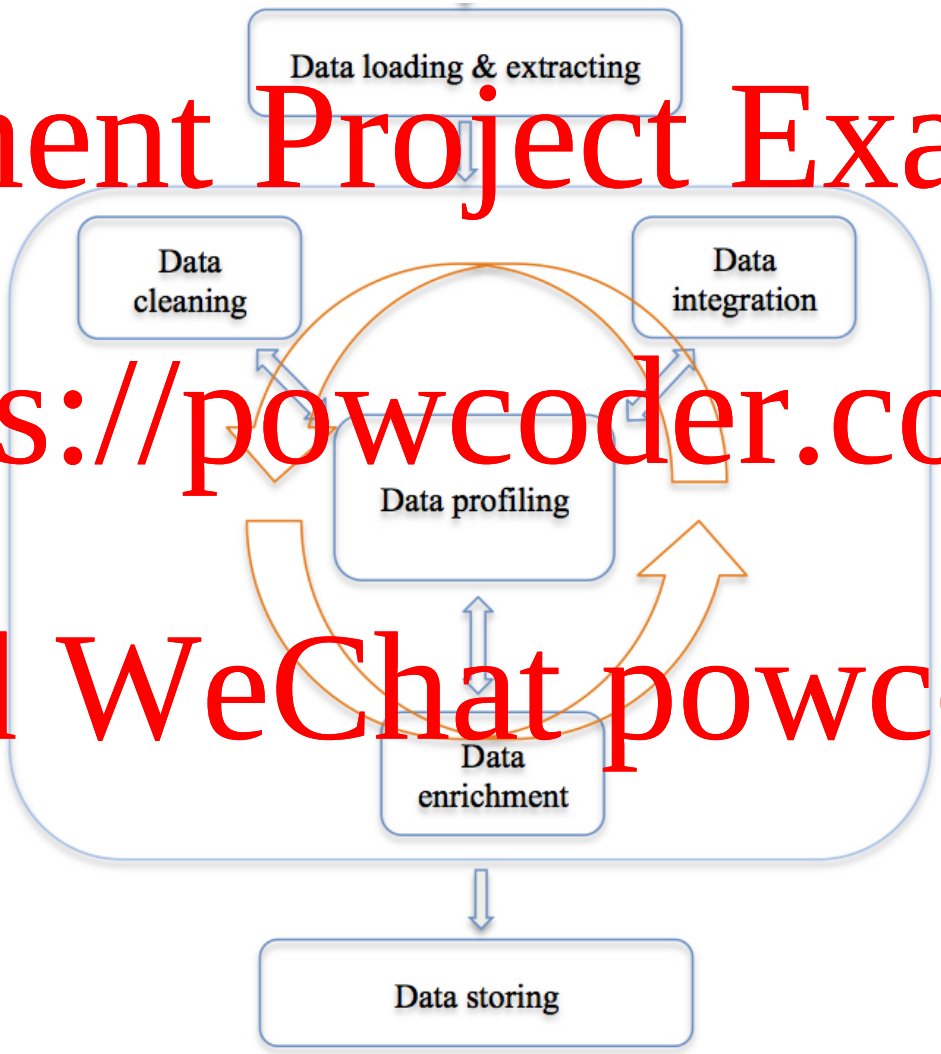
Data Cleansing — 1

Faculty of Information Technology, Monash University, Australia

FIT5196 week 6

Data Wrangling Process



Outline

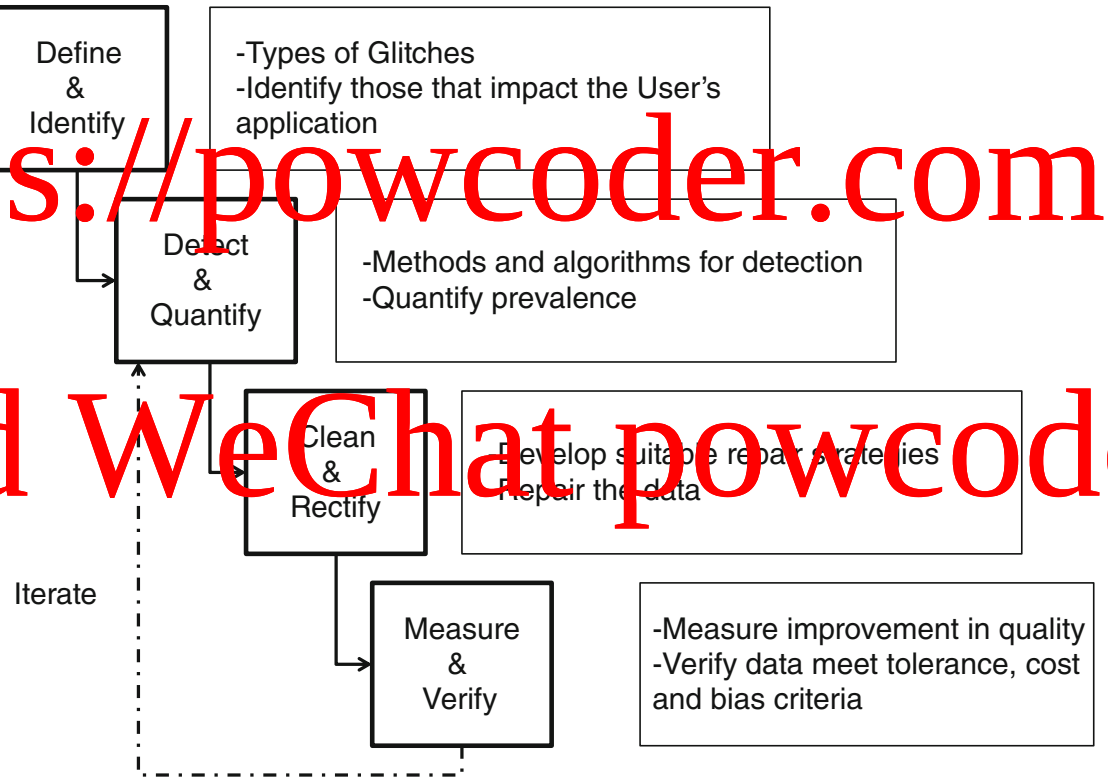
1. Data Anomalies
2. Exploratory Data Analysis
3. Summary

Data Cleansing

- Data Cleansing: A process of detecting and removing errors and inconsistencies from data in order to improve the quality of data

Data Cleansing

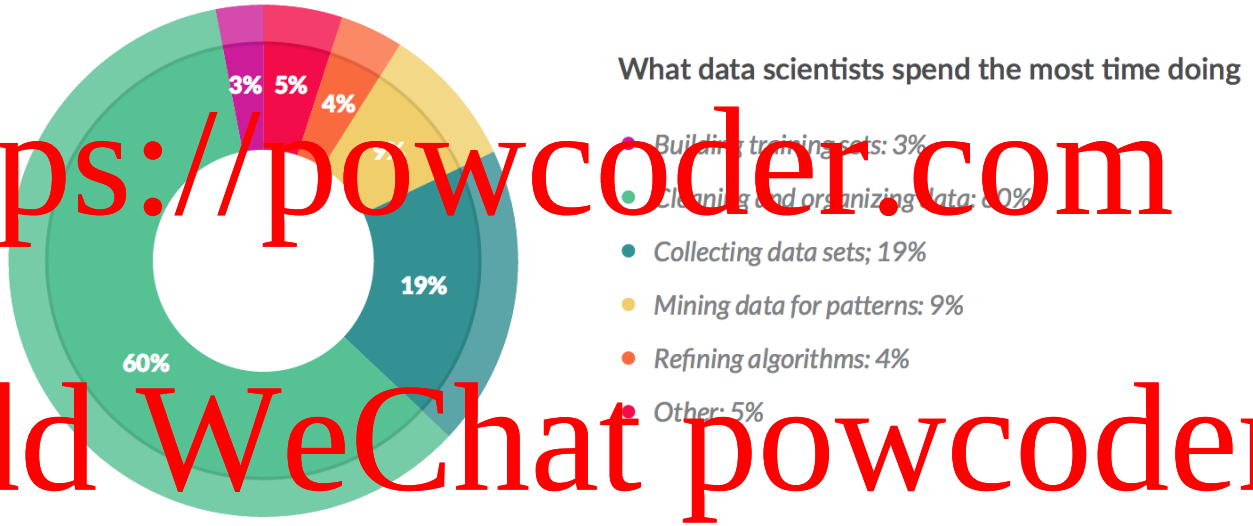
- Data Cleansing: A process of detecting and removing errors and inconsistencies from data in order to improve the quality of data
- Data Cleansing: An iterative process



