Data Processing

Álvaro Figueira, PhD.

MSc in Data Science - Data Visualization

[dcc] DEPARTAMENTO DE CIÊNCIA DE COMPUTADORES
FACULDADE DE CIÊNCIAS DA UNIVERSIDADE DO PORTO

FC FACULDADE DE CIÊNCIAS UNIVERSIDADE DO PORTO

U.PORTO

87

Introduction

- Data can be collected or be produced by sensors, also can be created by computational simulations
- Data can be raw (not processed) or can be derived using some process, such as, noise removal, scale, or interpolation

90

Introduction

- Data is composed of n instances/rows/records

$$(r_1, r_2, r_3, \ldots, r_n)$$

- Each one containing m (one or more) observations/variables (vector)

$$(v_1, v_2, v_3, \ldots, v_m)$$

- Each observation can be a number, symbol, string, or a more complex structure

91

Introduction

- A variable can be
  - Independent (iv_i): its value is not affected or controlled by any other variable, for instance, the time in a time-series
  - Dependent (dv_j): it is affected by the variation of one or more independent variables

- A data instance can be defined as

$$r_i = (iv_1, iv_2, \ldots, iv_{m_i}, dv_1, dv_2, \ldots, dv_{m_d})$$

- With m_i independent variables and m_d dependent variables $m=m_i+m_d$

92

The example

ID/Name	Nota	TTO	AV1	AV2	AV3	BT1	BT2	BT3	TIDZ1	TODZ1	TIDZ2	TODZ2	AT1	AT2	Download	Clicks	Forum	Sum	Ano	Class
56	17	743	0	2	0	6	5	0	0	3	0	2	91	0	2	154	37	24,047	2017	d
85	9	241	3	4	0	3	3	0	0	1	0	0	302	0	16	68	1	27,037	2017	b
28	13	202	3	11	1	3	-1	5	0	3	0	0	1641	0	11	59	1	37,049	2017	c
111	14	468	0	0	0	6	7	0	0	5	0	0	3124	0	6	102	12	28,067	2017	c
24	14	212	24	4	3	0	3	3	0	4	0	0	135	0	6	51	0	28,038	2017	c
57	10	56	3	2	5	3	5	1	0	0	0	0	4333	0	7	20	1	23,024	2017	b
154	16	262	3	0	5	3	7	1	0	3	0	0	1223	0	22	58	0	16,039	2017	d
14	10	498	4	1	0	2	6	0	0	3	0	2	10451	0	23	137	5	40,075	2017	b
123	9	427	4	1	5	2	6	1	0	1	1	1	5947	0	24	106	1	32,083	2017	b
86	11	305	1	0	0	5	7	6	0	2	0	0	60	0	7	72	2	29,062	2017	b
58	11	170	1	6	5	5	1	1	0	1	0	0	80	0	10	34	0	24,044	2017	b
8	12	152	3	2	0	3	5	0	0	2	0	1	76	0	2	38	1	18,034	2017	c
113	16	1033	5	0	5	1	7	1	0	2	0	1	175	0	17	244	13	50,127	2017	d
169	14	359	2	2	6	4	5	0	0	4	0	1	3963	0	9	89	3	26,056	2017	c
37	14	230	5	6	0	1	1	6	0	1	0	0	135	0	15	46	1	27,047	2017	c
155	12	199	4	10	0	2	-1	0	0	1	0	0	2828	0	13	53	0	27,046	2017	c
150	16	195	14	2	0	0	5	0	0	1	0	1	2925	0	13	49	0	35,056	2017	d
9	13	243	5	5	5	1	2	1	0	1	0	0	59	0	9	56	0	20,029	2017	c
6	14	247	5	4	0	1	3	0	0	1	0	0	139	0	11	54	5	21,041	2017	c
156	10	172	0	4	0	0	3	0	0	1	0	0	2782	0	3	52	5	25,038	2017	b
147	10	197	1	3	2	5	4	4	0	1	0	0	213	0	16	63	4	27,039	2017	b
97	13	469	1	0	5	5	7	1	0	4	0	0	302	0	16	105	5	28,082	2017	c
115	13	183	5	5	0	1	2	0	0	1	0	0	4446	0	11	42	0	29,041	2017	c
39	12	176	4	0	5	2	7	1	0	2	0	0	117	0	15	48	0	30,039	2017	c
59	16	328	9	2	5	-1	5	1	0	5	0	0	251	0	11	87	2	24,067	2017	d
5	11	112	5	6	5	1	1	1	0	1	0	0	2948	0	9	34	0	29,038	2017	b
7	8	206	4	5	4	2	2	2	0	1	0	0	4346	0	13	50	0	35,053	2017	b
71	12	517	0	0	0	6	7	6	0	3	0	0	84	0	19	119	34	34,060	2017	c
72	12	211	6	6	5	0	1	1	0	1	0	0	99	0	9	45	0	25,044	2017	c

Data Types

- Each variable represents a piece of information that can be classified as **ordinal** (numeric) and **nominal** (non-numeric)
- Ordinal data
  - Binary: 0 and 1 values
  - Discrete: integer values
  - Continuous: real values
- Nominal data
  - Categorical: value from a finite list of values (ex. Red, Blue, Green)
  - Ranked: categorical values with order (ex. Small, Medium, Large)
  - Arbitrary: infinite values without order (ex. Address)

