MONASH University

FIT1043 Introduction to Data Science

Week 4: Data Wrangling

Ts. Dr. Sicily Ting

School of Information Technology
Monash University Malaysia

With materials from Wray Buntine, Mahsa Salehi

Learning Outcomes

Week 4

By the end of this week you should be able to:

- *Explain open data and linked open data*
- *Explain how to access to new data sources through APIs*
- *Identify how different APIs work*
- Inspect data quality problems in datasets and recommend solutions to fix them
- *Use data wrangling operations in Python*

Data Wrangling

Manipulating data to make it directly usable for analysis

Week 4

- Common types of data that are at our disposal.
- Data wrangling and the goals of wrangling
- Data quality
- Data auditing
- Examples

Data

MONASH University

Databases

Relational Databases

- Oracle
- MySQL
- MariaDB
- MS-SQL

Containing data for

- Customer relationship management
- Mortgage, hire purchase, business loan system
- Salesforce automation
- Credit card system, ATM transactions, retail banking
- Human resource and payroll

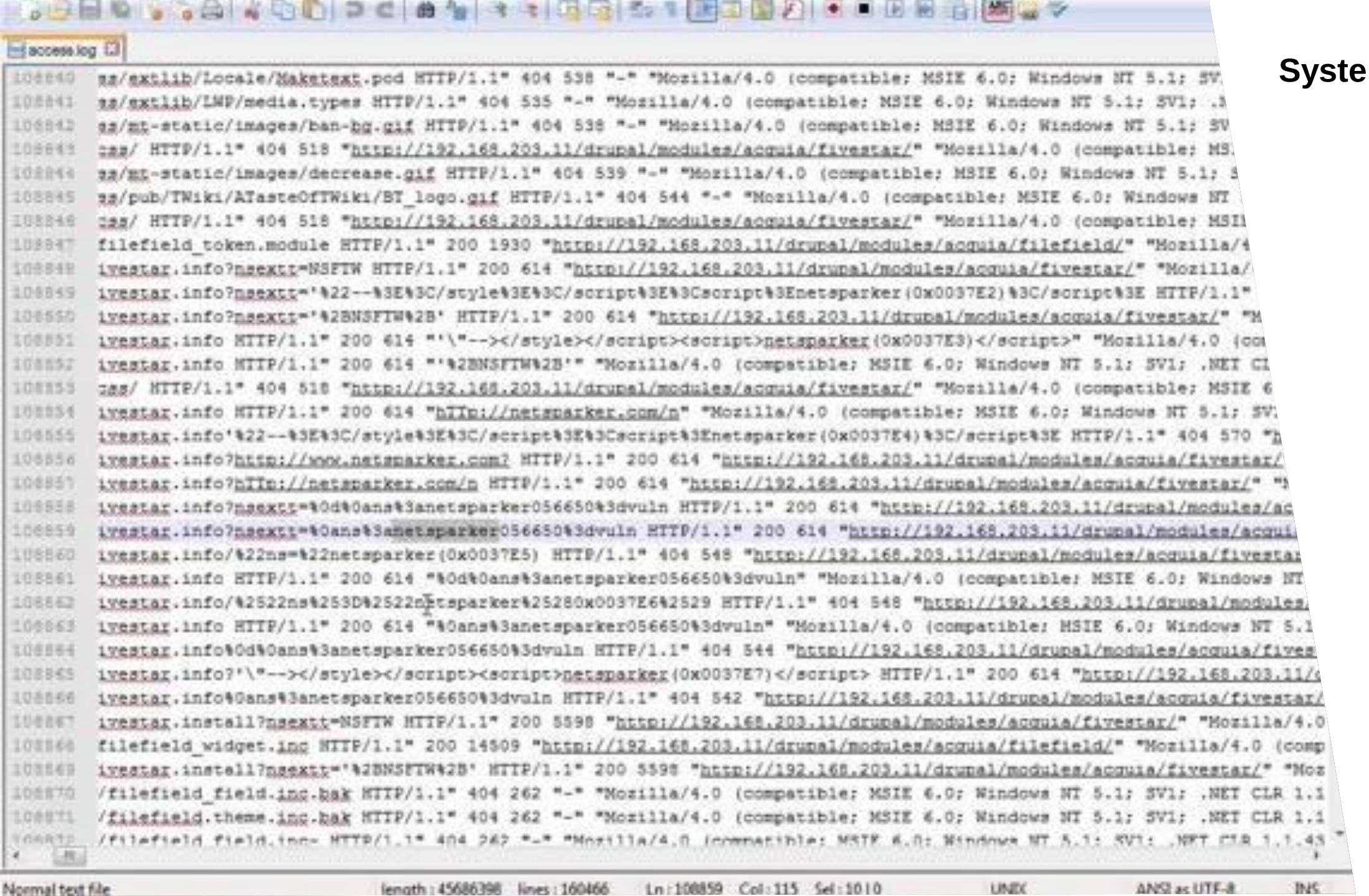
Files

Examples

- ***System log files***
- Spreadsheets
- PDF files
- Image files
- Raw text files
- Formatted text files

MONASH University

System Log Files



MONASH University

Web & Crowd Sources

Open Data, REST API, Web Scrapping, etc.

Huge!

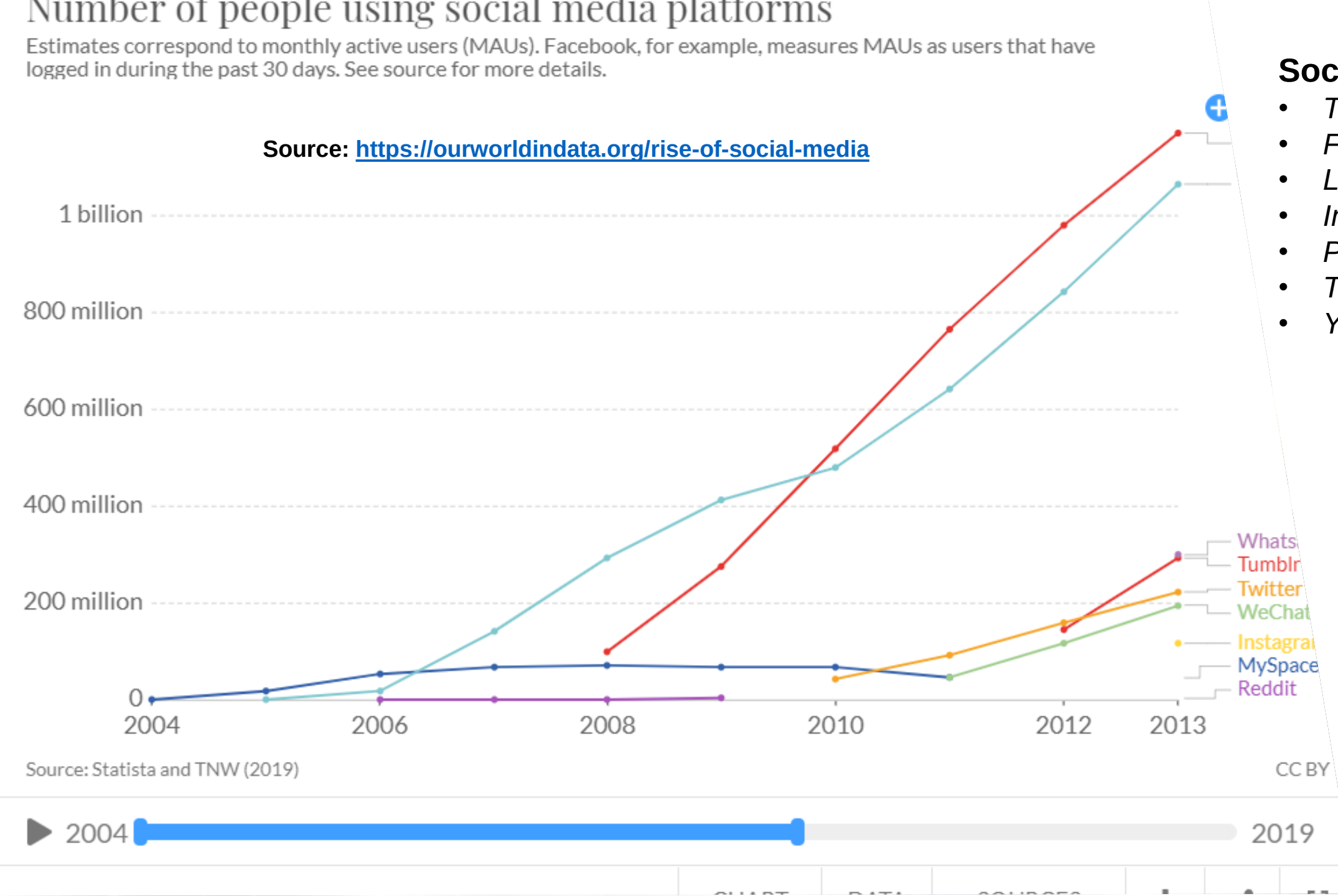
- What can we do with Web and other crowd sourced data?
- What type of web data?