An overview of data quality process, from "Data Glitches: Monsters in Your Data" by Tamraparni Dasu, in "Handbook of Data Quality" 2013

Data Cleansing

- Data Cleansing: A process of detecting and removing errors and inconsistencies from data in order to improve the quality of data
- Data Cleansing: An iterative process

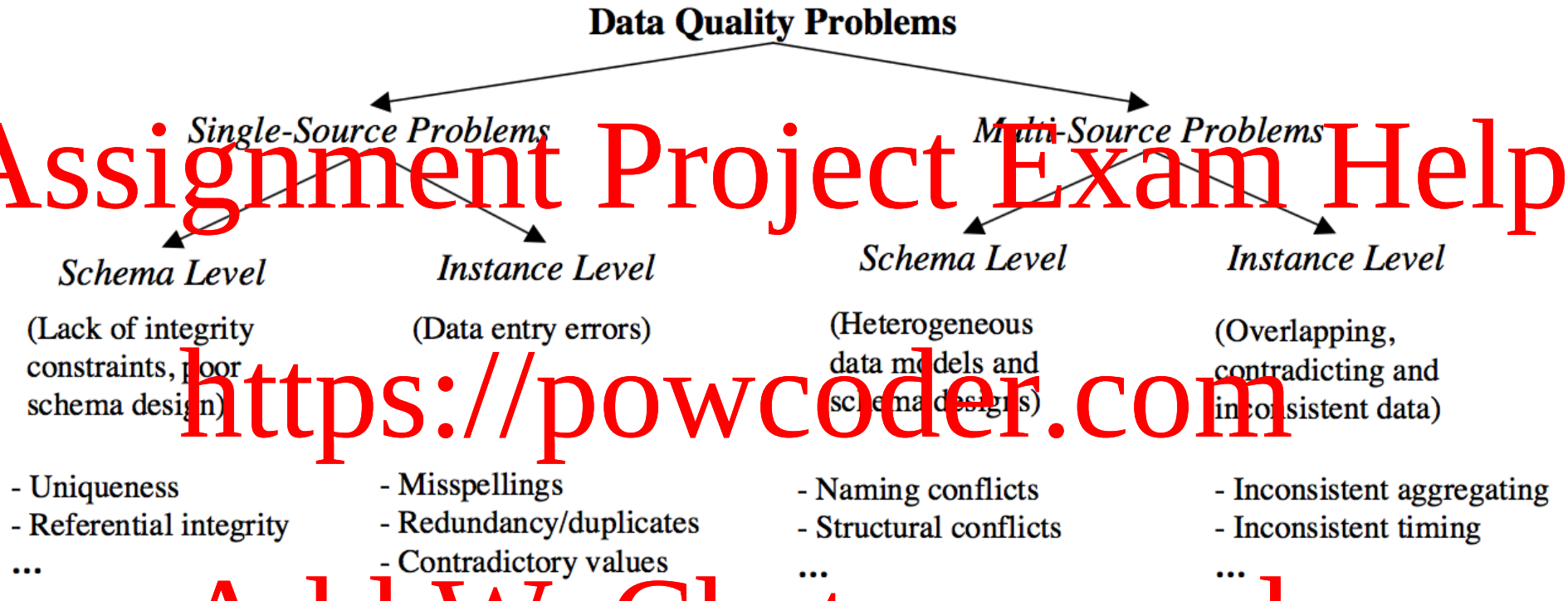


Data Anomalies (i.e., Glitches or Errors)

MONASH University

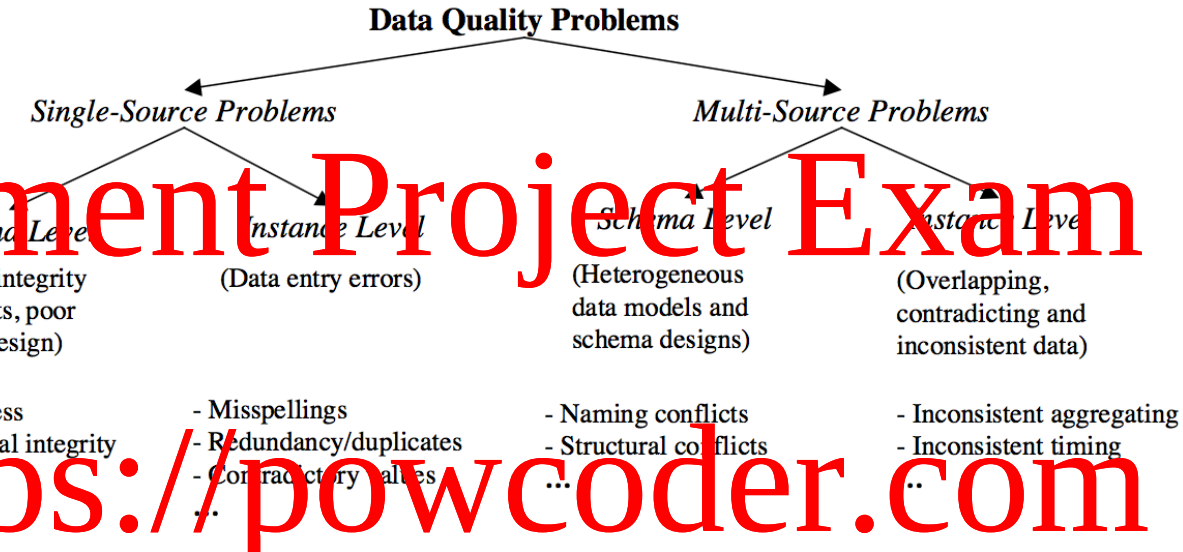
- Data Anomalies describes the distortion of the data because of any of the problems that might encounter in the life cycle of data that includes its capture, storage, update, transmission, access, archive, restore, deletion and purge.
- Some common data quality problems
  - Missing data
  - Inconsistent and faulty data
  - Outliers
  - Duplicates

Data Anomaly Classification: source-based[1]



[1] From "Data Cleaning: Problems and Current Approaches" by Rahm and Do

Data Anomaly Classification: source-based[1]

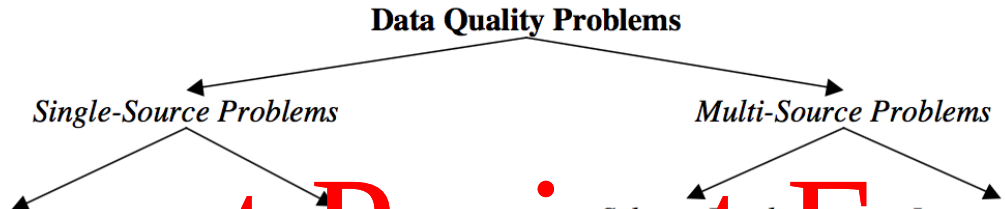


Scope/Problem		Dirty Data	Reasons/Remarks
Attribute	Illegal values	bdate=30.13.70	values outside of domain range
Record	Violated attribute dependencies	age=22, bdate=12.02.70	age = (current date – birth date) should hold
Record type	Uniqueness violation	emp=(name="John Smith", SSN="123456") emp=(name="Peter Miller", SSN="123456")	uniqueness for SSN (social security number) violated
Source	Referential integrity violation	emp=(name="John Smith", deptno=127)	referenced department (127) not defined

Table 1. Examples for single-source problems at schema level (violated integrity constraints)

[1] From "Data Cleaning: Problems and Current Approaches" by Rahm and Do

Data Anomaly Classification: source-based[1]

