Data Science
COMP5122M

Data preparation

Roy Ruddle

Private study

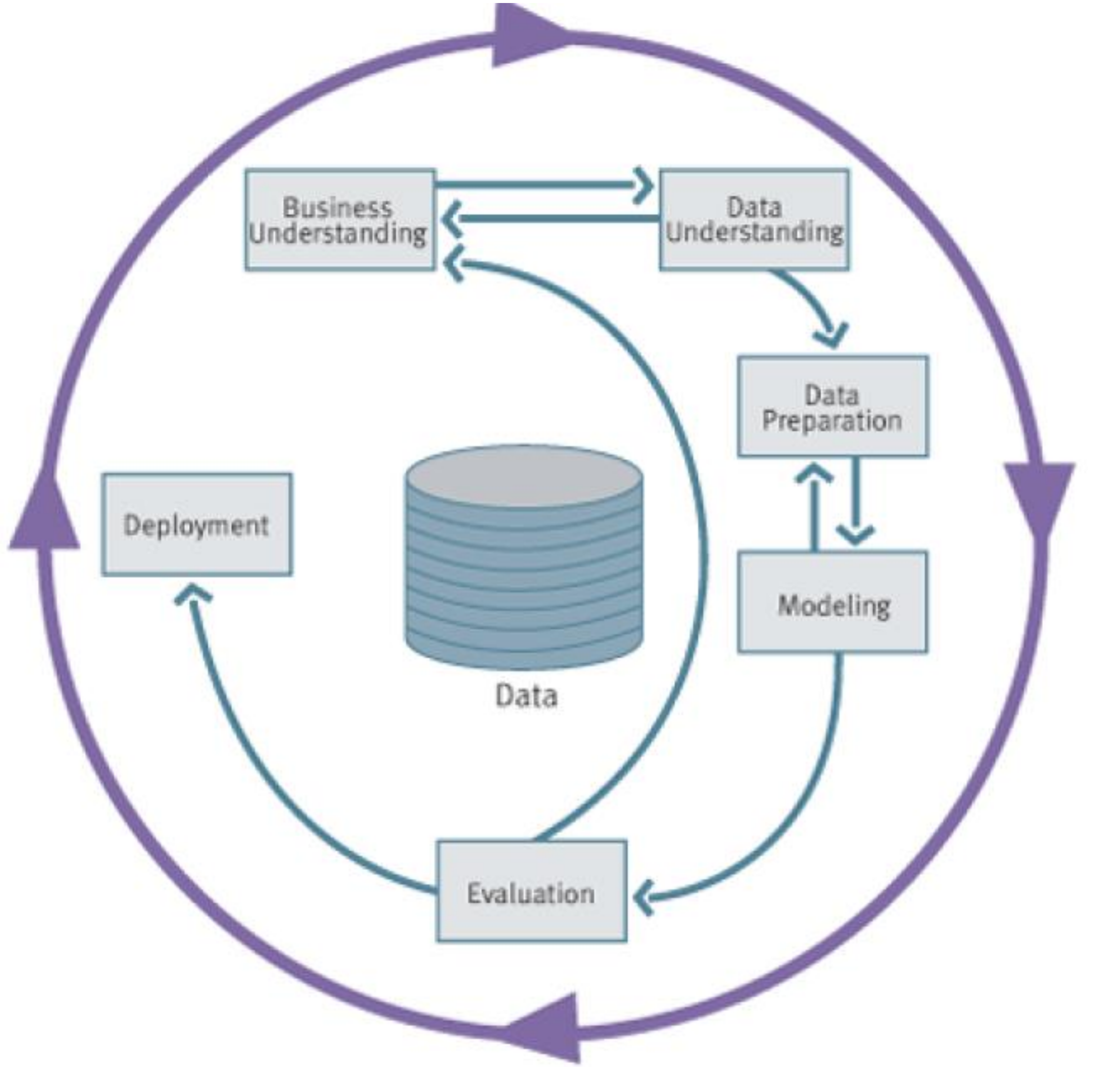
- See Minerva Announcements for up-to-date info
- Private study for this lecture
  - Data quality film https://tinyurl.com/VizDataQuality

What will you learn?

- A data preparation workflow
- How to use data profiling methods to
  - Characterise data and provide high-level insights
  - Investigate data quality so it may be cleaned
- Why it's important to use both calculations and visualizations
- How to write data validation rules

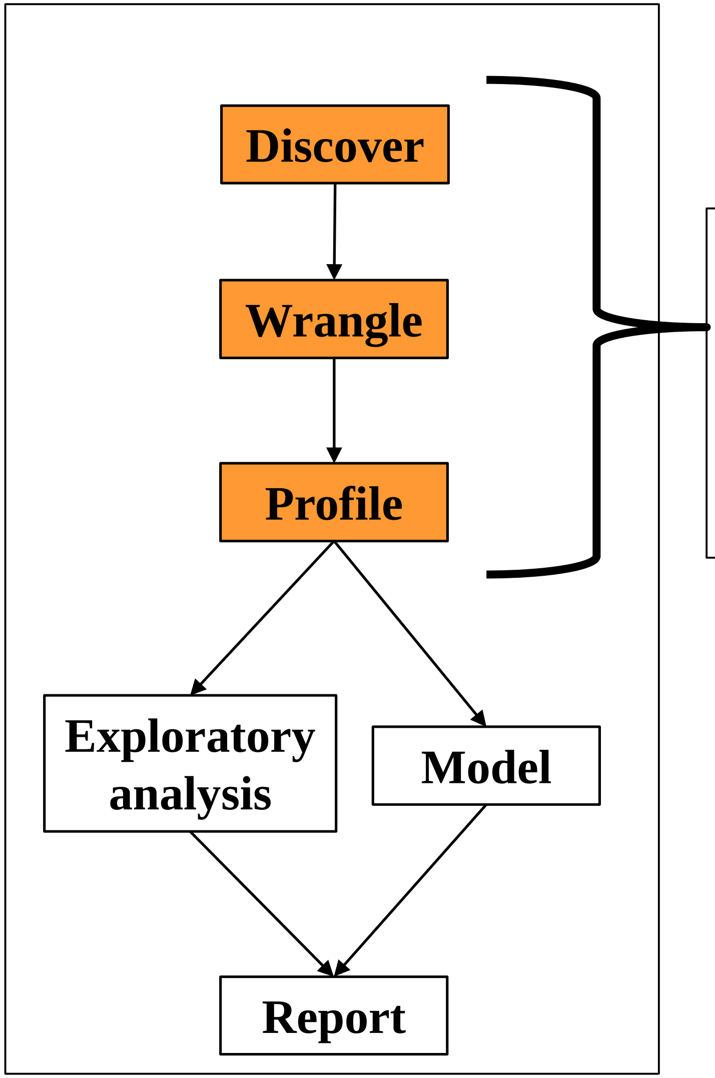
- "Everything that can be wrong will be wrong, on some occasion"