Data Types – try to explain the dataset

ID/Name	Nota	TTO	AV1	AV2	AV3	BT1	BT2	BT3	TIDZ1	TODZ1	TIDZ2	TODZ2	AT1	AT2	Download	Clicks	Forum	Sum	Ano	Class
56	17	743	0	2	0	6	5	0	0	3	0	2	91	0	2	154	37	24,047	2017	d
85	9	241	3	4	0	3	3	0	0	1	0	0	302	0	16	68	1	27,037	2017	b
28	13	202	3	11	1	3	-1	5	0	3	0	0	1641	0	11	59	1	37,049	2017	c
111	14	468	0	0	0	6	7	0	0	5	0	0	3124	0	6	102	12	28,067	2017	c
24	14	212	24	4	3	0	3	3	0	4	0	0	135	0	6	51	0	28,038	2017	c
57	10	56	3	2	5	3	5	1	0	0	0	0	4333	0	7	20	1	23,024	2017	b
154	16	262	3	0	5	3	7	1	0	3	0	0	1223	0	22	58	0	16,039	2017	d
14	10	498	4	1	0	2	6	0	0	3	0	2	10451	0	23	137	5	40,075	2017	b
123	9	427	4	1	5	2	6	1	0	1	1	1	5947	0	24	106	1	32,083	2017	b
86	11	305	1	0	0	5	7	6	0	2	0	0	60	0	7	72	2	29,062	2017	b
58	11	170	1	6	5	5	1	1	0	1	0	0	80	0	10	34	0	24,044	2017	b
8	12	152	3	2	0	3	5	0	0	2	0	1	76	0	2	38	1	18,034	2017	c
113	16	1033	5	0	5	1	7	1	0	2	0	1	175	0	17	244	13	50,127	2017	d
169	14	359	2	2	6	4	5	0	0	4	0	1	3963	0	9	89	3	26,056	2017	c
37	14	230	5	6	0	1	1	6	0	1	0	0	135	0	15	46	1	27,047	2017	c
155	12	199	4	10	0	2	-1	0	0	1	0	0	2828	0	13	53	0	27,046	2017	c
150	16	195	14	2	0	0	5	0	0	1	0	1	2925	0	13	49	0	35,056	2017	d
9	13	243	5	5	5	1	2	1	0	1	0	0	59	0	9	56	0	20,029	2017	c
6	14	247	5	4	0	1	3	0	0	1	0	0	139	0	11	54	5	21,041	2017	c
156	10	172	0	4	0	0	3	0	0	1	0	0	2782	0	3	52	5	25,038	2017	b
147	10	197	1	3	2	5	4	4	0	1	0	0	213	0	16	63	4	27,039	2017	b
97	13	469	1	0	5	5	7	1	0	4	0	0	302	0	16	105	5	28,082	2017	c
115	13	183	5	5	0	1	2	0	0	1	0	0	4446	0	11	42	0	29,041	2017	c
39	12	176	4	0	5	2	7	1	0	2	0	0	117	0	15	48	0	30,039	2017	c
59	16	328	9	2	5	-1	5	1	0	5	0	0	251	0	11	87	2	24,067	2017	d
5	11	112	5	6	5	1	1	1	0	1	0	0	2948	0	9	34	0	29,038	2017	b
7	8	206	4	5	4	2	2	2	0	1	0	0	4346	0	13	50	0	35,053	2017	b
71	12	517	0	0	0	6	7	6	0	3	0	0	84	0	19	119	34	34,060	2017	c
72	12	211	6	6	5	0	1	1	0	1	0	0	99	0	9	45	0	25,044	2017	c

95

Data Processing

- Sometimes it is preferable to visualize raw data (typically in medical applications), but for some applications it is necessary some kind of data pre-processing

16/06/17, 16:45	Álvaro Pedro de Barros Borges Reis Figueira	Hu Vi [redacted] s	Workshop: Submissão e revisão de artigos	Workshop	Fase de Avaliação reavaliada	The user with id '1032' has had their assessment attempt reevaluated for the workshop with the course module id '70597'.	web	188.37.109.192
16/06/17, 16:45	Álv Rei [redacted]	[redacted] ão	Workshop: Submissão e revisão de artigos	Workshop	Fase de Avaliação reavaliada	The user with id '1032' has had their assessment attempt reevaluated for the workshop with the course module id '70597'.	web	188.37.109.192
16/06/17, 16:26	Ric [redacted]	[redacted]	Unidade: Comunicação Técnica (FCUP-DPI1001-2016/2017-2S)	Sistema	Disciplina visualizada	The user with id '27545' viewed the course with id '419'.	web	185.101.177.17
16/06/17, 15:27	Jos [redacted]	[redacted]	Unidade: Comunicação Técnica (FCUP-DPI1001-2016/2017-2S)	Sistema	Disciplina visualizada	The user with id '23041' viewed the course with id '419'.	web	89.155.159.107
16/06/17, 15:27	Jos [redacted]	[redacted]	Página: Notas dos testes	Página	Módulo de disciplina visualizado	The user with id '23041' viewed the 'page' activity with the course module id '71704'.	web	89.155.159.107
16/06/17, 15:27	Jos [redacted]	[redacted]	Unidade: Comunicação Técnica (FCUP-DPI1001-2016/2017-2S)	Sistema	Disciplina visualizada	The user with id '23041' viewed the course with id '419'.	web	89.155.159.107
16/06/17, 15:24	Álv Rei [redacted]	[redacted]	Workshop: Submissão e revisão de artigos	Workshop	Módulo de disciplina visualizado	The user with id '1032' viewed the 'workshop' activity with the course module id '70597'.	web	188.37.109.192
16/06/17, 15:24	Álv Rei [redacted]	[redacted]	Unidade: Comunicação Técnica (FCUP-DPI1001-2016/2017-2S)	Sistema	Disciplina visualizada	The user with id '1032' viewed the course with id '419'.	web	188.37.109.192
16/06/17, 11:21	Sim [redacted]	[redacted]	Unidade: Comunicação Técnica (FCUP-DPI1001-2016/2017-2S)	Sistema	Disciplina visualizada	The user with id '23091' viewed the course with id '419'.	web	193.136.39.100
16/06/17, 10:36	Car de [redacted]	[redacted]	Unidade: Comunicação Técnica (FCUP-DPI1001-2016/2017-2S)	Sistema	Disciplina visualizada	The user with id '23125' viewed the course with id '419'.	web	193.136.24.134
16/06/17, 09:55	Car de [redacted]	[redacted]	Página: Notas dos testes	Página	Módulo de disciplina visualizado	The user with id '23125' viewed the 'page' activity with the course module id '71704'.	web	193.136.24.134
16/06/17, 09:55	Car de [redacted]	[redacted]	Unidade: Comunicação Técnica (FCUP-DPI1001-2016/2017-2S)	Sistema	Disciplina visualizada	The user with id '23125' viewed the course with id '419'.	web	193.136.24.134
16/06/17, 08:18	Rui Bal [redacted]	[redacted]	Ficheiro: Apresentacao-CT-2017	Ficheiro	Módulo de disciplina visualizado	The user with id '12521' viewed the 'resource' activity with the course module id '62986'.	web	188.80.45.99