Data on the Web

- News: Sports, Finance, World, Nation
- Blogs
- Corporates: Products & Services
- Government

GDELT

MONASH University



Source: https://ourworldindata.org/rise-of-social-media

Social Media

- *Twitter*
- *Facebook*
- *LinkedIn*
- *Instagram*
- *Pinterest*
- *TikTok*
- *YouTube*

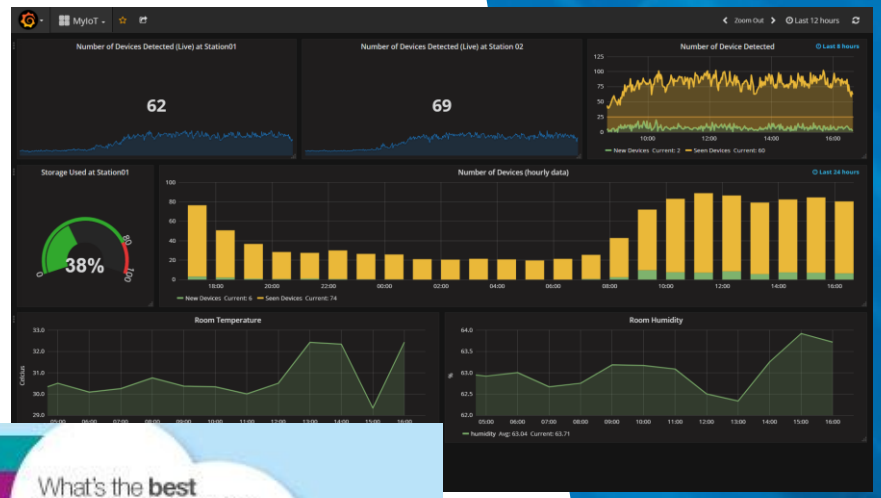
MONASH University

Internet of Things (IoT)

The Future

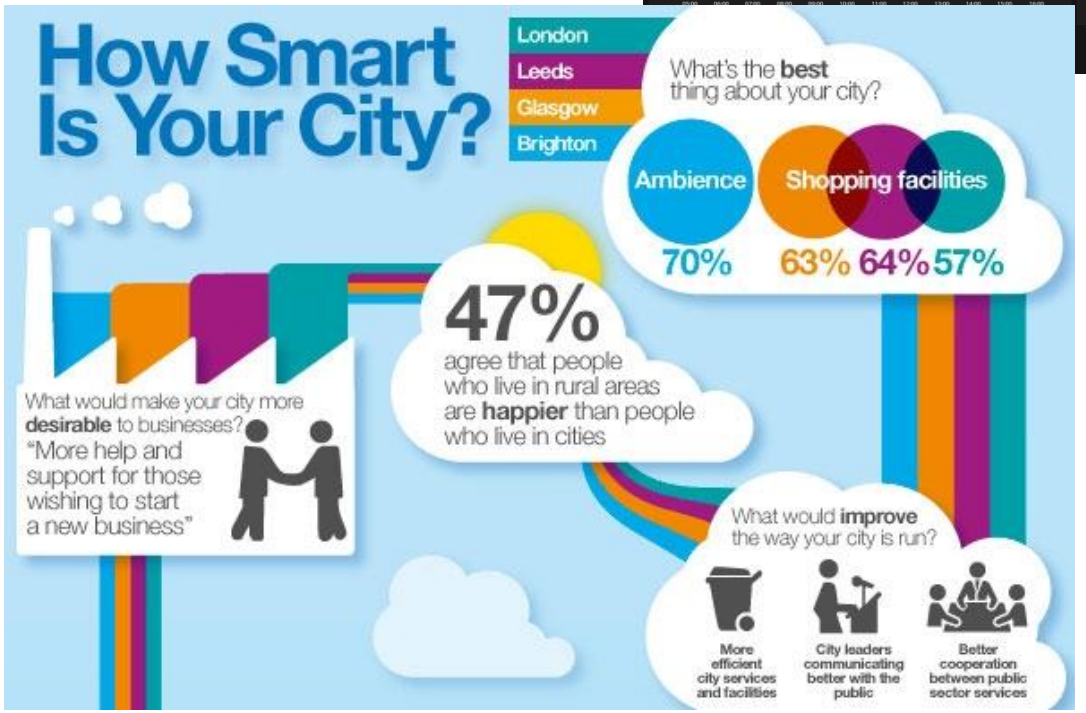
Where?

