (i.e., completely similar)
  - Range [0, 1] or [0, ∞) , depending on the definition

Data Matrix and Dissimilarity Matrix

Data matrix
A data matrix of n data points with l dimensions

Input Features: X
(attributes)

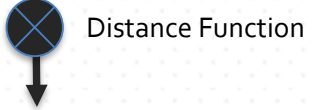
Data Objects (Instances)

	x_1	x_2	..	x_m
1	x_{11}	x_{12}	...	x_{1l}
2	x_{21}	x_{22}	...	x_{2l}
.	⋮	⋮	⋱	⋮
.				
n	x_{n1}	x_{n2}	...	x_{nl}

Data Matrix and Dissimilarity Matrix

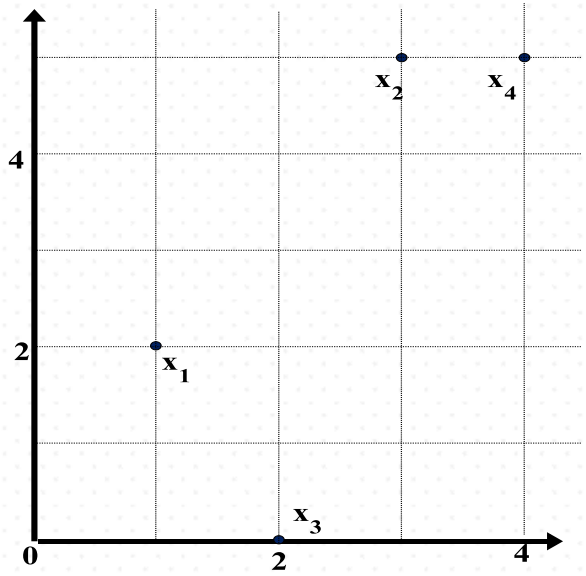
- Dissimilarity (distance) matrix
  - n data points, but registers only the distance d(i, j)
  - Usually symmetric, thus a triangular matrix
  - Distance functions are usually different for real, boolean, categorical, ordinal variables
  - Weights can be associated with different variables based on applications and data semantics

	x_1	x_2	..	x_m
1	x_{11}	x_{12}	...	x_{1l}
2	x_{21}	x_{22}	...	x_{2l}
.	⋮	⋮	⋱	⋮
n	x_{n1}	x_{n2}	...	x_{nl}



$$\begin{pmatrix} 0 & & & \\ d(2,1) & 0 & & \\ & \ldots & & \\ d(n,1) & d(n,2) & \ldots & 0 \end{pmatrix}$$

Example: Euclidean Distance



Data Matrix

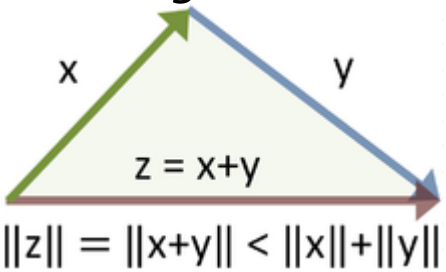
point	attribute	attribute
x1	1	2
x2	3	5
x3	2	0
x4	4	5

Dissimilarity Matrix (by Euclidean Distance)