CRISP data mining process



Cross-Industry Standard Process for Data Mining (Shearer, 2000).

Data science workflow[1]



"I spend more than half of my time integrating, cleansing and transforming data without doing any actual analysis"[2]

I.e., doing data preparation

[1] Alspaugh et al. (2018). IEEE Transactions on Visualization and Computer Graphics.
[2] Kandel et al.. (2012). IEEE Transactions on Visualization and Computer Graphics.

Data preparation workflow

- Discover
  - What data sources and level of detail (see Data Understanding lecture)
  - What spatio-temporal coverage and cost?
- Wrangle
  - Read in data, reformat, transform, link
- Profile
  - Rigorous investigation of data quality

Profiling tasks

Characterisation tasks

Cardinalities

- Number of distinct values
- Number of rows
- Value lengths

Distribution

- First digit
- Frequency measures (count, etc.)
- Mean, median etc.
- Outliers
- Range (percentile, etc.)
- Variance, etc.

Patterns

- Character types
- Clusters
- Correlation
- Cross tabulation
- Curve fitting
- Data format
- Data type
- Example values
- Precision
- Principal Components Analysis
- Trends
- Value patterns

Data quality tasks

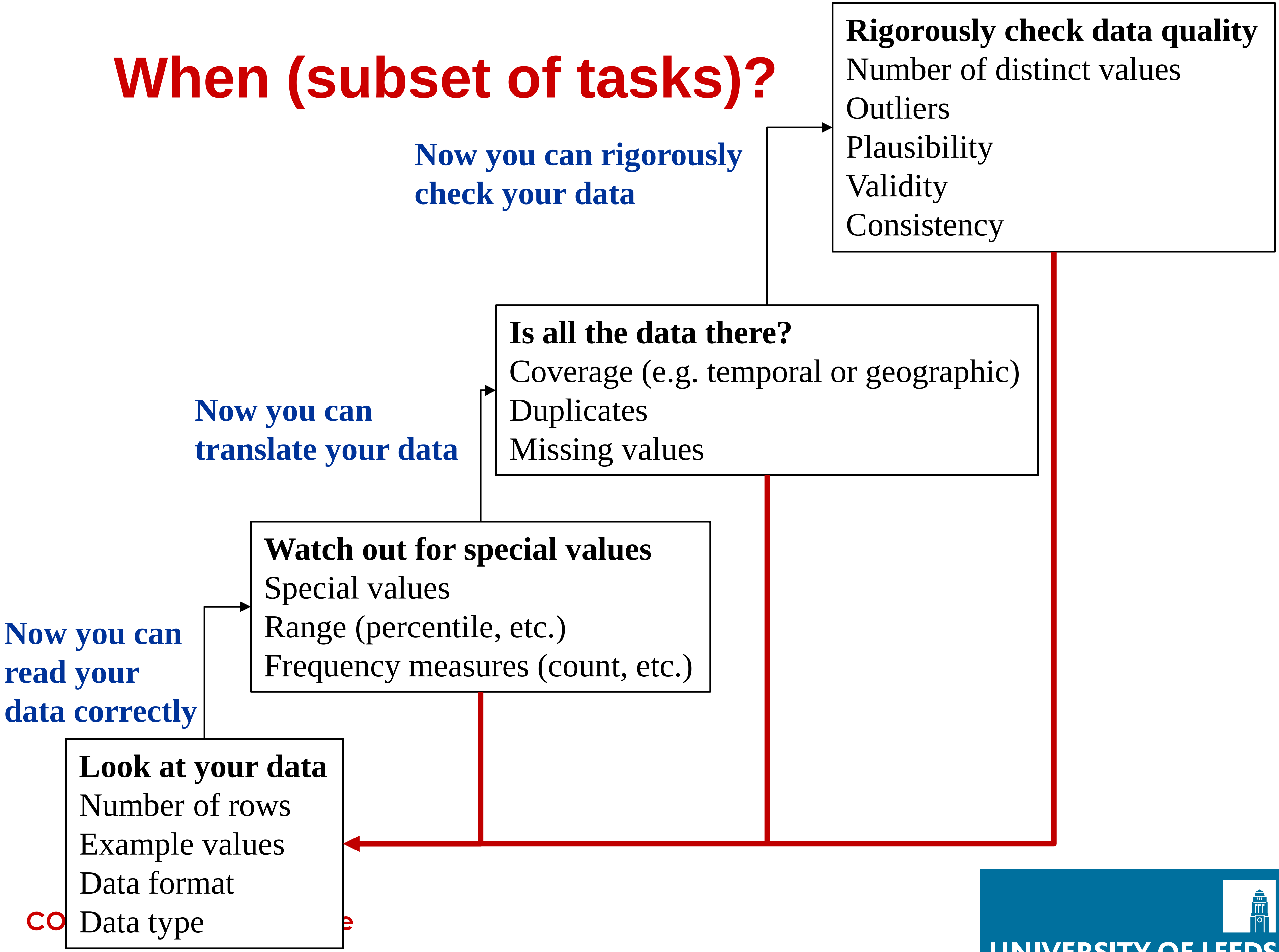
Completeness

- Coverage (e.g. temporal or geographic)
- Duplicates
- Missing records
- Missing values
- Rate of recording
- Recency

Correctness

- Accuracy
- Bias
- Consistency
- Integrity
- Misleadingness
- Noise
- Outlier
- Plausibility
- Special values
- Use of default values
- Validity
- Variation

When (subset of tasks)?



Look at your data
Number of rows
Example values
Data format
Data type

Now you can read your data correctly

Watch out for special values
Special values
Range (percentile, etc.)
Frequency measures (count, etc.)

Now you can translate your data

Is all the data there?
Coverage (e.g. temporal or geographic)
Duplicates
Missing values

Now you can rigorously check your data

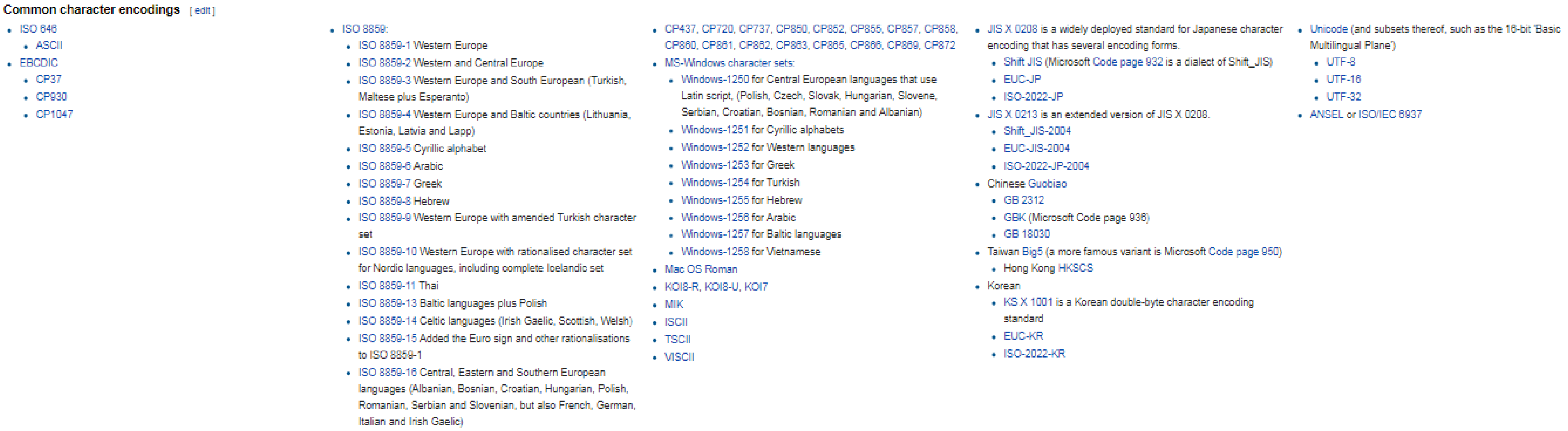
Rigorously check data quality
Number of distinct values
Outliers
Plausibility
Validity
Consistency

Step 1: Look at your data

- How is it encoded?
- Determine file size & number of rows
- Check data types
- Check data formats
- Print example values

Text file encodings

- There are many different character encodings



Common character encodings [edit]

- ISO 646