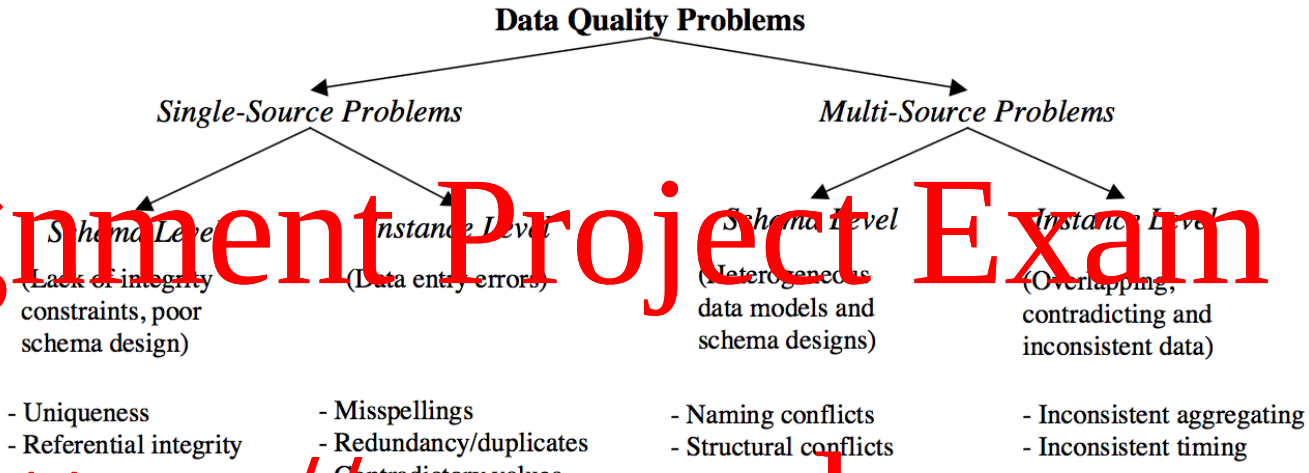


Scope/Problem		Dirty Data	Reasons/Remarks
Attribute	Missing values	phone=9999-999999	unavailable values during data entry (dummy values or null)
	Misspellings	city="Liipzig"	usually typos, phonetic errors
	Cryptic values, Abbreviations	experience="B"; occupation="DB Prog."	
	Embedded values	name="J. Smith 12.02.70 New York"	multiple values entered in one attribute (e.g. in a free-form field)
	Misfielded values	city="Germany"	
Record	Violated attribute dependencies	city="Redmond", zip=77777	city and zip code should correspond
Record type	Word transpositions	$name_1$= "J. Smith", $name_2$="Miller P."	usually in a free-form field
	Duplicated records	emp_1=(name="John Smith",...); emp_2=(name="J. Smith",...)	same employee represented twice due to some data entry errors
	Contradicting records	emp_1=(name="John Smith", bdate=12.02.70); emp_2=(name="John Smith", bdate=12.12.70)	the same real world entity is described by different values
Source	Wrong references	emp=(name="John Smith", deptno=17)	referenced department (17) is defined but wrong

Table 2. Examples for single-source problems at instance level

[1] From "Data Cleaning: Problems and Current Approaches" by Rahm and Do

Data Anomaly Classification: source-based[1]



Customer (source 1)

CID	*Name*	*Street*	*City*	*Sex*
11	Kristen Smith	2 Hurley Pl	South Fork, MN 48503	0
24	Christian Smith	Hurley St 2	S Fork MN	1

Client (source 2)

Cno	*LastName*	*FirstName*	*Gender*	*Address*	*Phone/Fax*
24	Smith	Christoph	M	23 Harley St, Chicago IL, 60633-2394	333-222-6542 / 333-222-6599
493	Smith	Kris L.	F	2 Hurley Place, South Fork MN, 48503-5998	444-555-6666

Customers (integrated target with cleaned data)

No	*LName*	*FName*	*Gender*	*Street*	*City*	*State*	*ZIP*	*Phone*	*Fax*	*CID*	*Cno*
1	Smith	Kristen L.	F	2 Hurley Place	South Fork	MN	48503-5998	444-555-6666		11	493
2	Smith	Christian	M	2 Hurley Place	South Fork	MN	48503-5998			24	
3	Smith	Christoph	M	23 Harley Street	Chicago	IL	60633-2394	333-222-6542	333-222-6599		24

Figure 3. Examples of multi-source problems at schema and instance level

[1] From "Data Cleaning: Problems and Current Approaches" by Rahm and Do

Data Anomalies Classification: type-based

- Syntactical Anomalies: format and values
- Semantic Anomalies: comprehensiveness and non-redundancy
- Coverage Anomalies: missing values

Data Anomalies Classification: type-based

- Syntactical Anomalies: format and values
  - Lexical errors: data format discrepancies in terms of database; spelling errors, typos in terms of linguistics.
  - Domain format errors: inconsistent value format of an attribute, e.g., Buntine, Wray Lindsay v.s. Wray L. Buntine
  - Irregularities: the non-uniform use of values, units and abbreviations?e.g., salary in difference currencies.
- Semantic Anomalies: comprehensiveness and non-redundancy
- Coverage Anomalies: missing values

Data Anomalies Classification: type-based

- Syntactical Anomalies: format and values
- Semantic Anomalies: comprehensiveness and non-redundancy
  - Integrity constraint violations
  - Contradictions: violation of dependencies between attributes, e.g., AGE and DOB.
  - Duplicates: observations representing the same entity.
  - Invalid observations
- Coverage Anomalies: missing values

Data Anomalies Classification: type-based

- Syntactical Anomalies: format and values
- Semantic Anomalies: comprehensiveness and non-redundancy
- Coverage Anomalies: missing values
  - Missing values: due to omissions while collecting the data
  - Missing observations:

Taxonomy of Dirty Data[2]

Dirty data manifests itself in three different ways:

- missing data
- not missing but wrong data,
- not missing and not wrong but unusable,

[2] Kim et al, "A Taxonomy of Dirty Data", DMKD 2003

Taxonomy of Dirty Data[2]

Dirty data manifests itself in three different ways:

- missing data
  - Missing data where there is no Null-not-allowed constraint
  - Missing data where Null-not-allowed constraint should be enforced
- not missing but wrong data,
- not missing and not wrong but unusable.

[2] Kim et al, "A Taxonomy of Dirty Data", DMKD 2003

Taxonomy of Dirty Data[2]

Dirty data manifests itself in three different ways:

- missing data
- not missing but wrong data, due to
  - Integrity constraints
    - violation of data type constraint, including value range
    - violation of non-null uniqueness constraint, i.e., duplicated data
    - violation of referential integrity
    - Wrong categorical data
    - Outdated temporal data
    - Inconsistent spatial data
  - Data Entry error involving a single table
    - Data entry error involving a single field: erroneous entry, misspelling, extraneous data
    - Data entry error involving multiple fields: entry into wrong fields, wrong derived-field data
- not missing and not wrong but unusable,

[2]Kim et al, "A Taxonomy of Dirty Data", DMKD 2003

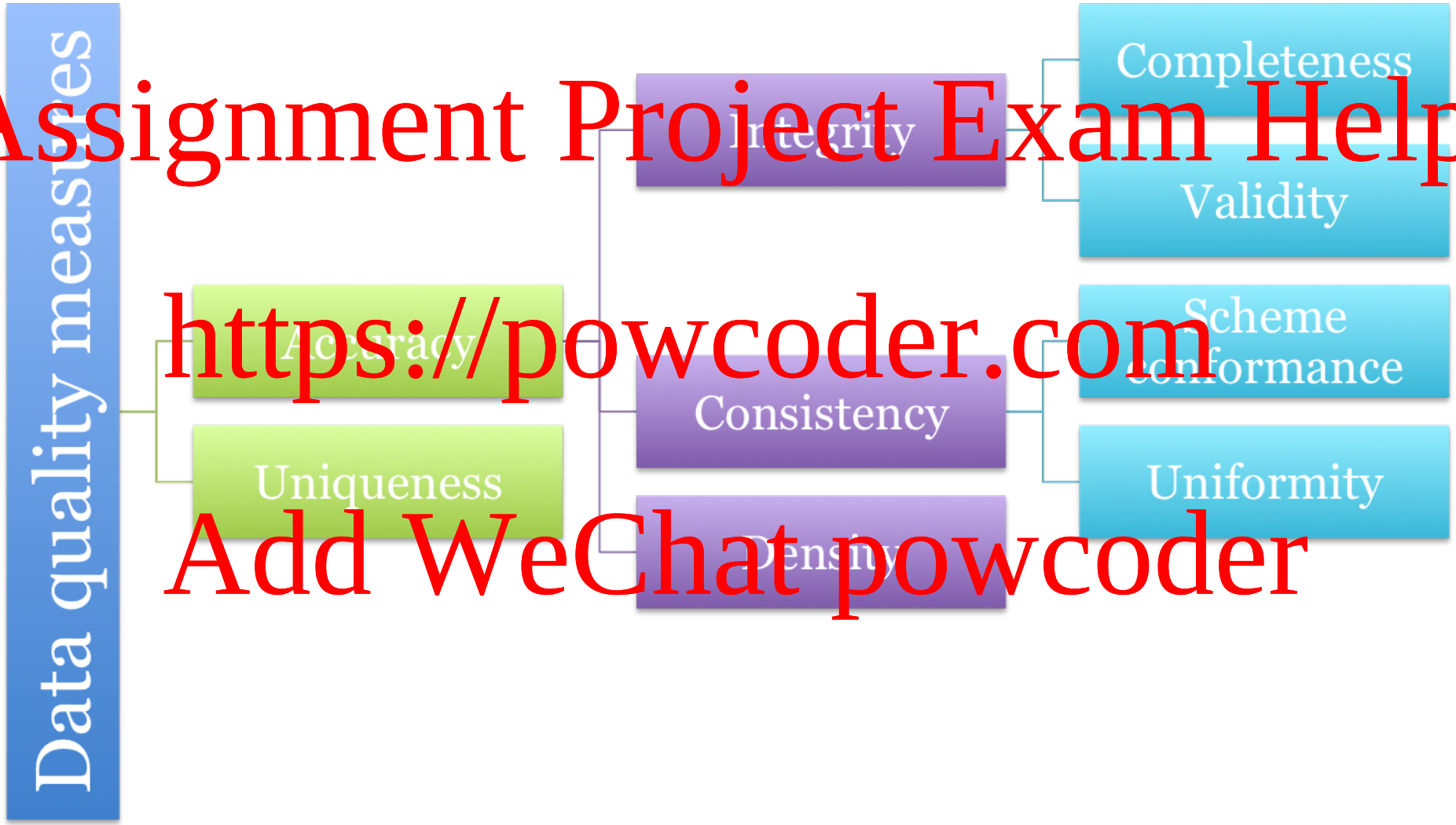
Taxonomy of Dirty Data[2]