	x1	*x2*	*x3*	*x4*
x1	0			
x2	3.61	0		
x3	2.24	5.1	0	
x4	4.24	1	5.39	0

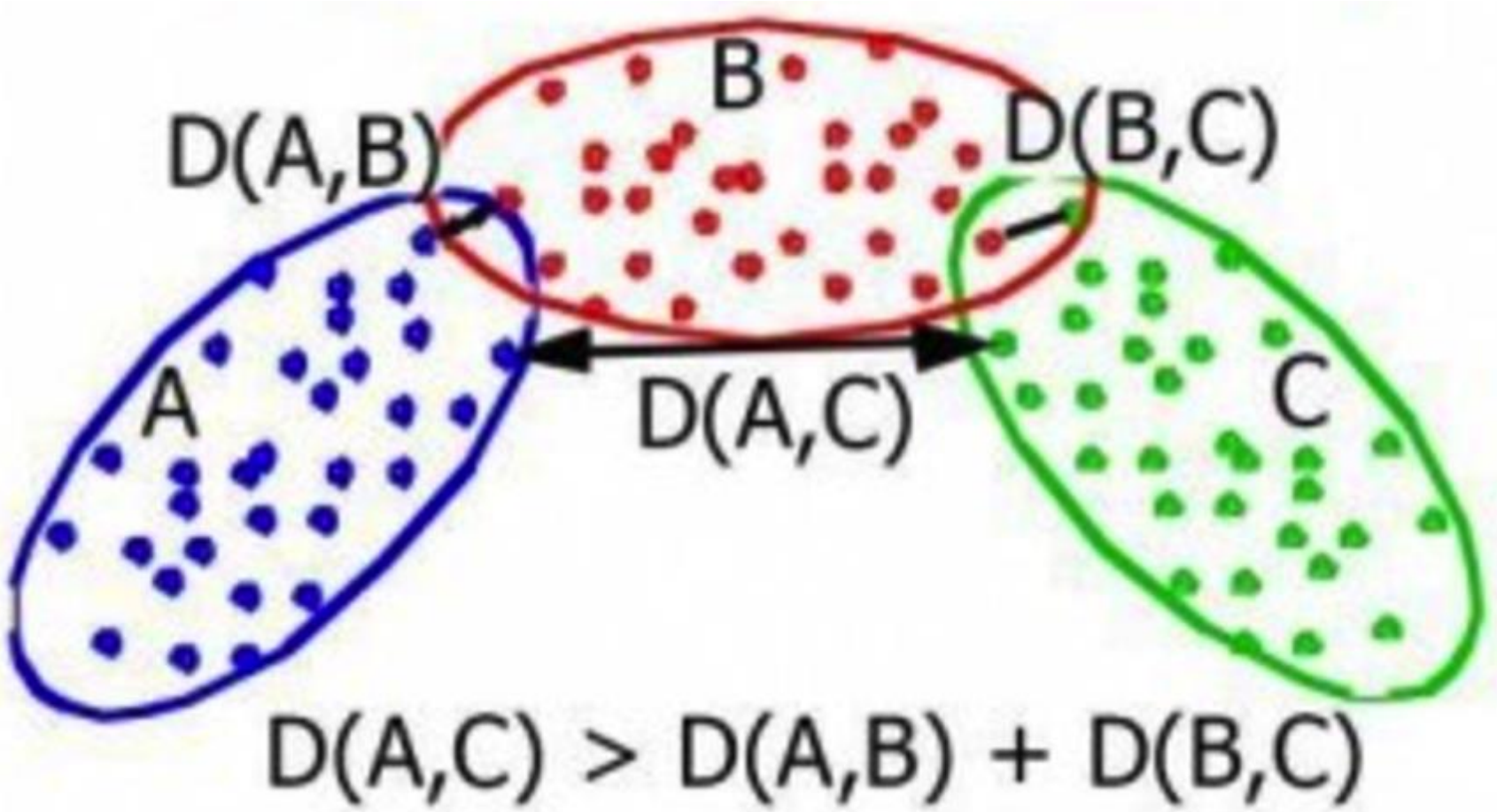
Distance Measures

- Properties
  - $d(i, j) > 0$ if $i \neq j$, and $d(i, i) = 0$ (Positivity)
  - $d(i, j) = d(j, i)$ (**Symmetry**)
  - $d(i, j) \leq d(i, k) + d(k, j)$ (**Triangle Inequality**)
    - *the triangle inequality states that for any triangle, the sum of the lengths of any two sides must be greater than or equal to the length of the remaining side*



- A distance that satisfies these properties is a metric
  - Note: There are nonmetric dissimilarities, e.g., *set difference*