  - ASCII
- EBCDIC
  - CP37
  - CP930
  - CP1047
- ISO 8859:
  - ISO 8859-1 Western Europe
  - ISO 8859-2 Western and Central Europe
  - ISO 8859-3 Western Europe and South European (Turkish, Maltese plus Esperanto)
  - ISO 8859-4 Western Europe and Baltic countries (Lithuania, Estonia, Latvia and Lapp)
  - ISO 8859-5 Cyrillic alphabet
  - ISO 8859-6 Arabic
  - ISO 8859-7 Greek
  - ISO 8859-8 Hebrew
  - ISO 8859-9 Western Europe with amended Turkish character set
  - ISO 8859-10 Western Europe with rationalised character set for Nordic languages, including complete Icelandic set
  - ISO 8859-11 Thai
  - ISO 8859-13 Baltic languages plus Polish
  - ISO 8859-14 Celtic languages (Irish Gaelic, Scottish, Welsh)
  - ISO 8859-15 Added the Euro sign and other rationalisations to ISO 8859-1
  - ISO 8859-16 Central, Eastern and Southern European languages (Albanian, Bosnian, Croatian, Hungarian, Polish, Romanian, Serbian and Slovenian, but also French, German, Italian and Irish Gaelic)
- CP437, CP720, CP737, CP850, CP852, CP855, CP857, CP858, CP860, CP861, CP862, CP863, CP865, CP866, CP869, CP872
- MS-Windows character sets:
  - Windows-1250 for Central European languages that use Latin script, (Polish, Czech, Slovak, Hungarian, Slovene, Serbian, Croatian, Bosnian, Romanian and Albanian)
  - Windows-1251 for Cyrillic alphabets
  - Windows-1252 for Western languages
  - Windows-1253 for Greek
  - Windows-1254 for Turkish
  - Windows-1255 for Hebrew
  - Windows-1256 for Arabic
  - Windows-1257 for Baltic languages
  - Windows-1258 for Vietnamese
- Mac OS Roman
- KOI8-R, KOI8-U, KOI7
- MIK
- ISCII
- TSCII
- VISCII
- JIS X 0208 is a widely deployed standard for Japanese character encoding that has several encoding forms.
  - Shift JIS (Microsoft Code page 932 is a dialect of Shift_JIS)
  - EUC-JP
  - ISO-2022-JP
- JIS X 0213 is an extended version of JIS X 0208.
  - Shift_JIS-2004
  - EUC-JIS-2004
  - ISO-2022-JP-2004
- Chinese Guobiao
  - GB 2312
  - GBK (Microsoft Code page 936)
  - GB 18030
- Taiwan Big5 (a more famous variant is Microsoft Code page 950)
  - Hong Kong HKSCS
- Korean
  - KS X 1001 is a Korean double-byte character encoding standard
  - EUC-KR
  - ISO-2022-KR
- Unicode (and subsets thereof, such as the 16-bit 'Basic Multilingual Plane')
  - UTF-8
  - UTF-16
  - UTF-32
- ANSEL or ISO/IEC 6937