Dirty data manifests itself in three different ways:

- missing data
- not missing but wrong data
- not missing and not wrong but unusable, due to
  - Different data for the same entity across multiple databases
  - Ambiguous data due to: the use of abbreviation (Dr. for doctor or drive),
  - Incomplete context (e.g. Sydney of Australia or Canada)
  - The use of abbreviation (e.g., ste for suite, rd for road, st for street, etc)
  - Alias/nick name (e.g., Bill Clinton, President Clinton)
  - Encoding formats (e.g, ASCII, ...)
  - Representations (e.g. negative number, precision, fraction)
  - Measurement units (e.g., data, time, currency, weight, area, etc.)
  - Uses of special characters (e.g., space, dash, parenthesis in phone numbers) in concatenated data

[2]Kim et al, "A Taxonomy of Dirty Data", DMKD 2003

Quality measures



Adapted from "Problems, Methods, and Challenges in Comprehensive Data Cleansing" By Muller and Fretag

Data Anomalies: Looking for Errors

MONASH University

ID	Landgrabbed	ISO	Landgrabber	Base	Sector	Hectares	Production	Projected investment	Status of deal	Start	End
A23	Algeria	DZA	Al Qudra	UAE	Finance	31000.00	Milk, olive oil, potatoes		Done	06/2005	11/2015
A5	Algeria		Al Qudra	UAE	real estate	31000.00	Milk, olive oil, potatoes		Done	07/1905	11/2012
A3	Angola		CAMC Engineering Co. Ltd	China	Construction	1[illegible]00	Rice	US$77 million	Don	06/2010	05/2005
A1	Philippines		Kuwait	Kuwait	Government	20000	Maize, rice		In process	10/2015	12/1917
A34	菲律宾		Zuellig Group	Malaysia	Agribusiness, health care	30000	Maize		In process	06/2016	08/2020
A45	Philippines		Oman	Oman	Government	10,000	Rice	150m	Processing	06/1909	09/1917
A34	Philippines	PHL	Brunei Investment Authority	Brunei	Gover	10,000	Rice		Proposed	03/2016	
A56	Philippines		China	China		100,[illegible]000	various		Suspended	02/2000	11/2001
A54	Philippines		Green Future Innovation	Japan		11,000	Sugar cane	US$120 million	Done	06/2014	09/2015
A4	argantin	ARG	Beidahuang	CH		320000	Maize, soybeans, wheat	US$1,500 million	Suspended	12/1900	07/1901
A65	Tanzânia		Nirmal Seeds	Indië	Agribusiness	30000	Seeds		In process	03/2013	06/2016
	tanzania		Yes Ban	India	Finance	50000	Rice, wheat		In process	06/2010	06/2017
A3	Tansania		Export Trading Group	Singapore	Agribusiness	80[illegible]0	Rice		Don	12/2015	10/2018
A23	Brazil	BRA	Clean Energy Brazil	UK		30,000	Sugar cane		Done	03/2012	09/2013
N67		BRA	Adecoagro	US	Agribusiness	165,000	Cattle, cofee, grains, soybeans, sugar cane	98,000,000	Done	10/2010	07/2005
A67	brazil	BRA	Archer Daniels Midland	US	Agribusiness	12,000	Oil palm		In process	06/2014	01/2015
A56	Brasil		Black River Asset Management	United states	Finance	50,000	Crops	20000000	Done	02/2010	2015

Data Anomalies: Some Errors

ID	Landgrabbed	ISO	Landgrabber	Base	Sector	Hectares	Production	Projected investment	Status of deal	Start	End
A23	Algeria	DZA	Al Qudra	UAE	Finance	31000.00	Milk, olive oil, potatoes		Done	06/2005	01/2015
A45	Algeria		Al Qudra	UAE	real estate	31000.00	Milk, olive oil, potatoes		Done	06/1905	01/2012
A1	Angola		CAMC Engineering Co. Ltd	China	Construction	50[illegible].00	Rice	US$77 million	Done	06/2010	01/2015
A1	Philippines		Kuwait	Kuwait	Government	20000	Maize, rice		In process	10/2015	12/1917
A34	菲律宾		Zuellig Group	Malaysia	Agribusiness, health care	30000	Maize		In process	06/2016	
A45	Philippines		Oman	Oman	Government	10,000	Rice	150m	Processing	06/1909	09/1917
A34	Philippines	PHL	Brunei Investment Authority	Brunei	Gover	10,000	Rice		Proposed	03/2016	
A56	Philippines		China	China		00,2[illegible]0,00	various		Suspended	02/2000	11/2001
A54	Philippines		Green Future Innovation	Japan		11,000	Sugar cane	US$120 million	Done	06/2014	09/2015
A4	argantin	ARG	Beidahuang	CH		320000	Maize, soybeans, wheat	US$1,500 million	Suspended	12/1900	07/1901
A65	Tanzânia		Nirmal Seeds	Indië	Agribusiness	30000	Seeds		In process	03/2013	06/2016
	tanzania		Yes Bank	India	Finance	50000	Rice, wheat		In process	06/2010	
A3	Tansania		Export Trading Grou	Singapore	Agribusiness	00	Rice		Done	12/2015	10/2018
A23	Brazil	BRA	Clean Energy Brazil	UK		30,000	Sugar cane		Done	03/2012	09/2013
N67		BRA	Adecoagro	US	Agribusiness	165,000	Cattle, cofee, grains, soybeans, sugar cane	98,000,000	Done	10/2010	07/2005
A67	brazil								In process	06/2014	
A56	Brasil		Black River Asset Management	United states	Finance	50,000	Crops	20000000	Done	02/2010	2015

Data Cleansing Blueprint

1. Data Auditing (or Analysis): detect errors and inconsistencies in the data
   - Data profiling: focuses on the instance analysis of individual attributes
   - Data mining: descriptive data mining, e.g. clustering, summarisation, association discovery, etc.
2. Definition of transformation workflow: define a sequence of operations on the data, used to detect and eliminate anomalies
   - Early data cleaning steps: correct single-source instance problems
   - Later data cleaning steps: deal with schema/data integration and clean multi-source problems.
3. Verification: test and evaluate the correctness and effectiveness of a transformation workflow
4. Data transformation: Execute the transformation steps
5. Post-processing and controlling: inspect the results to verify the correctness of the specified operations.