Non Metric Dissimilarity:
~~Triangle Inequality~~



Numeric Data: Minkowski Distance

- Minkowski distance: A popular distance measure

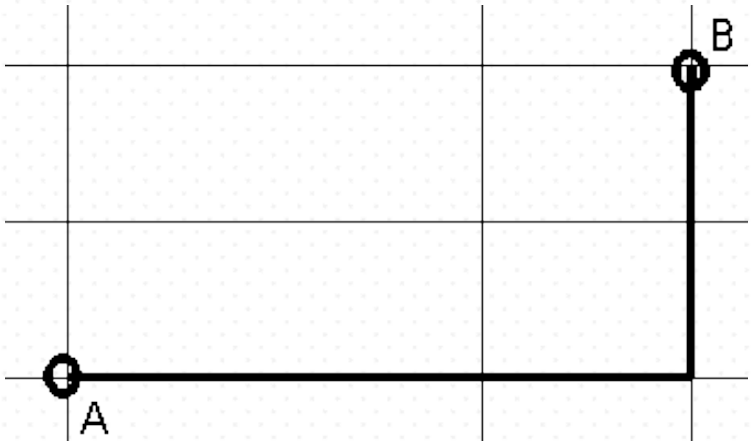
$$d(i, j) = \sqrt[p]{| x_{i1} - x_{j1} |^p + | x_{i2} - x_{j2} |^p + \cdots + | x_{il} - x_{jl} |^p}$$

where $i = (x_{i1}, x_{i2}, \ldots, x_{il})$ and $j = (x_{j1}, x_{j2}, \ldots, x_{jl})$ are two l-dimensional data objects, and p is the order (the distance so defined is called **L-*p* norm**)

Special Cases of Minkowski Distance

- p = 1: (L_1 norm) Manhattan (or city block) distance
  - E.g., the Hamming distance: the number of bits that are different between two binary vectors

$$d(i, j) = | x_{i1} - x_{j1} | + | x_{i2} - x_{j2} | + \cdots + | x_{il} - x_{jl} |$$



Special Cases of Minkowski Distance

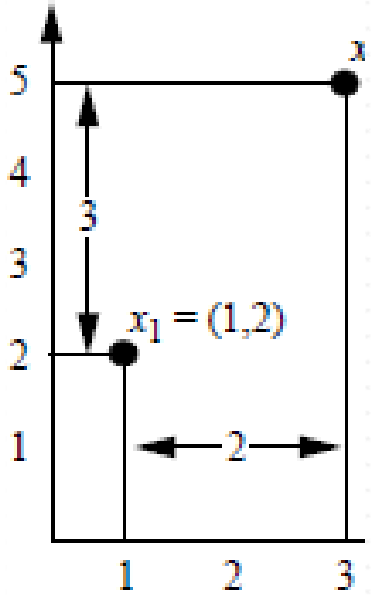
- p = 2: (L_2 norm) Euclidean distance

$$d(i, j) = \sqrt{| x_{i1} - x_{j1} |^2 + | x_{i2} - x_{j2} |^2 + \cdots + | x_{il} - x_{jl} |^2}$$

Special Cases of Minkowski Distance

- $p \to \infty$: (L_{max} norm, L_∞ norm) "supremum" distance
  - The maximum difference between any component (attribute) of the vectors

$$d(i,j) = \lim_{p\to\infty} \sqrt[p]{|x_{i1}-x_{j1}|^p + |x_{i2}-x_{j2}|^p + \cdots + |x_{il}-x_{jl}|^p} = \max_{f=1}^{l} |x_{if}-x_{if}|$$



In Class Activity

- Manhattan

$$d(i,j) = |x_{i1} - x_{j1}| + |x_{i2} - x_{j2}| + \cdots + |x_{il} - x_{jl}|$$

- Euclidean distance

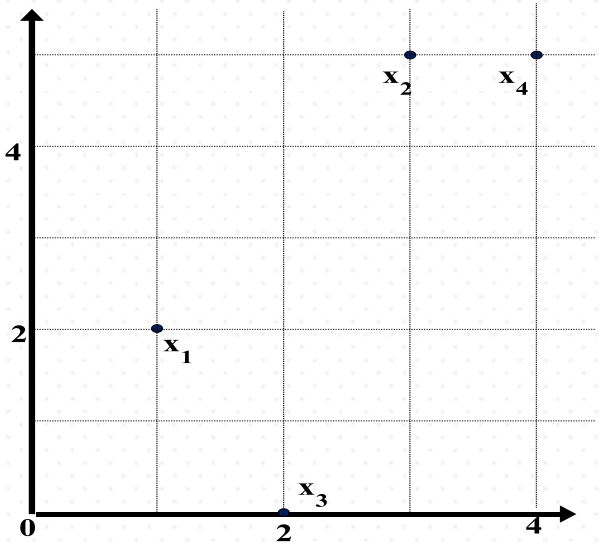
$$d(i,j) = \sqrt{|x_{i1} - x_{j1}|^2 + |x_{i2} - x_{j2}|^2 + \cdots + |x_{il} - x_{jl}|^2}$$