96

96

Metadata and Statistics

ID/Name	Nota	TTO	AV1	AV2	AV3	BT1	BT2	BT3	TIDZ1	TODZ1	TIDZ2	TODZ2	AT1	AT2	Download	Clicks	Forum	Sum	Ano	Class
56	17	743	0	2	0	6	5	0	0	3	0	2	91	0	2	154	37	24,047	2017	d
85	9	241	3	4	0	3	3	0	0	1	0	0	302	0	16	68	1	27,037	2017	b
28	13	202	3	11	1	3	-1	5	0	3	0	0	1641	0	11	59	1	37,049	2017	c
111	14	468	0	0	0	6	7	0	0	5	0	0	3124	0	6	102	12	28,067	2017	c

- Metadata can help to interpret the data, providing information such as
  - Reference points for measures
  - Employed units
  - Symbols employed as missing values
  - Measurements resolution
- UCI Machine Learning Repository
  - http://archive.ics.uci.edu/ml/index.php (Iris dataset)
- R datasets
  - https://stat.ethz.ch/R-manual/R-devel/library/datasets/html/00Index.html

97

97

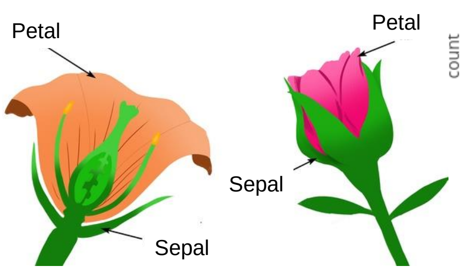
Metadata and Statistics

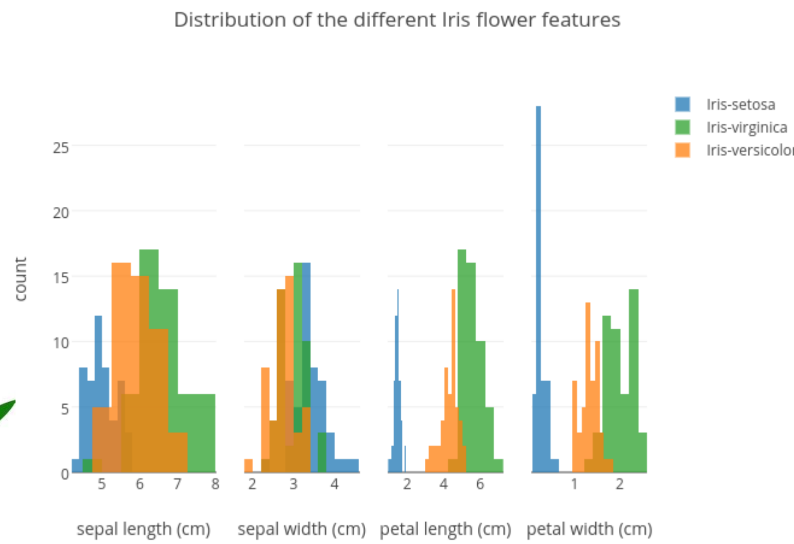
- Statistical analysis can provide useful information such as
  - **Outlier** detection (possible wrong measurements)
  - Identification of similar **clusters**
  - Identification of **redundant** variables/instances using **correlation**
- Histograms and violin plots can be used to analyze the **data distribution**

98

98

Histogram





99

99

Simple Method for Outlier Detection

- Suppose that the variable *j* has a Gaussian distribution $N(\mu_j, \sigma_j)$. Then, first transform the data so that it presents $\mu_j = 0$ e $\sigma_j = 1$
- If x_{ij} is the value of the variable *j* on the instance *i*, and *c* is a constant, the probability that $|x_{ij}| \geq c$ rapidly decreases as *c* increases
- If $\alpha = prob(|x_{ij}| \geq c)$, a data instance will be an outlier if