Outline

1 Data Anomalies

2 Exploratory Data Analysis

3 Summary

Exploratory Data Analysis

- Two types of variables:
  - categorial variable
  - numerical variable
- Two types of EDA
  - Non-graphical: summary statistics
  - Graphical: various plots
- EDA
  - Univariate
  - Multivariate

Univariate non-graphical methods: Categorial data

- Categorical variables: values or observations that can be sorted into groups or categories.
  - Examples: Sex, Eye colour and blood type
- The characteristics of interest for a categorial variable
  - the range of values
  - the frequency of occurrence for each value
  - univariate non-graphical technique: tabulation of the frequencies

	sex	embarked	class	who	deck	embark_town	alive	name
count	892	890	892	892	204	890	892	892
unique	4	3	3	3	7	3	2	89
top	male	S	Third	man	C	Southampton	no	Behr, Mr. Karl Howell
freq	574	644	491	538	60	643	550	2

Univariate non-graphical methods: Quantitative data

- Numerical variables: values or observations that can be measured, and these numerical values can be placed in ascending or descending order.
  - Examples: salary, height, weight, etc.
- The characteristics of the population distribution of a numerical variable
  - center tendency: "location" of a distribution, dealing with typical or middle values.
  - spread: an indicator of how far away from the centre we are still likely to find data values.
  - shape: Skewness and Kurtosis
  - outliers: values that are outside of the areas of a distribution that would commonly occur.

Univariate non-graphical methods: Quantitative data

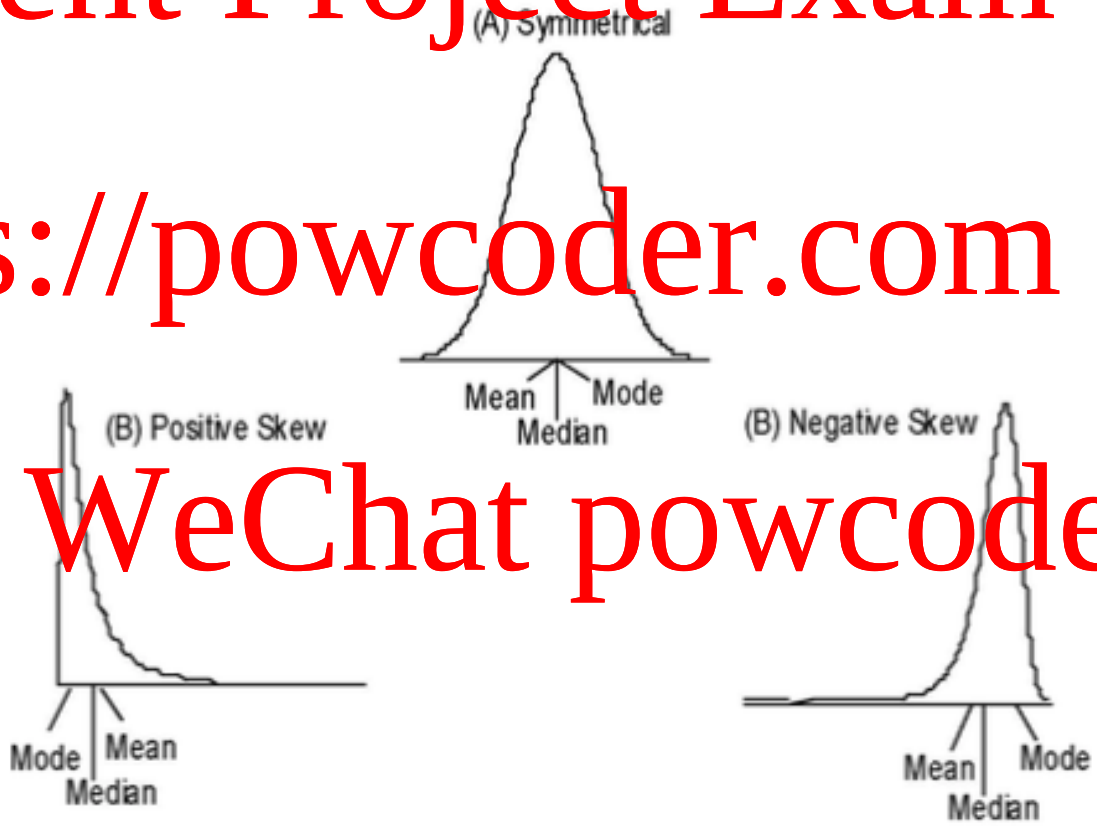
- The characteristics of the population distribution of a numerical variable
  - center tendency
    - Mean: the arithmetic average of a set of values

$$\mu = \frac{x_1 + x_2 + \cdots + x_n}{n}$$

    - Median: the middle value after all the values are put in an order list.
    - Mode: the most frequent occurring value in a set of values

Univariate non-graphical methods: Quantitative data

- The characteristics of the population distribution of a numerical variable
  - center tendency

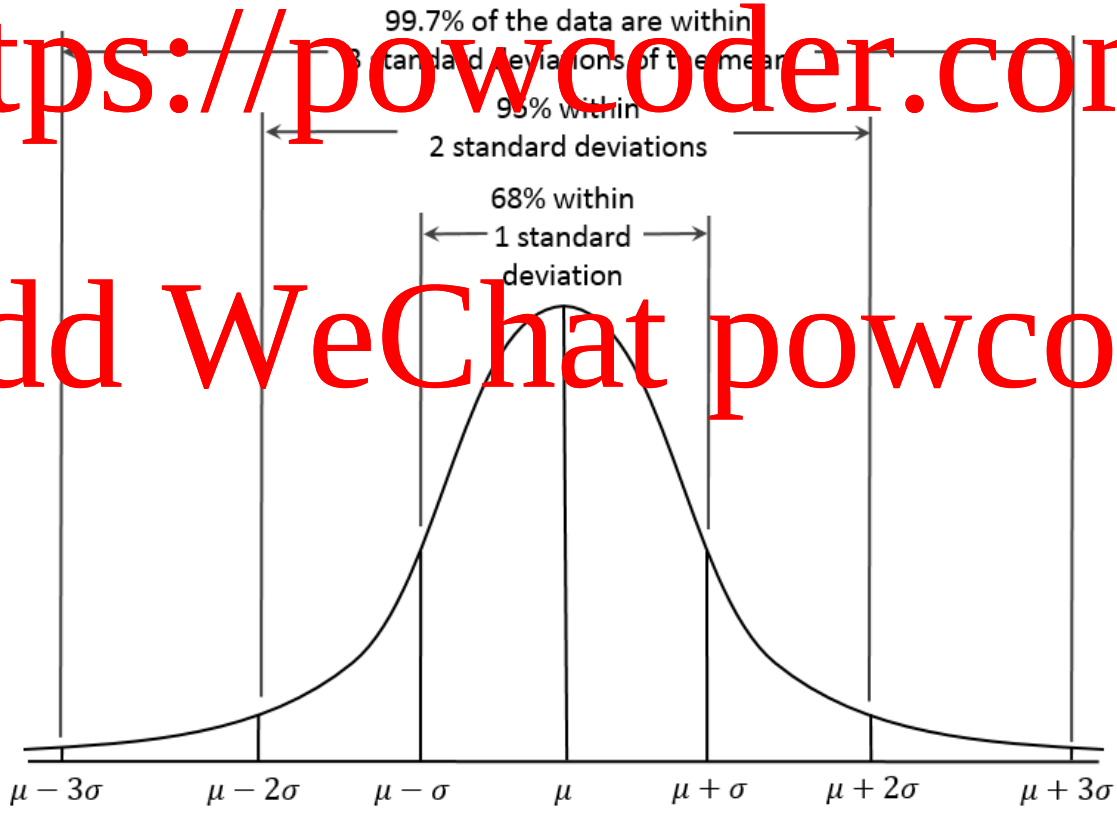


This figure is from "Summary Statistics"

Univariate non-graphical methods: Quantitative data

- The characteristics of the population distribution of a numerical variable
  - spread
    - Range: the difference between the smallest and largest values in the data set
    - Standard Deviation and Variance

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



This figure is from Wikipedia

Univariate non-graphical methods: Quantitative data

- The characteristics of the population distribution of a numerical variable
  - spread
    - Range: the difference between the smallest and largest values in the data set.
    - Standard Deviation and Variance

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

    - The Interquartile Range:

Q0	the minimum
Q1	bigger than 25% of the data points
Q2	the median
Q3	bigger than 75% of the data points
Q4	the maximum

    - The inter-quartile range (IQR):

$$IQR = Q3 - Q1$$

Univariate non-graphical methods: Quantitative data

- The characteristics of the population distribution of a numerical variable
  - The output of Pandas describe() function