- Supremum distance (Chebyshev distance)

$$d(i,j) = \lim_{p \to \infty} \sqrt[p]{|x_{i1} - x_{j1}|^p + |x_{i2} - x_{j2}|^p + \cdots + |x_{il} - x_{jl}|^p} = \max_{f=1}^{l} |x_{if} - x_{if}|$$

Example: Minkowski Distance at Special Cases

point	attribute 1	attribute 2
x1	1	2
x2	3	5
x3	2	0
x4	4	5



Manhattan (L_1)

L	x1	x2	x3	x4
x1	0	X	X	X
x2		0	X	X
x3			0	X
x4				0

Euclidean (L_2)

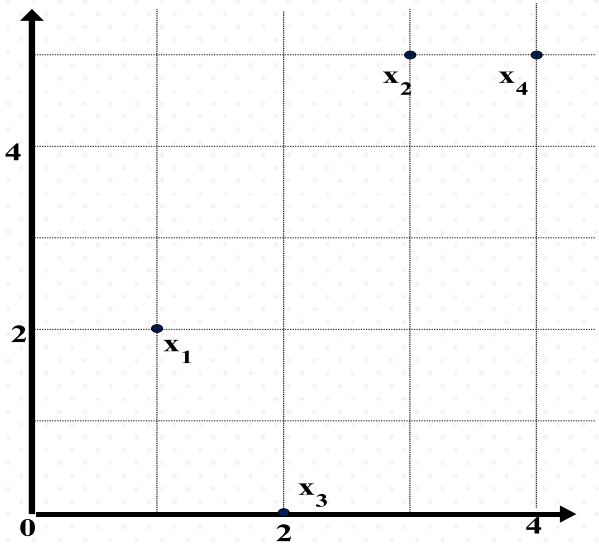
L	x1	x2	x3	x4
x1	0	X	X	X
x2		0	X	X
x3			0	X
x4				0

Supremum (L_∞)

L	x1	x2	x3	x4
x1	0	X	X	X
x2		0	X	X
x3			0	X
x4				0

Example: Minkowski Distance at Special Cases

point	attribute 1	attribute 2
x1	1	2
x2	3	5
x3	2	0
x4	4	5



Manhattan (L_1)

L	x1	x2	x3	x4
x1	0			
x2	5	0		
x3	3	6	0	
x4	6	1	7	0

Euclidean (L_2)

L2	x1	x2	x3	x4
x1	0			
x2	3.61	0		
x3	2.24	5.1	0	
x4	4.24	1	5.39	0

Supremum (L_∞)

L_∞	x1	x2	x3	x4
x1	0			
x2	3	0		
x3	2	5	0	
x4	3	1	5	0

Real World Considerations

point	attribute 1	attribute 2
x1	1	2
x2	3	5
x3	2	0
x4	4	5

Real World Considerations

userId	Call duration	SMS Total	Data Counter MB
1	25000	24	4
2	40000	27	5
3	55000	32	7
4	27000	25	5
5	53000	30	5

Real World Considerations

userId	Call duration	SMS Total	Data Counter MB
1	25000	24	4
2	40000	27	5
3	55000	32	7
4	27000	25	5
5	53000	30	5

Distance User 1-2

$$\sqrt{(25000-40000)^2+(24-27)^2+(4-5)^5}$$

15000.0003333333

Real World Considerations

userId	Call duration	SMS Total	Data Counter MB
1	25000	24	4
2	40000	27	5
3	55000	32	7
4	27000	25	5
5	53000	30	5

	1	2	3	4	5
1	0				
2	15000	0			
3	30000	15000	0		
4	2000	13000	28000	0	
5	28000	13000	2000	26000	0

Solution: Normalization

The goal of normalization is to make an entire set of values have a particular property.