https://en.wikipedia.org/wiki/Character_encoding

Encodings: Why should I care?

- If you use anything other than the most basic English text, people may not be able to read your data unless you state the character encoding
  - Your software may crash
  - Your may get error messages

Pandas.read_csv(): UnicodeEncodeError: 'ascii' codec can't encode character '\xab' in position 0: ordinal not in range(128)

  - Your data may contain spurious characters
- E.g., common names and words in many languages
  - Jürgen Klopp (German football manager)
  - Écrire (to write; French)

What is character encoding?

- Data values are created from characters
  - E.g., “Male” or “1.6846”
- Characters are stored on disk and in memory as one or more bytes
- An encoding maps the byte values to a specific character
  - Without the mapping, the data looks like garbage
- E.g.
  - ASCII encoding (1977/1986): 128 characters
  - Other encodings
    - More characters
    - Sometimes require more memory

https://www.w3.org/International/questions/qa-what-is-encoding

ASCII character set (1977/1986)

- **American Standard Code for Information Interchange**
- **128 characters**
  - **No British pound sign £**
  - **Let alone a Euro sign €**
- **ASCII vs. Windows-1252**
  - **https://en.wikipedia.org/wiki/Windows-1252**

Reccommendations

- Reading text files
  - Is the encoding specified?
  - Check it!
- Saving text files
  - Use the UTF-8 encoding
    - 1 – 4 bytes per character
    - It includes any character you are likely to need
    - Removes the need to track and convert between various character encodings

```
Pandas.to_csv(encoding='utf_8‘)
```

Detecting a file’s encoding

- Python chardet package
  - Reads file completely or incrementally
  - Returns
    - Encoding (e.g., ‘ascii’, ‘windows-1252’, ‘utf-8’)
    - Confidence (0.0 – 1.0)
  - Based on heuristics, so not guaranteed to be correct
- Now part of the Anaconda installation of Python/Pandas

See VLE: teaching chardet.py

File size & number of rows

- E.g., Leeds off-street parking fines
  - https://datamillnorth.org/dataset/off-street-parking-fines

File name	File size
Quarter 1 - 2016/17	1.15 kB
Quarter 2 - 2016/17	1.28 kB
Quarter 3 - 2016/17	2.45 kB
Quarter 4 - 2016/17	4.43 kB
Quarter 1 - 2017/18	54.50 kB
Quarter 2 - 2017/18	233.99 kB
Quarter 3 - 2017/18	630.08 kB
Quarter 4 - 2017/18	722.24 kB

Check the data types

Name	Gender	Date of birth	Number of attempts	Highest mark
John Smith	1	01/02/2001	3	51
Fiona May	2	09/07/2001		80
Emily Jones	2	06/05/2001	1	63

Variable	Data type
Name	object
Gender	int64
Date of birth	object
Number of attempts	float64
Highest mark	int64

What is wrong?

See VLE: teaching data type.py

Check data formats (especially dates)

- Check the format yourself
  - Don’t rely on heuristics
  - Don’t assume that all your data files use the same format, even if the files come from one source

Name	Date of birth	Day first	Year first	Country
John Smith	01/02/2001	True	False	UK
Fiona May	2001/02/01	False	True	General
Emily Jones	02/01/2001	False	False	USA

- Set appropriate parameters

```
Pandas. to_datetime(df['Date of birth'], dayfirst=True, yearfirst=False)
```

Example values

- Print out a few lines
  - What delimiter is used?

```
Name,Age,Gender
John Smith,25,Male
```

```
Name|Age|Gender
John Smith|25|Male
```

  - How are missing values indicated?

```
Name,Age,Gender
John Smith,,Male
```

```
Name,Age,Gender
John Smith,NULL,Male
```

  - What character is used to denote a quote?

```
Name,Age,Gender
John Smith,25,Male
```

```
Name,Age,Gender
“John Smith”,25,”Male”
```

- Choose appropriate parameters, e.g.

```
Pandas.read_csv(sep=‘,’, na_values=‘NULL’, quoting=csv.QUOTE_NONE)
```

Now you can read your data correctly

- Did any warnings occur?