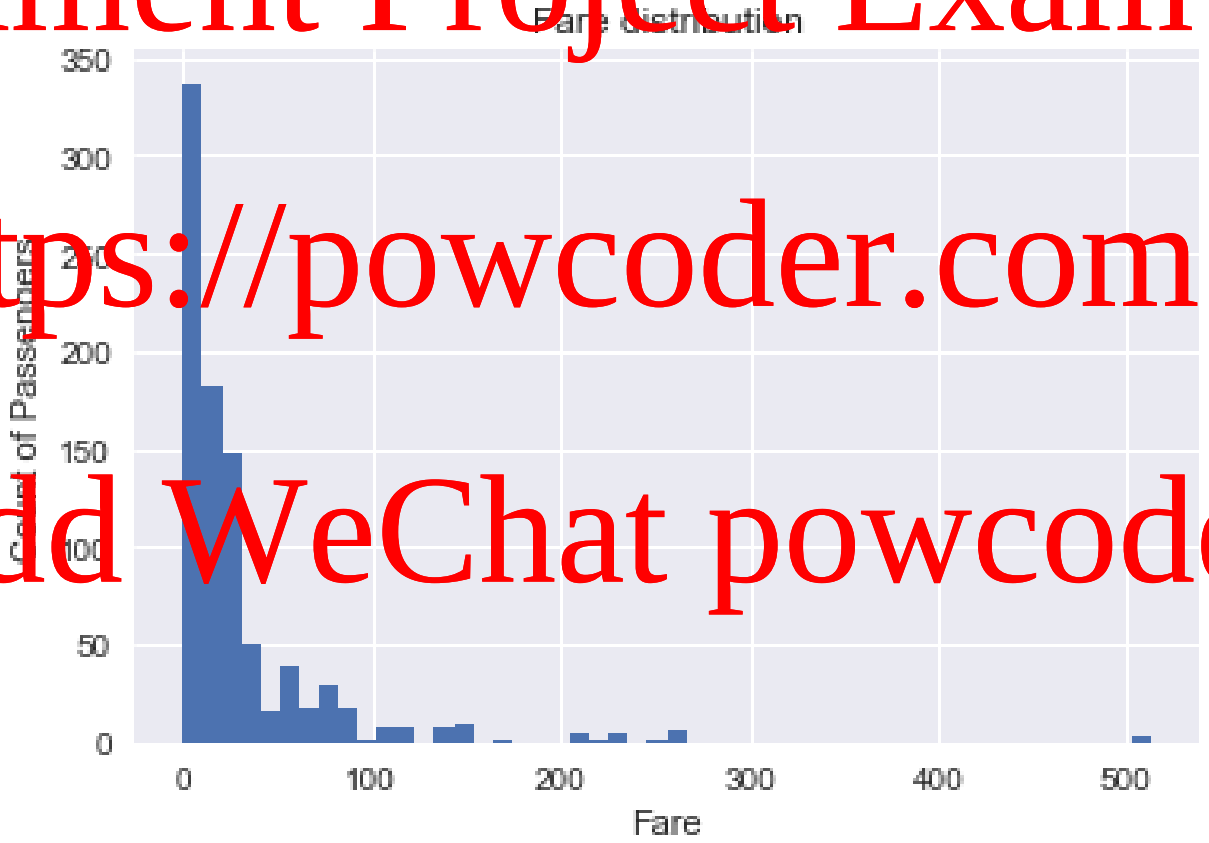
	survived	pclass	age	sibsp	parch	fare
count	892.000000	892.000000	715.000000	892.000000	892.000000	892.000000
mean	0.384529	2.307175	29.720517	0.522422	0.381166	32.201737
std	0.486757	0.836750	14.490914	1.102264	0.806706	49.665589
min	0.000000	1.000000	0.420000	0.000000	0.000000	0.000000
25%	0.000000	2.000000	20.750000	0.000000	0.000000	7.917700
50%	0.000000	3.000000	28.000000	0.000000	0.000000	14.454200
75%	1.000000	3.000000	38.000000	1.000000	0.000000	31.000000
max	1.000000	3.000000	80.000000	8.000000	6.000000	512.329200

Univariate graphical methods

- Histograms: a quick way of learning the characteristics of your data, including central tendency, spread, shape, outliers, etc.
- Boxplots (or Box-and-Whiskers Plot): display five-point summaries and potential outliers in graphical form

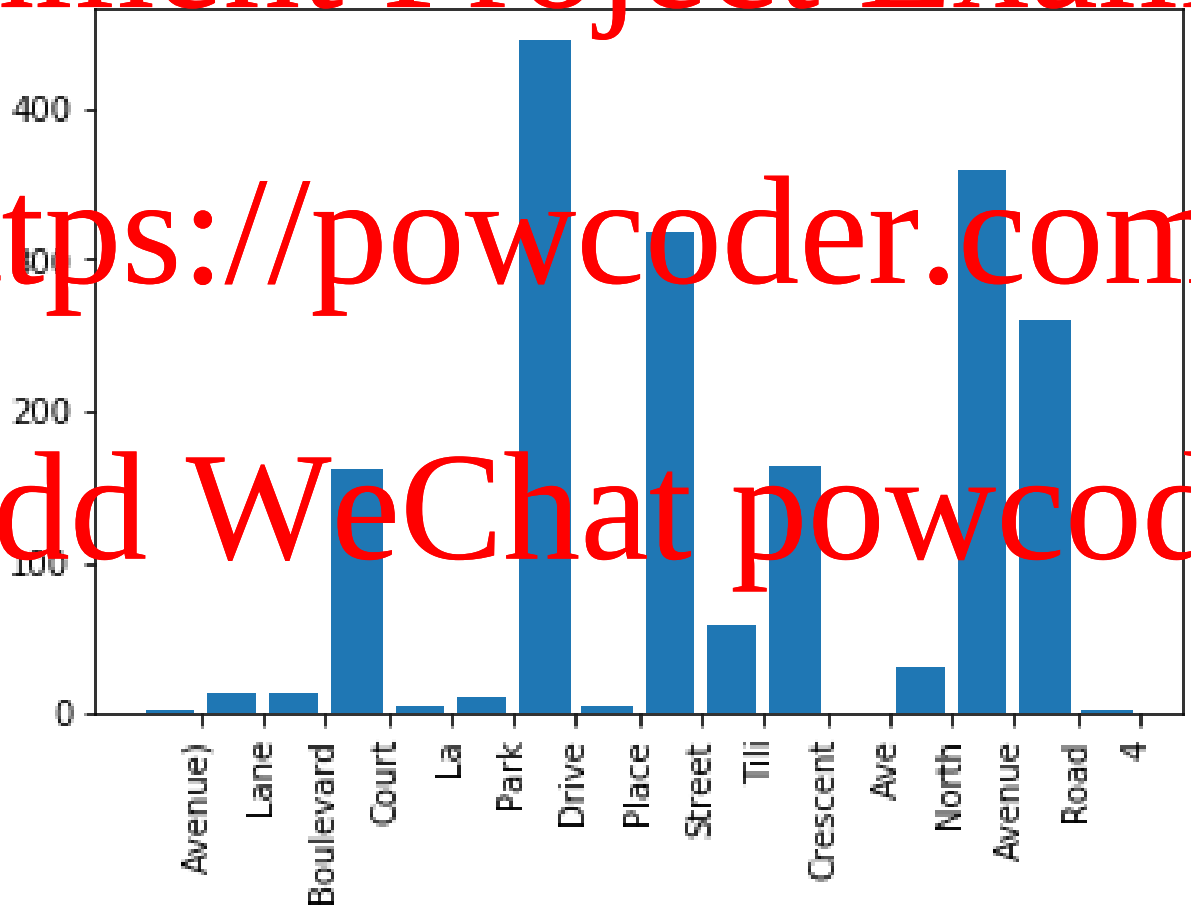
Univariate graphical methods

- Histograms: a quick way of learning the characteristics of your data, including central tendency, spread, shape, outliers, etc.