Data Transformation: Normalization

- Normalization is often performed on data to remove amplitude variation and only focus on the underlying distribution shape.
- Makes training less sensitive to the scale of features
- Sometimes used in order to speed up the convergence.

Data Transformation: Normalization

- Min-max normalization
- Z-score normalization
- Normalization by decimal scaling

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.

Min-Max Normalization: Example

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 (zero-mean) 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 :

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

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

Decimal Scaling Normalization

- Suppose that the recorded values of F range from – 986 to 917. The maximum absolute value of F is 986. To normalize by decimal scaling, we therefore divide each value by 1,000 (i.e., j = 3) so that –986 normalizes to –0.986 and 917 normalizes to 0.917.

iPython Examples

Proximity Measure for Binary Attributes

A contingency table for binary data

Object *j*

Object *i*	1	0	sum
1	q	r	$q+r$
0	s	t	$s+t$
sum	$q+s$	$r+t$	p

- Distance measure for *symmetric* binary variables:

$$d(i,j) = \frac{r+s}{q+r+s+t}$$

- Distance measure for *asymmetric* binary variables:

$$d(i,j) = \frac{r+s}{q+r+s}$$

Proximity Measure for Binary Attributes

A contingency table for binary data

Object *i* \ Object *j*	1	0	sum
1	q	r	$q+r$
0	s	t	$s+t$
sum	$q+s$	$r+t$	p

- Jaccard coefficient (similarity measure for *asymmetric* binary variables):

$$sim_{Jaccard}(i,j) = \frac{q}{q+r+s}$$

- Note: Jaccard coefficient is the same as "coherence":

$$coherence(i,j) = \frac{sup(i,j)}{sup(i)+sup(j)-sup(i,j)} = \frac{q}{(q+r)+(q+s)-q}$$

Ordinal Variables

- An ordinal variable can be discrete or continuous
- Order is important, e.g., rank (e.g., freshman, sophomore, junior, senior)
- Can be treated like interval-scaled
  - Replace an ordinal variable value by its rank and map the range of each variable onto [0, 1]:
    - Example: freshman: 0; sophomore: 1/3; junior: 2/3; senior 1
      - Then distance: d(freshman, senior) = 1, d(junior, senior) = 1/3
  - Compute the dissimilarity using methods for interval-scaled variables

Example: Ordinal Data

Object	Ordinal Data
1	Excellent
2	Fair
3	Good
4	Excellent

Example: Rank

Object	Ordinal Data
1	Excellent
2	Fair
3	Good
4	Excellent

Object	Ordinal Data
1	3
2	1
3	2
4	3

Example: Normalize

Object	Ordinal Data
1	Excellent
2	Fair
3	Good
4	Excellent

Object	Ordinal Data
1	3
2	1
3	2
4	3

Object	Ordinal Data
1	1
2	0
3	.5
4	1

Example: Compute Distance

Object	Ordinal Data
1	Excellent
2	Fair
3	Good
4	Excellent

Object	Ordinal Data
1	3
2	1
3	2
4	3

Object	Ordinal Data
1	1
2	0
3	.5
4	1

Distance

0			
1	0		
.5	.5	0	
0	1	.5	0

Proximity Measure for Categorical Attributes

- Categorical data, also called nominal attributes
  - Example: Color (red, yellow, blue, green), profession, etc.
- Method 1: Simple matching
  - m: # of matches, p: total # of variables

$$d(i, j) = \frac{p - m}{p}$$

- Method 2: Use a large number of binary attributes
  - Creating a new binary attribute for each of the M nominal states