Pandas.read_csv(): DtypeWarning: Columns (3, 5) have mixed types. Specify dtype option on import or set low_memory=False

- By default, Pandas infers the data types after reading part of a file
  - Inference can be wrong, e.g.,
    - Variable has many missing values
    - Variable is categorical, but the first values contain only digits
  - Or the data may actually be inconsistent
    - E.g., record split between two lines

```
ID,Name
1010,John Smith
...
101A,Emily Jones
```

```
Name,Gender,Age
John Smith,Male,26
Fiona May,Female
19
Emily Jones,Female,20
```

Solutions for data type warnings

- Specify the data types

```
Pandas.read_csv(dtype=xxx)
```

- Read the whole file

```
Pandas.read_csv(low_memory=False)
```

Step 2: Watch out for special values

- E.g., NHS data is cleaned to set
  - 01/01/1800 (missing date)
  - 01/01/1801 (invalid date)
  - See Visualizing the Quality of Data
    - https://tinyurl.com/VizDataQuality (18:08 - 19:07)
- Range, percentiles are informative

Variable	Min	25th	50th	75th	Max.
Measurement	0	1	8	60	99,999,999,999
Dose	-32.1	1.6	11	77	348,500
Year of birth	1900	1941	1962	2001	2106

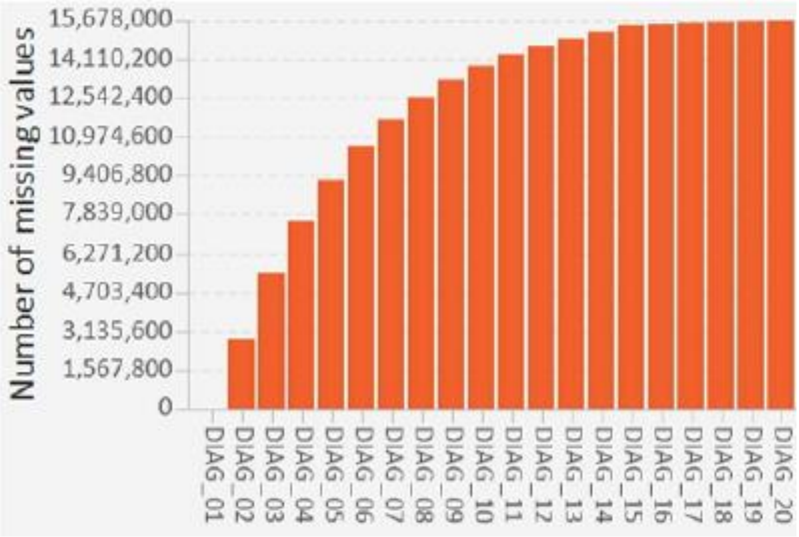
Step 3: Is all the data there?

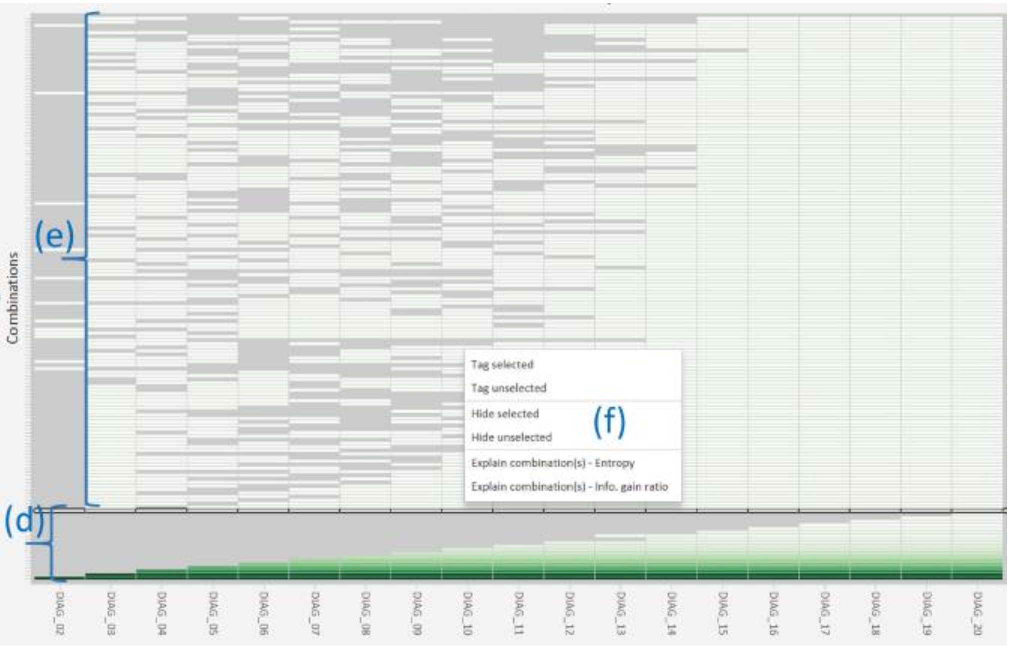
- **Missing values**
  - **Terrible statistical terminology**
  - **Advantages of visualization**
- **Coverage**