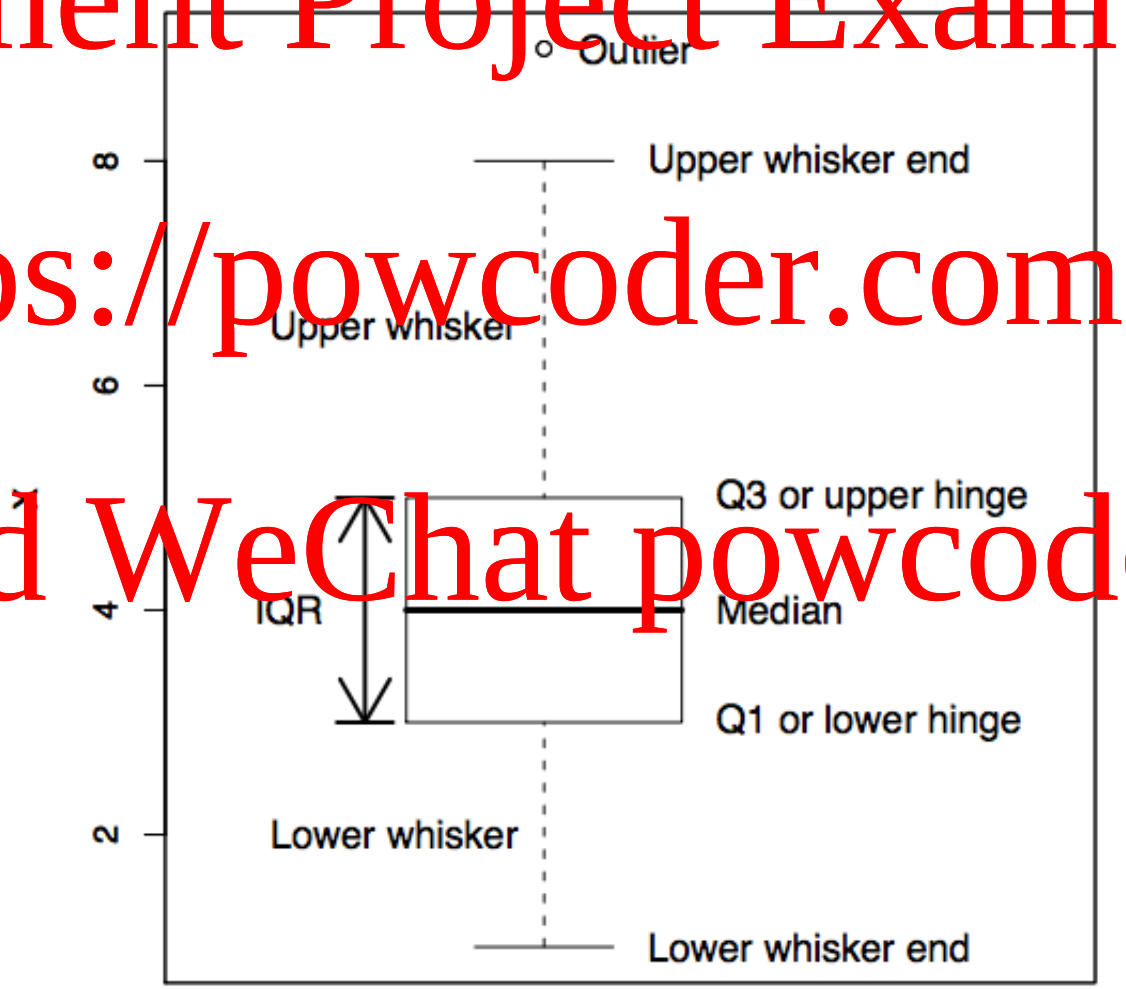
Univariate graphical methods

- Histograms: a quick way of learning the characteristics of your data, including central tendency, spread, shape, outliers, etc.



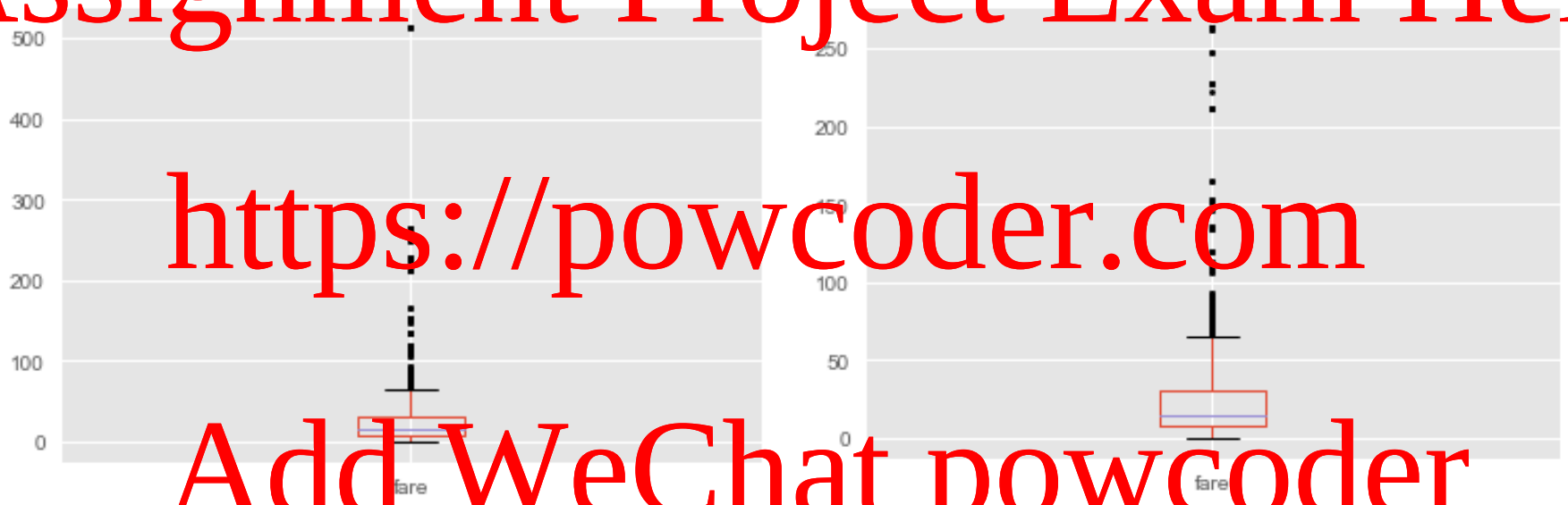
Univariate graphical methods

- Boxplots (or Box-and-Whiskers Plot): display five-point summaries and potential outliers in graphical form



Univariate graphical methods

- Boxplots (or Box-and-Whiskers Plot): display five-point summaries and potential outliers in graphical form



Multivariable non-graphical methods: Categorical data

- Cross-tabulation: a two-way table with column headings that match the levels of one variable and row headings that match the levels of the other variable, then filling in the counts of all subjects that share a pair of levels

sex	female	male
child	54	58
man	0	520
woman	260	0

Multivariable non-graphical methods: Quantitative variables

- Covariance: measures how much two variables "co-vary", i.e., how much (and in what direction) should we expect one variable to change when the other changes

$$Cov(X, Y) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n-1}$$

Note $Cov(X, X) = Var(X)$

- Correlation:

$$Cor(X, Y) = \frac{Cov(X, Y)}{\sigma_x \sigma_y}$$

  - value range between −1 and +1,
  - -1 being a perfect negative linear correlation,
  - +1 being a perfect positive linear correlation,
  - and 0 indicating that X and Y are uncorrelated.

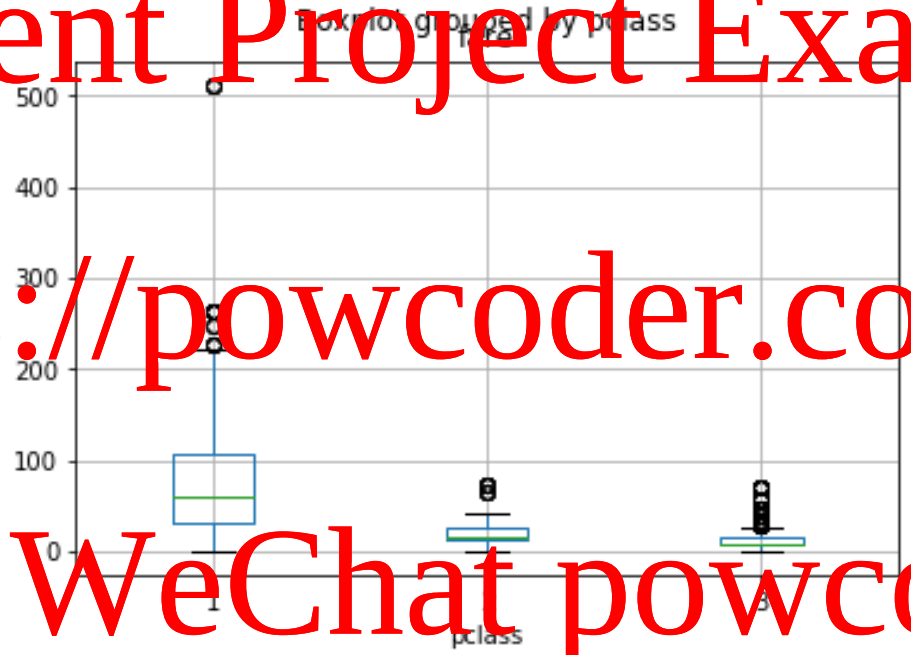
Multivariable non-graphical methods: correlation matrix