Example

Object	Ordinal Data
1	Blue
2	Red
3	Green
4	Blue

P=1

$$d(i,j)=\frac{p-m}{p}$$

0			
1	0		
1	1	0	
0	1	1	0

Cosine Similarity of Two Vectors

- **Cosine measure:** If d_1 and d_2 are two vectors (e.g., term-frequency vectors), then

$$cos(d_1, d_2) = \frac{d_1 \bullet d_2}{\| d_1 \| \times \| d_2 \|}$$

where • indicates vector dot product, ||*d*||: the length of vector *d*

Cosine Similarity of Two Vectors

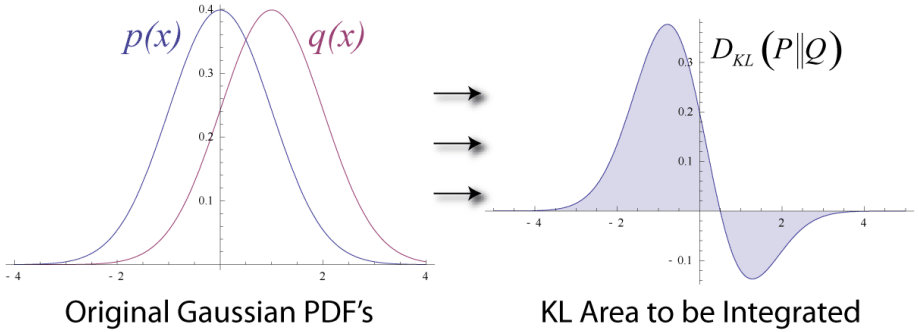
- A **document** can be represented by a bag of terms or a long vector, with each attribute recording the *frequency* of a particular term (such as word, keyword, or phrase) in the document

Document	*team*	*coach*	*hockey*	*baseball*	*soccer*	*penalty*	*score*	*win*	*loss*	*season*
Document1	5	0	3	0	2	0	0	2	0	0
Document2	3	0	2	0	1	1	0	1	0	1
Document3	0	7	0	2	1	0	0	3	0	0
Document4	0	1	0	0	1	2	2	0	3	0

- Other vector objects: Gene features in micro-arrays
- Applications: Information retrieval, biologic taxonomy, gene feature mapping, etc.

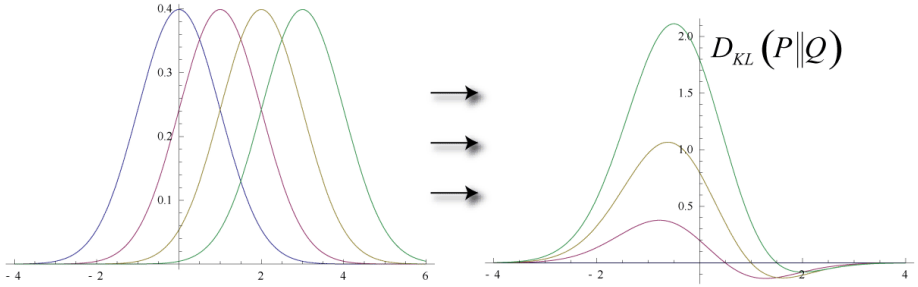
KL Divergence: Comparing Two Probability Distributions

- *The Kullback-Leibler (KL) divergence:* Measure the **difference** between two probability distributions over the same variable *x*
  - From information theory, closely related to *relative entropy*, *information divergence*, and *information for discrimination*
- $D_{KL}(p(x) \| q(x))$: divergence of $q(x)$ from $p(x)$, measuring the **information lost when *q*(*x*) is used to approximate *p*(*x*)**



Original Gaussian PDF's

KL Area to be Integrated



Discrete form

$$D_{KL}(p(x)||q(x)) = \sum_{x \in X} p(x) \ln \frac{p(x)}{q(x)}$$

Continuous form

$$D_{KL}(p(x)||q(x)) = \int_{-\infty}^{\infty} p(x) \ln \frac{p(x)}{q(x)} dx$$

Discussion

- Can you use Matrix Multiplication to compute Cosine Similarity between every pair of objects?
- Can you use KL divergence to find suspiciousness?
- Can you use KL divergence to find representative phrases for specific topics?

Summary

- Data attribute types: nominal, binary, ordinal …
- Gain insight into the data by:
  - Basic data description: central tendency, outlierness
  - Data visualization
  - Measure data similarity and dissimilarity
- Above steps are the beginning of data preprocessing

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