- Utilities – Water, Power, Street Lighting, Traffic Lights, Waste Disposal
- Vehicle to Vehicle, Autonomous Driving
- Monitoring Systems



Mobile Phone Data

- Location
- Browsing history
- Usage history
- Personal information



MONASH University

Data Wrangling

MONASH University

Ideal Data

	C	D	E	F	G	H	I	J	K	L
	77516	Bachelors	13	Never-married	Adm-clerical	Not-in-famil	White	Male	2174	
no	83311	Bachelors	13	Married-civ-spouse	Exec-managerial	Husband	White	Male	0	
	215646	HS-grad	9	Divorced	Handlers-cleaners	Not-in-famil	White	Male	0	0
	234721	11th	7	Married-civ-spouse	Handlers-cleaners	Husband	Black	Male	0	0
	338409	Bachelors	13	Married-civ-spouse	Prof-specialty	Wife	Black	Female	0	0
	284582	Masters	14	Married-civ-spouse	Exec-managerial	Wife	White	Female	0	0
	160187	9th	5	Married-spouse-abser	Other-service	Not-in-famil	Black	Female	0	0
no	209642	HS-grad	9	Married-civ-spouse	Exec-managerial	Husband	White	Male	0	0
	45781	Masters	14	Never-married	Prof-specialty	Not-in-famil	White	Female	14084	0
	159449	Bachelors	13	Married-civ-spouse	Exec-managerial	Husband	White	Male	5178	0
	280464	Some-colleg	10	Married-civ-spouse	Exec-managerial	Husband	Black	Male	0	0
	141297	Bachelors	13	Married-civ-spouse	Prof-specialty	Husband	Asian-Pac-Isl	Male	0	0
	122272	Bachelors	13	Never-married	Adm-clerical	Own-child	White	Female	0	0
	205019	Assoc-acdm	12	Never-married	Sales	Not-in-famil	Black	Male	0	0
	121772	Assoc-voc	11	Married-civ-spouse	Craft-repair	Husband	Asian-Pac-Isl	Male	0	0
	245487	7th-8th	4	Married-civ-spouse	Transport-moving	Husband	Amer-Indian	Male	0	0

MONASH University

CT Chest Hi Resolution 30/11/04 at 2156 CT-04-

REPORT:

Clinical note: transformed AML. Ongoing fevers.? Source. ? fungal infe

Report:

Axial 1.25 mm slices at 10 mm intervals taken in inspiration with selec
images in the prone position.

No mediastinal or hilar lymphadenopathy. Heart size is normal. Borderlir
enlargement of the main pulmonary outflow tract. There is smooth interlo
septal thickening throughout both lungs, which may be secondary to fluid
overload. There is a background of emphysematous changes , predominantly
the upper lobes. A 5 x 8 mm nodule is identified in the right upper lobe
(image 10). It is well-circumscribed with no evidence of surrounding
ground-glass opacity. No calcification or cavitation of this lesion. The
visualised portions of the liver and spleen appear normal, allowing for la
of intravenous contrast.

Conclusion:

Single nodule in right upper lobe has a non-specific appearance but given t
clinical history, this could represent a focus of fungal infection.