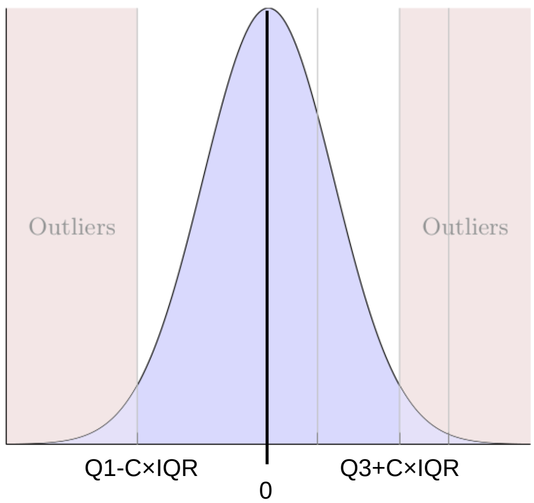
$$|x_{ij}| \geq c$$

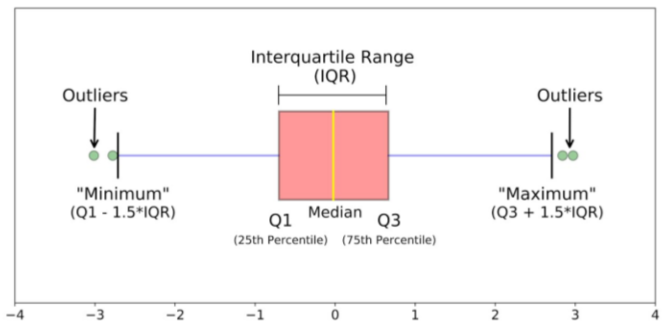
- **Simple method:** x is outlier if x is out of [Q1 – 1.5 × IQR , Q3 + 1.5 × IQR]
- Usually C=1.5 , 2, 2.5, ….
- IQR is the "interquartile range"

100

100

Simple Method for Outlier Detection





Simple Method to Detect Redundant Variables

- Let x_i and x_j two variables, first calculate their correlation

$$cor(x_i, x_j) = \frac{cov(x_i, x_j)}{var(x_i)\, var(x_j)}$$

- With $cov(x_i,x_j)$ given by

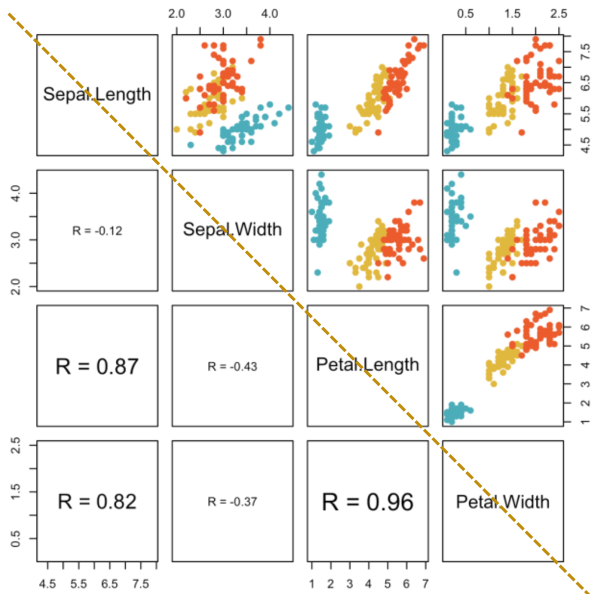
$$cov(x_i, x_j) = \frac{1}{m-1} \sum_{k=1}^{m} (x_{ki} - \mu_i)(x_{kj} - \mu_j)$$

- And $var(x_j)$ given by

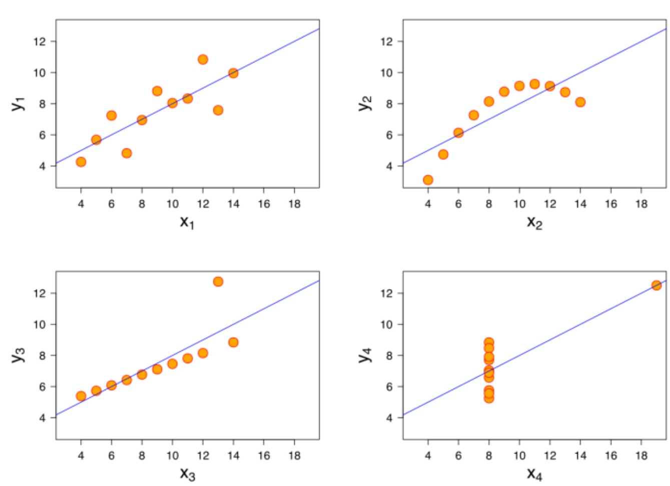
$$var(x_i) = \sigma^2$$

- Values of $|cor(x_i,x_j)|$ close to *1* indicates high correlation, so either x_i or x_j can be discarded

Simple Method to Detect Redundant Variables



Correlation does not describe everything!



They all have correlation of 0.816

By Anscombe.svg: SchutzDerivative works of this file:(label using subscripts): Avenue - Anscombe.svg, CC BY-SA 3.0, https://commons.wikimedia.org/w/index.php?curid=9838454

Missing Values and Data Cleaning

- On "real" data it is normal that some data be missing or be wrong
- Common strategies to address such issue
  - Discard the data instance with defect. Note it can represent an important data loss
  - Assign a sentinel value. However, the sentinel value cannot be used on the calculations
  - Calculate a replacement value. However, data imputation might be risky

105

Data Imputation

- Two simple data imputation methods
  - Assign the average value: can hide outlier values
  - Assign a value based on the nearest neighbors: on the neighborhood calculation it is difficult to know if there are more relevant attributes

106

Data Imputation

It is probably one of the most difficult steps in data mining. Must be done with care!