Statistical terminology

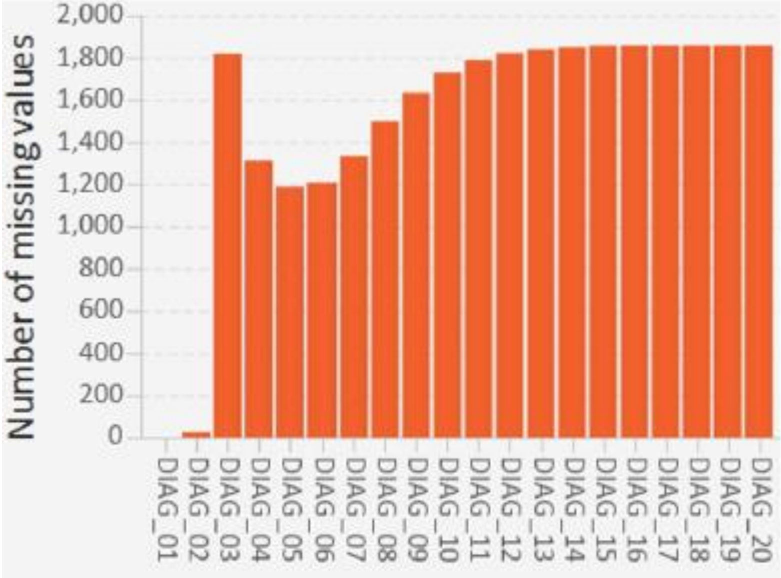
- Missing at random (MAR)
  - Related to other variables
  - Term is misleading!
- Missing completely at random (MCAR)
  - Haphazard
  - Unrelated to values of variable, or other variables
- Missing not at random (MNAR)
  - Related to values of the variable itself

Visualizing missing values





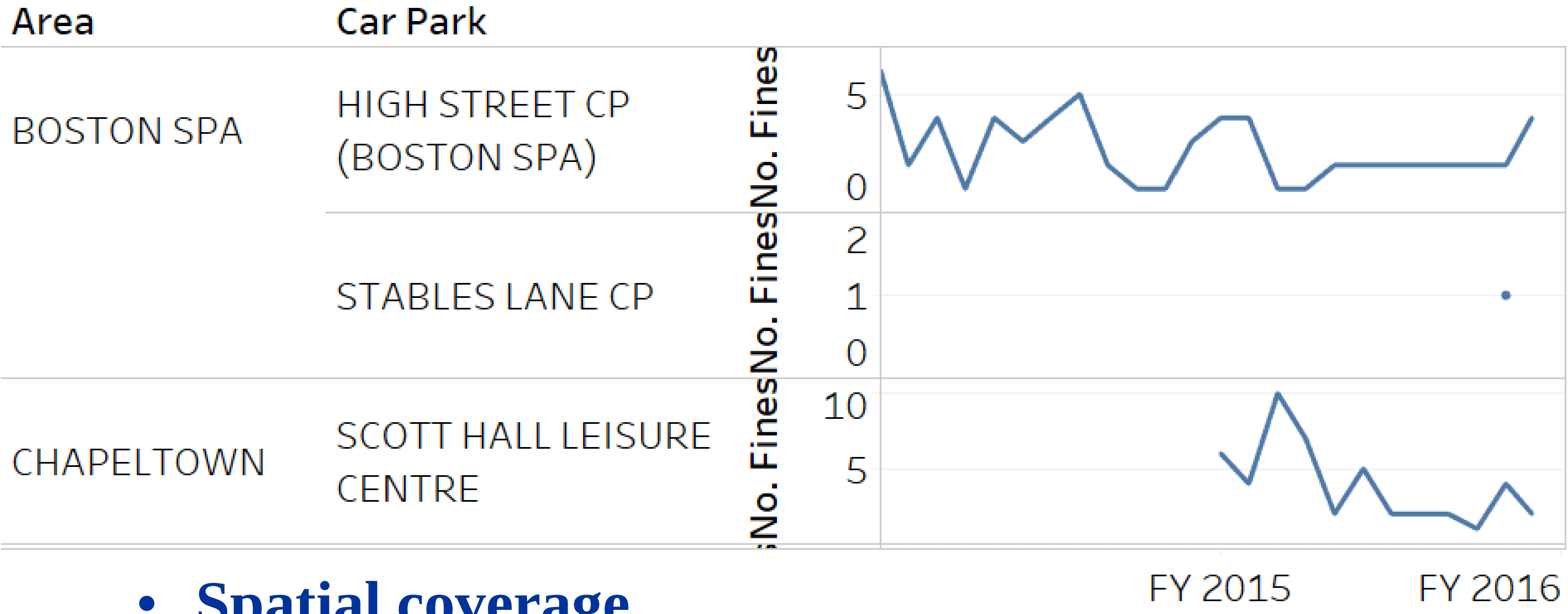
Gaps in diagnosis codes (NHS hospital data)