MONASH University

	price	bedrooms	bathrooms	sqft_living	sqft_lot
price	1	0.323447334198355	0.525048875929692	0.695853418039519	0.093951220218312
bedrooms	0.323447334198355	1	0.521726344148986	0.590938504384772	0.03469110139448
bathrooms	0.525048875929692	0.521726344148986	1	0.750759932730059	0.0891087018348975
sqft_living	0.695853418039519	0.590938504384772	0.750759932730059	1	0.186531766823882
sqft_lot	0.093951220218312	0.03469110139448 15	0.0891087018348975	0.186531766823882	1

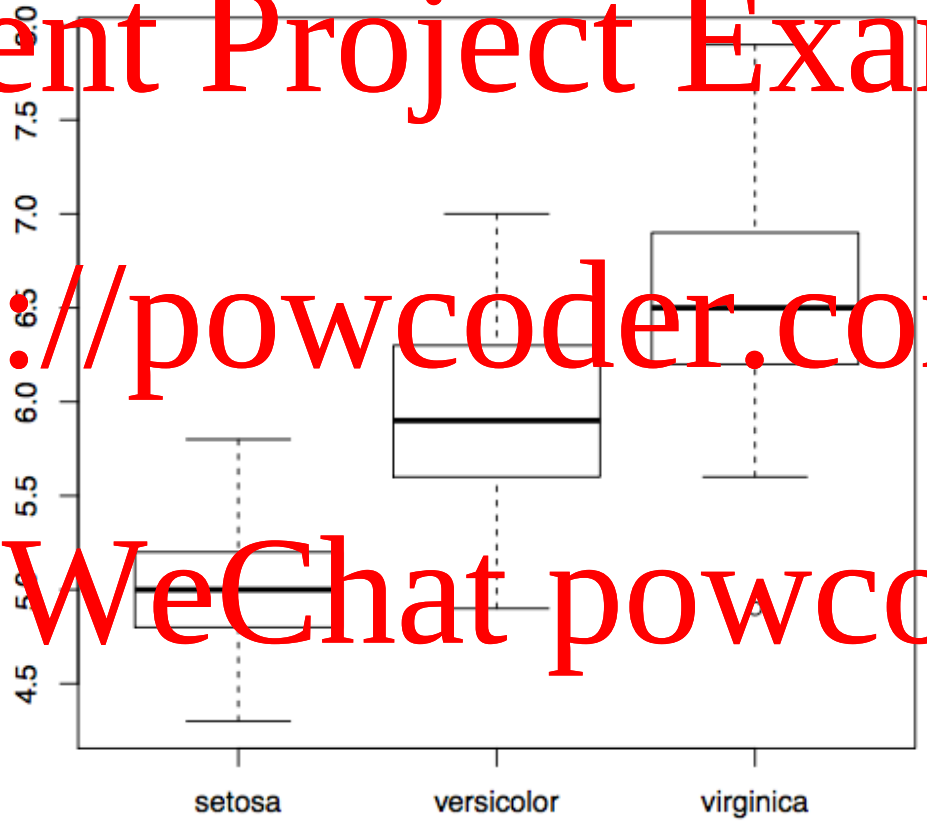
Multivariable graphical methods

- Side-by-Side boxplot



Multivariable graphical methods

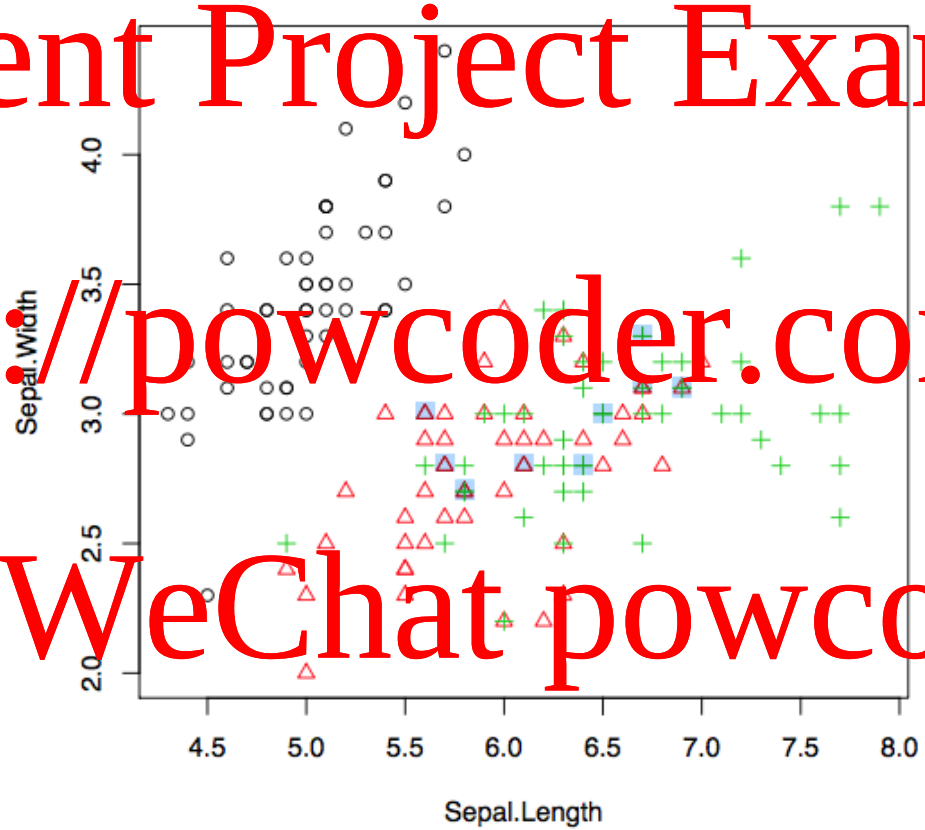
- Side-by-Side boxplot



This figure is from "R and Data Mining: Examples and Case Studies"

Multivariable graphical methods

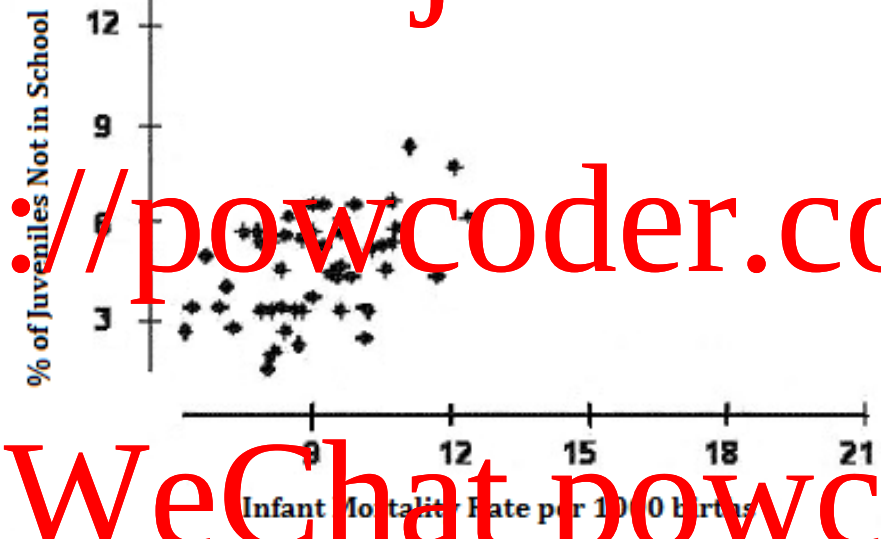
- Scatterplot



This figure is from "R and Data Mining: Examples and Case Studies"

Multivariable graphical methods

- Scatterplot



This figure is from https://onlinecourses.science.psu.edu/stat100/node/36

Summary

1. What we discussed
   - Data anomalies
   - Exploratory Data Analysis
2. Python Plot Tutorial:
   - http://pandas.pydata.org/pandas-docs/stable/visualization.html
   - http://matplotlib.org/users/tutorials.html
3. Attend tutorial for week 6 in next week
4. **Assessment 1** is due this week (2 Sept)!