Fisher's *Iris* Data

Sepal length	Sepal width	Petal length	Petal width	Species
5.1	3.5	1.4	0.2	*I. setosa*
4.9	3.0	1.4	0.2	*I. setosa*
4.7	3.2	1.3	0.2	*I. setosa*
4.6	3.1	1.5	0.2	*I. setosa*
5.0	3.6	1.4	0.2	*I. setosa*
5.4	3.9	1.7	0.4	*I. setosa*
4.6	3.4	1.4	$\boxed{*}$	*I. setosa*
5.0	3.4	1.5	0.2	*I. setosa*
4.4	2.9	1.4	0.2	*I. setosa*
4.9	3.1	1.5	0.1	*I. setosa*
5.4	3.7	1.5	0.2	*I. setosa*
4.8	3.4	1.6	0.2	*I. setosa*
4.8	3.0	1.4	0.1	*I. setosa*

		duration	
16/06/17, 10:36	Person X	-	Unidade: Comunicação Técnica (FCUP-DPI1001-2016/2017-2S)
16/06/17, 09:55	Person X	-	Página: Notas dos testes
16/06/17, 09:55	Person X	-	Unidade: Comunicação Técnica (FCUP-DPI1001-2016/2017-2S)

107

107

Normalization

- On applications that involve instances' comparison, one scenario can distort the result and introduce bias
  - When the Euclidean norm of the vectors (line or column) they represent are too different
- A possible solution is normalization
  - Transform the data so that they present a desired statistical property

108

108

Normalization

- One process consists on transforming the data so that the values range is in *[0,1]*

- If the maximum X_j^{max} and minimum X_j^{min} values are know, so

$$x_{ij} = (x_{ij} - X_j^{min}) / (X_j^{max} - X_j^{min})$$

- On specific cases it could be interesting to use the known maximum and minimum values, such as on percentages

111

111

Normalization

- Another known normalization is the standardization. It transform the data so that the average is 0 and the standard deviation is 1

$$x_{ij} = (x_{ij} - \mu_j) / \sigma_j$$

- However, normalization can distort the data, for instance, in the presence of outliers, and the data can be flattened

112

112

Interpolation

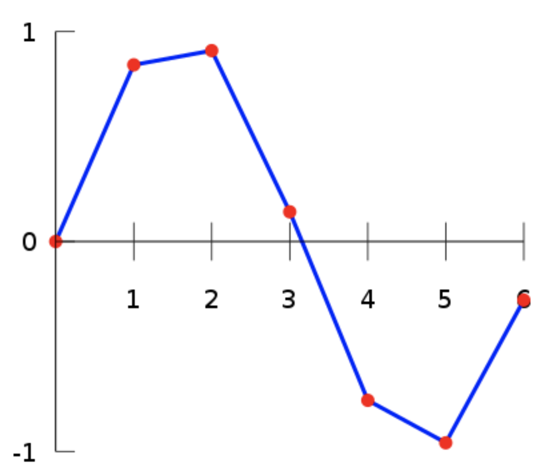
- Sometimes it is necessary to fill the "space" between samples, this is done through interpolation
- Given x_j and x_k, the linear interpolation between them can be computed using

$$x = (1 - \alpha) * x_j + \alpha * x_k$$

- With α ranging in $[0,1]$

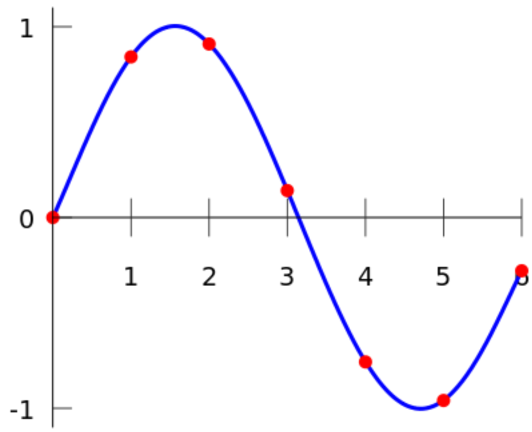
113

Linear Interpolation



114

Polynomial Interpolation

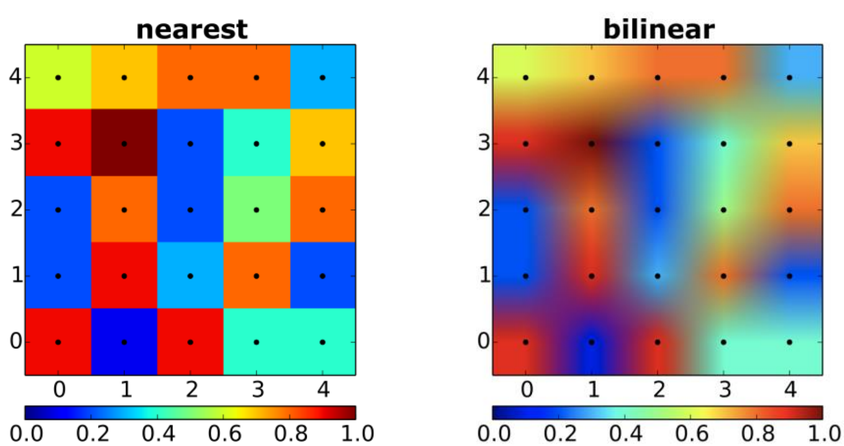


$$f(x) = -0.0001521x^6 - 0.003130x^5 + 0.07321x^4 - 0.3577x^3 + 0.2255x^2 + 0.9038x.$$

115

115

Interpolation in Higher Dimensions



116

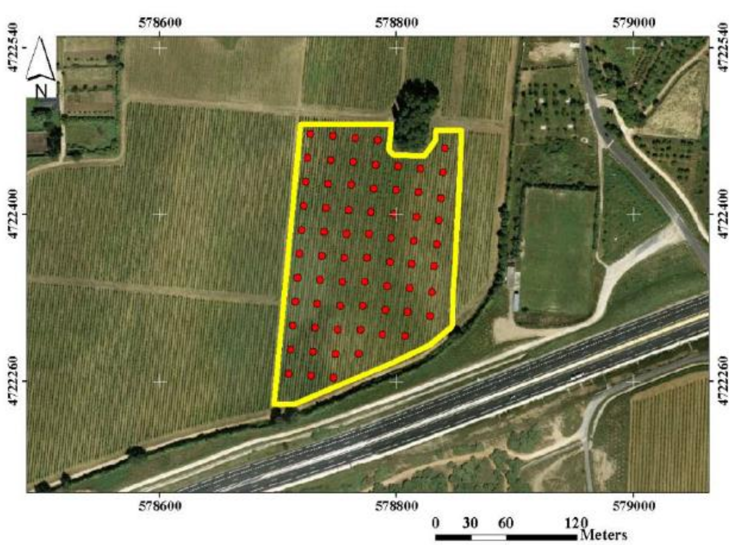
116

Sampling

- It is possible to reduce the data using a sampling strategy that preserves some data property, e.g. distribution