See Visualizing the Quality of Data
https://tinyurl.com/VizDataQuality

Coverage

- **Temporal coverage (Tableau Challenge #3)**



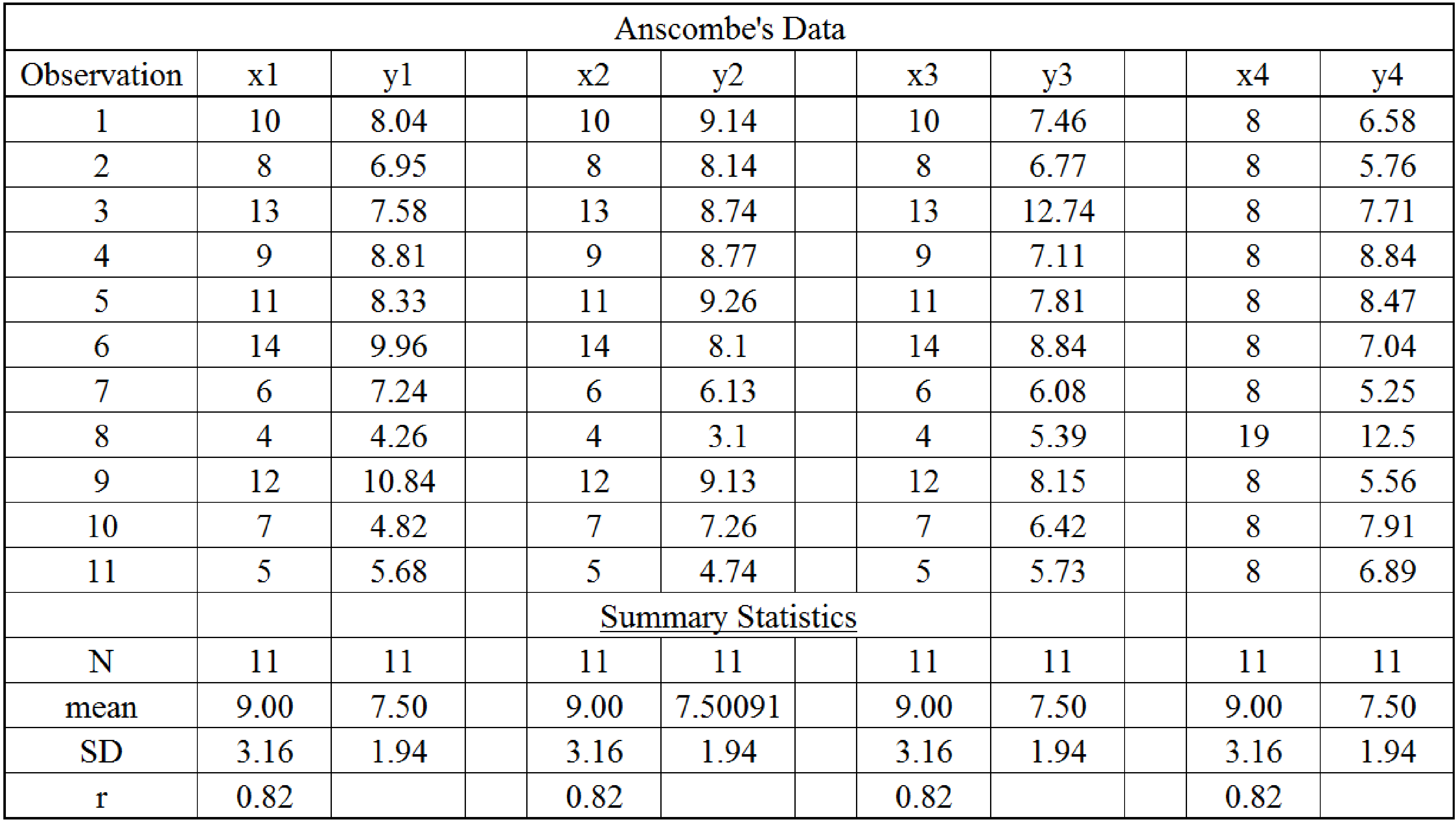
- **Spatial coverage**
  - **See Visualizing the Quality of Data https://tinyurl.com/VizDataQuality**

Step 4: Rigorously check data quality

- **See Visualizing the Quality of Data**
  https://tinyurl.com/VizDataQuality

Why do we need visualization?

- Anscombe's quartet
  - Four datasets that have nearly identical statistical properties