- It can be done by simple selecting regularly spaced data, but can result on information loss ("maps with holes")

Regular Sampling Grid



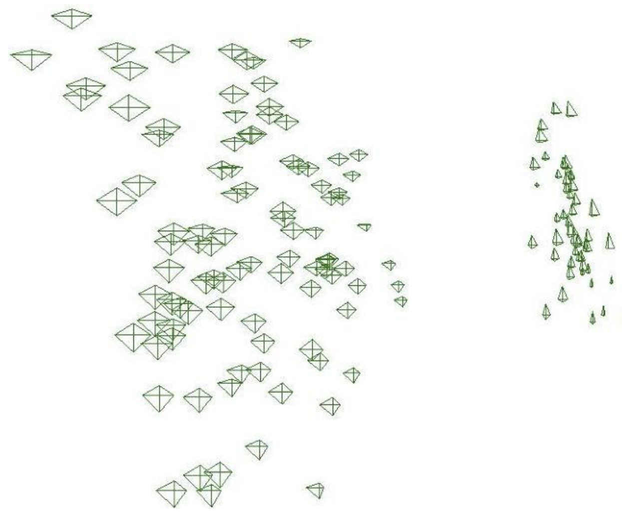
Sampling

- Another strategy involves the average on a neighborhood or a random selection on a certain region

Dimensionality Reduction

- Sometimes it is necessary to reduce the data dimensionality so we can use certain visualization techniques
- This reduction should preserve, as much as possible, the information contained on the original data
- Such reduction can be made by hand, selecting attributes, or using some established technique
  - Principal Component Analysis (PCA)
  - Multidimensional Scaling (MDS)
  - Self-organizing Maps (SOM)

Dimensionality Reduction



PCA of the Iris dataset.
The glyphs represent the 4 original variables: each line from the center is proportional to an attribute values.

Mapping Nominal Values to Numbers

- In case of ranked nominal values (e.g. air quality), there is a straightforward mapping, map each category into a consecutive integer
- In case of categorical values (car type), they can be transformed (expanded) into binary values, one column for each different category

Mapping Nominal Values to Numbers

Vehicle Name	Small/Sporty/ Compact/Large Sedan	Sports Car	SUV	Wagon	Minivan	Pickup	AWD	RWD	Retail Price	Dealer Cost	Engine Size (l)	Cyl	HP	City MPG	Hwy MPG	Weight	Wheel Base	Len	Width
Toyota 4Runner SR5 V6	0	0	1	0	0	0	0	0	27710	24801	4	6	245	18	21	4035	110	189	74
Toyota Avalon XL 4dr	1	0	0	0	0	0	0	0	26560	23693	3	6	210	21	29	3417	107	192	72
Toyota Avalon XLS 4dr	1	0	0	0	0	0	0	0	30920	27271	3	6	210	21	29	3439	107	192	72
Toyota Camry LE 4dr	1	0	0	0	0	0	0	0	19560	17558	2.4	4	157	24	33	3086	107	189	71
Toyota Camry LE V6 4dr	1	0	0	0	0	0	0	0	22775	20325	3	6	210	21	29	3296	107	189	71
Toyota Camry Solara SE 2dr	1	0	0	0	0	0	0	0	19635	17722	2.4	4	157	24	33	3175	107	193	72
Toyota Camry Solara SE V6 2dr	1	0	0	0	0	0	0	0	21965	19819	3.3	6	225	20	29	3417	107	193	72
Toyota Camry Solara SLE V6 2dr	1	0	0	0	0	0	0	0	26510	23908	3.3	6	225	20	29	3439	107	193	72
Toyota Camry XLE V6 4dr	1	0	0	0	0	0	0	0	25920	23125	3	6	210	21	29	3362	107	189	71
Toyota Celica GT-S 2dr	0	1	0	0	0	0	0	0	22570	20363	1.8	4	180	24	33	2500	102	171	68
Toyota Corolla CE 4dr	1	0	0	0	0	0	0	0	14085	13065	1.8	4	130	32	40	2502	102	178	67
Toyota Corolla LE 4dr	1	0	0	0	0	0	0	0	15295	13889	1.8	4	130	32	40	2524	102	178	67
Toyota Corolla S 4dr	1	0	0	0	0	0	0	0	15030	13650	1.8	4	130	32	40	2524	102	178	67
Toyota Echo 2dr auto	1	0	0	0	0	0	0	0	11560	10896	1.5	4	108	33	39	2085	93	163	65
Toyota Echo 2dr manual	1	0	0	0	0	0	0	0	10760	10144	1.5	4	108	35	43	2035	93	163	65
Toyota Echo 4dr	1	0	0	0	0	0	0	0	11290	10642	1.5	4	108	35	43	2055	93	163	65
Toyota Highlander V6	0	0	1	0	0	0	1	0	27930	24915	3.3	6	230	18	24	3935	107	185	72
Toyota Land Cruiser	0	0	1	0	0	0	1	0	54765	47986	4.7	8	325	13	17	5390	112	193	76
Toyota Matrix XR	0	0	0	1	0	0	0	0	16695	15156	1.8	4	130	29	36	2679	102	171	70
Toyota MR2 Spyder convertible 2dr	0	1	0	0	0	0	0	1	25130	22787	1.8	4	138	26	32	2195	97	153	67
Toyota Prius 4dr (gas/electric)	1	0	0	0	0	0	0	0	20510	18926	1.5	4	110	59	51	2890	106	175	68
Toyota RAV4	0	0	1	0	0	0	1	0	20290	18553	2.4	4	161	22	27	3119	98	167	68
Toyota Sequoia SR5	0	0	1	0	0	0	1	0	35695	31827	4.7	8	240	14	17	5270	118	204	78
Toyota Sienna CE	0	0	0	0	1	0	0	0	23495	21198	3.3	6	230	19	27	4120	119	200	77
Toyota Sienna XLE Limited	0	0	0	0	1	0	0	0	28800	25690	3.3	6	230	19	27	4165	119	200	77
Toyota Tacoma	0	0	0	0	0	1	0	1	12800	11879	2.4	4	142	22	27	2750	103	*	*
Toyota Tundra Access Cab V6 SR5	0	0	0	0	0	1	1	0	25935	23520	3.4	6	190	14	17	4435	128	*	*
Toyota Tundra Regular Cab V6	0	0	0	0	0	1	0	1	16495	14978	3.4	6	190	16	20	3925	128	*	*