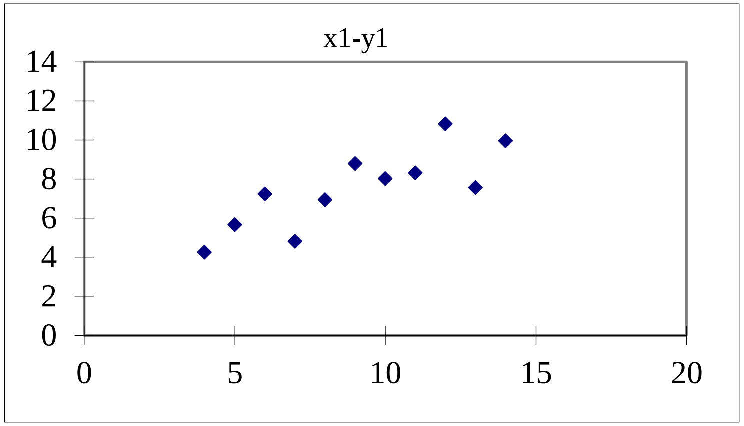


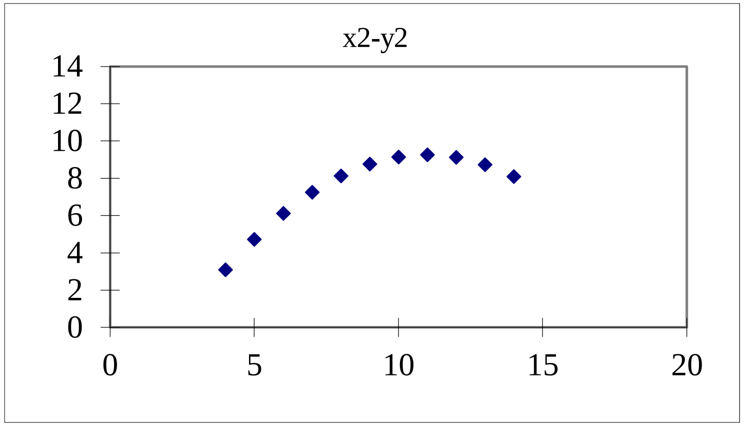
Anscombe's Data

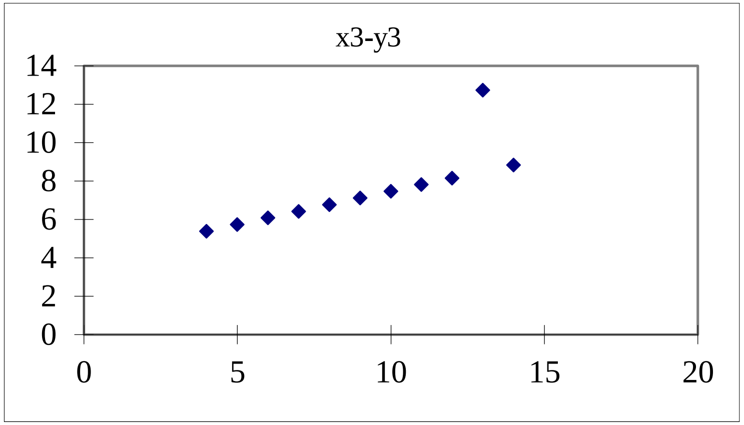
Observation	x1	y1		x2	y2		x3	y3		x4	y4
1	10	8.04		10	9.14		10	7.46		8	6.58
2	8	6.95		8	8.14		8	6.77		8	5.76
3	13	7.58		13	8.74		13	12.74		8	7.71
4	9	8.81		9	8.77		9	7.11		8	8.84
5	11	8.33		11	9.26		11	7.81		8	8.47
6	14	9.96		14	8.1		14	8.84		8	7.04
7	6	7.24		6	6.13		6	6.08		8	5.25
8	4	4.26		4	3.1		4	5.39		19	12.5
9	12	10.84		12	9.13		12	8.15		8	5.56
10	7	4.82		7	7.26		7	6.42		8	7.91
11	5	5.68		5	4.74		5	5.73		8	6.89
				Summary Statistics							
N	11	11		11	11		11	11		11	11
mean	9.00	7.50		9.00	7.50091		9.00	7.50		9.00	7.50
SD	3.16	1.94		3.16	1.94		3.16	1.94		3.16	1.94
r	0.82			0.82			0.82			0.82	

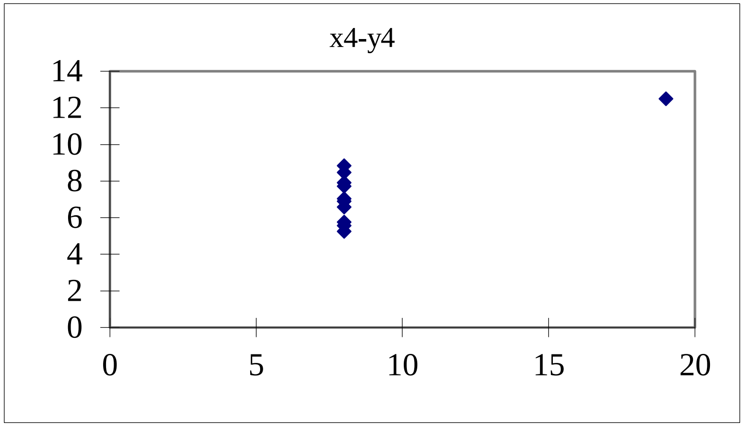
Anscombe’s quartet visualizations

- **What do the scatter plots tell you?**









How to write data validation rules

- Subject-matter specialists typically use free text to describe valid values and explain how to clean them
  - E.g., during interview
- Data scientist may need to write validation & cleaning rules as pseudocode
  - Less ambiguous than text
  - Easier to implement in software

Example NHS validation rule

The main specialty (**mainspef**) is the speciality of the doctor who treated you. The mainspef is a three-digit code, which ranges from 100 (General Surgery) to 960 (Allied Health Professional Episode).

Validation

If the main specialty is null or contains an invalid entry, it is overwritten with the appropriate value for not known (&)

```
For all records
  If mainspef is NULL or mainspef is not a valid code
    Set mainspef equal to ‘&’
```

Example NHS correction rule

Rule 150: Epitype reset to 3

When the episode type is not coded as an NHS hospital birth record, the admission method (admimeth), date of birth (dob), episode order (epiorder) and episode start date (epistart) are examined to see whether they indicate that the record is a birth record.

If so, the episode type (epitype) is changed to reflect this.

```
For all records
  If epitype = 1,2,4
  and dob <> null and epistart <> null
  and dob = epistart
  and admimeth = 82
  and epiorder = 01
    Set epitype = 3
```

```
NB: <> means "not equal to"
```

Tips for writing rules (1)

- **Write rules in the style of pseudocode**
- **Neat layout (indentation like Python)**
- **What is the difference between?**

```
If date < 01/01/1900 or date > today
   Set date equal to ‘&’
```

```
If date < 01/01/1900 and date > today
   Set date equal to ‘&’
```

Tips for writing rules (2)

- **What is the difference between?**

```
If patient_ID is blank
   set patient_ID equal to ‘&’

If patient_ID <> blank
   do some other validation
```

```
If patient_ID is blank
   set patient_ID equal to ‘&’
Else
   do some other validation
```

Transforming data for modelling

- Part of data wrangling, but should be done after data cleaning
- Depends on your modelling software
- Particular challenges for dates/times and categorical variables

Transforming dates/times

- Software May not directly support dates or times
- Transform to value after start date/time?
  - E.g., days since xxx
  - This is how computer operating systems and software work
    - Try typing this into Excel

```
=DATEVALUE("01/01/2021")
```

Transforming categorical variables

- **Software may**
  - **Directly handle categories (A, B, etc.), or**
  - **Require categories to be represented as numbers (1, 2, etc.), or**
  - **Require categories to be one-hot encoded**

Name	Nationality
John Smith	English
Fiona May	Scottish
Emily Jones	Welsh

Direct

Name	Nationality
John Smith	1
Fiona May	2
Emily Jones	3

Represent as numbers

Name	English	Scottish	Welsh
John Smith	1	0	0
Fiona May	0	1	0
Emily Jones	0	0	1

One-hot encoding

Quiz

- On Minerva
  - Assessment -> Formative test: Data understanding & preparation
  - Available from 10am today (complete it by 23:59 hrs on Thursday)
  - At the end, click OK to review results, see which questions you got correct/wrong and why