Mapping Nominal Values to Numbers

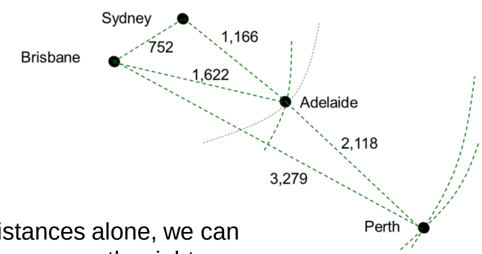
- For non-ranked values the problem is more complex (e.g. car name)
- If there is only one arbitrary nominal variable, we can use correspondence analysis
  - A numerical value can be assigned using the other variables to calculate a distance matrix, applying MDS (multidimensional scaling) to calculate unidimensional coordinates

Multidimensional Scaling

- Multidimensional scaling (MDS) is a technique for visualizing distances between objects, where the distance is known between pairs of the objects.

The distance matrix below shows the distance, in kilometers, between four Australian cities.

	Sydney	Adelaide	Brisbane
Adelaide	1,166		
Brisbane	752	1,622	
Perth	3,279	2,118	3,606

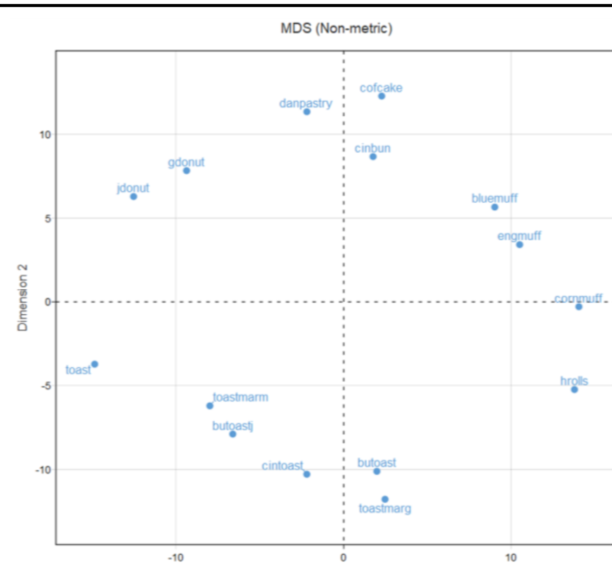


From these distances alone, we can reconstruct the map on the right.

Multidimensional Scaling (II)

- The distance matrix below shows the *perceived dissimilarities* between 15 breakfast baked goods, where a high number means that the subject rated them as being very dissimilar, and a lower number indicates the pair of baked breakfast goods are highly similar.

	toast	butoast	engmuff	jdonut	cintoast	bluemuff	hrolls	toastmarm	butoastj	toastmarg	cinbun	danpastry	gdonut	cofcake
butoast	15													
engmuff	25	15												
jdonut	3	24	22											
cintoast	14	3	17	22										
bluemuff	24	17	2	21	19									
hrolls	28	8	4	27	18	8								
toastmarm	7	7	20	11	6	18	23							
butoastj	8	6	21	12	5	19	22	2						
toastmarg	16	2	16	25	4	18	9	8	7					
cinbun	26	17	10	17	12	7	18	20	19	18				
danpastry	21	25	11	5	19	10	22	17	16	26	2			
gdonut	20	18	24	2	23	22	25	11	12	17	4	11		
cofcake	16	22	11	13	21	7	21	21	20	23	6	7	11	
cornmuff	27	11	3	26	16	4	5	25	24	12	12	16	24	16



Paul E. and Vithala R. Rao (1972), Applied Multidimensional Scaling: A Comparison of Approaches and Algorithms. New York: Holt, Rinehart and Winston.

Multidimensional Scaling (III)

- When reading an MDS map, we can consider only distances. Unlike a geographic map, there is no concept of up or down, or north and south.
- All examples represent the same.

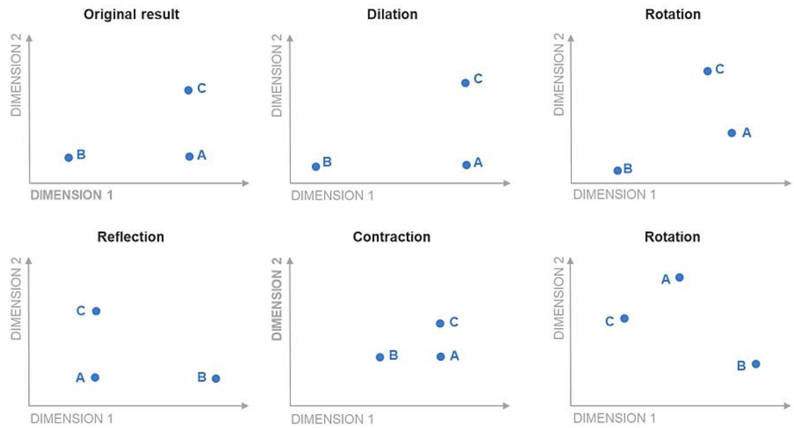


Figure from: Lehman, Donald (1989): Market Research and Analysis, 3rd Edition, Irwin.

127

127

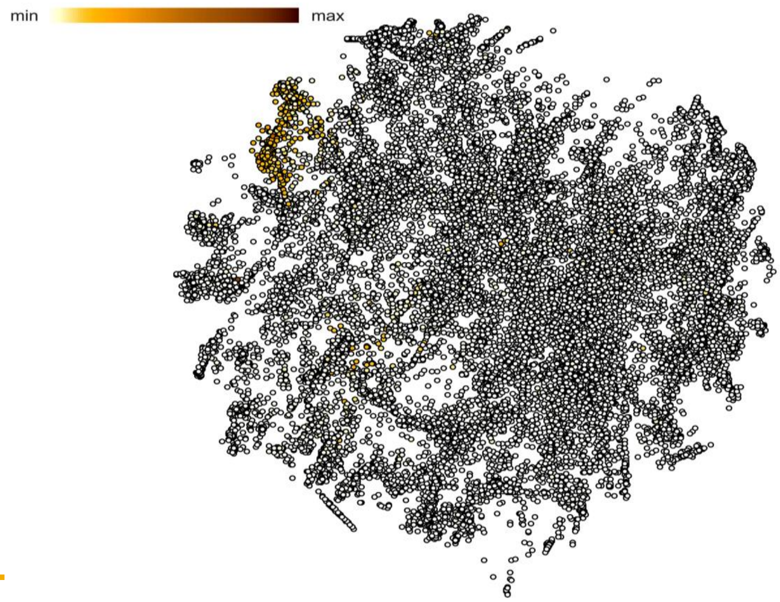
Aggregation and Summarization

- It can be useful to group data instances, using representatives for the groups (aggregation)
  - The average can be shown or some other extra information, such as, the number of instances in a group
- The core idea of aggregation is to provide information to help users to decide if a group needs to be further inspected
  - Variability analysis, outlier detection, and others are essential

128

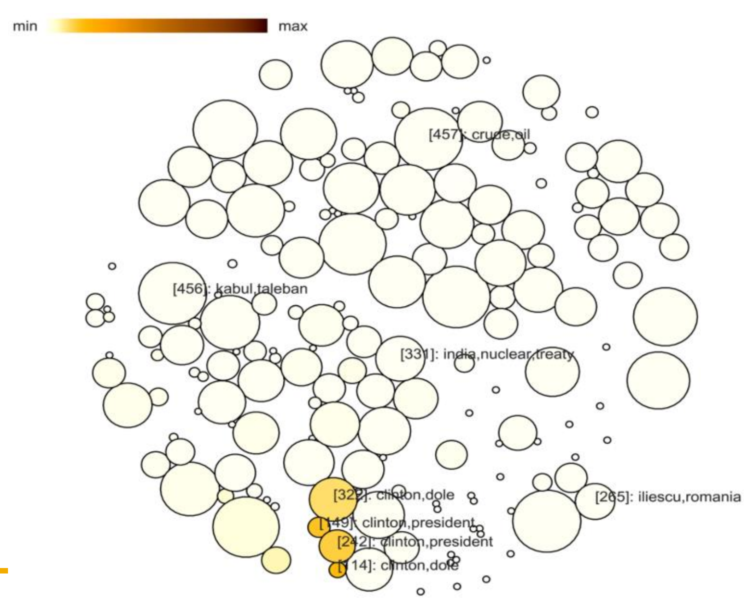
128

Aggregation and Summarization



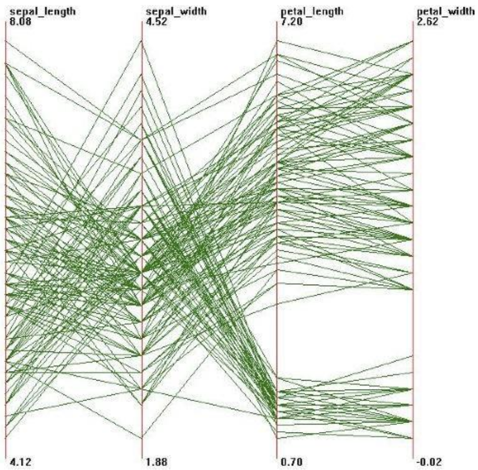
129

Aggregation and Summarization

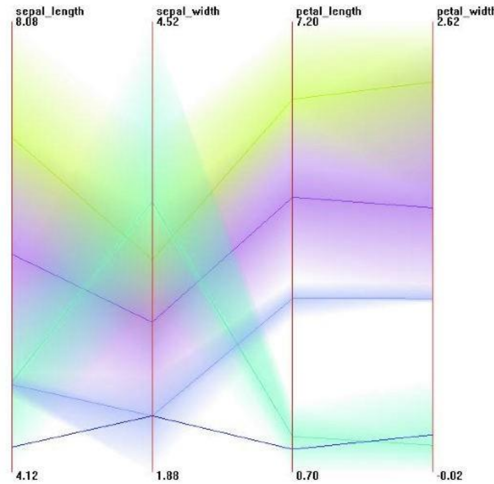


130

Aggregation and Summarization



Original



After aggregation

131

131

Final Observation

If the data were transformed through some process, this needs to be informed to the user or analyst!

132

132