Reported by: D

PJL/PJL

A1.2f

Result type: CT Chest Hi Resolution
Result date: 11 January 2005 12:21
Result status: Auth (Verified)
Result title: CTCHEHI

Data Reality

MONASH University

Why Wrangle?

Working with Raw Data is challenging

- Data comes in all shapes and sizes
- Different files have different formatting
- Mistakes in data entries

We need techniques to cleanse and prepare the data!

Goals of Data Wrangling

Transform Data into Usable Data

Raw data ⇒ **Data Wrangling** ⇒ Tidy data ⇒ Data Analysis ⇒ Knowledge

Data + Wrangling + Analysis = Data Product

What is Data Wrangling?

Process of transforming "raw" data into data that can be analysed to generate valid actionable results and insights

Steps

- Data pre-processing
- Data preparation
- Data cleansing
- Data transformation
- Etc.

Data Quality

MONASH University

Source of Data Quality Issues

Causes

- Interpretability issue
- Data format issue
- Inconsistent and faulty data
- Missing and incomplete data
- Outliers
- Duplicates

Interpretability Issue

Data quality problems

- Is there a proper documentation about the data?
  - Without proper documentation (i.e., a data dictionary), it is not possible for us to use the synthetic data.

- We might be able to guess the meaning of each column,
  - But in general, we need a data dictionary to explain the fields

Is this data interpretable?

```
32,1,1,95,0,?,0,127,0,.7,1,?,?,1
34,1,4,115,0,?,?,154,0,.2,1,?,?,1
35,1,4,?,0,?,0,130,1,?,?,?,7,3
36,1,4,110,0,?,0,125,1,1,2,?,6,1
38,0,4,105,0,?,0,166,0,2.8,1,?,?,2
38,0,4,110,0,0,0,156,0,0,2,?,3,1
38,1,3,100,0,?,0,179,0,-1.1,1,?,?,0
38,1,3,115,0,0,0,128,1,0,2,?,7,1
38,1,4,135,0,?,0,150,0,0,?,?,3,2
38,1,4,150,0,?,0,120,1,?,?,?,3,1
40,1,4,95,0,?,1,144,0,0,1,?,?,2
```

MONASH University

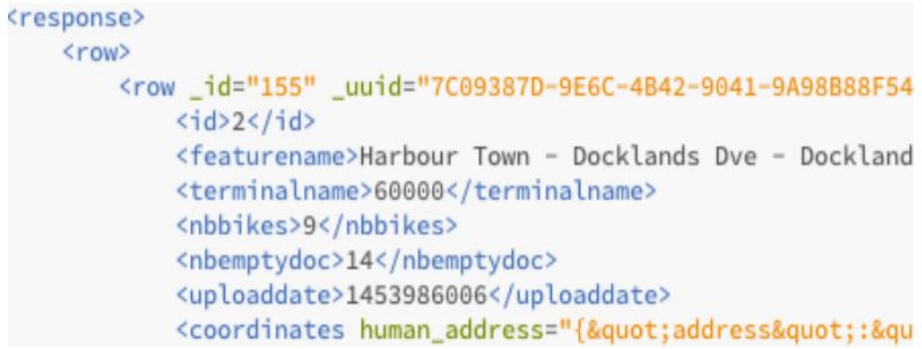
Data format issue

Data quality problems

- Data from different sources often have different data formats and are generated from different processes.
- It is challenging to integrate and manipulate data in different formats.

```
{
  "meta" : {
    "view" : {
      "id" : "tdvh-n9dv",
      "name" : "Melbourne bike share",
      "attribution" : "City of Melbourne, Australia",
      "averageRating" : 0,
      "category" : "Transport & Movement",
      "createdAt" : 1428898164,
      "description" : "Melbourne Bike Share is a joint RACV/Victoria
      "displayType" : "table",
```

JavaScript Object Notation (JSON)



```
<response>
    <row>
        <row _id="155" _uuid="7C09387D-9E6C-4B42-9041-9A98B88F54
            <id>2</id>
            <featurename>Harbour Town - Docklands Dve - Dockland
            <terminalname>60000</terminalname>
            <nbbikes>9</nbbikes>
            <nbemptydoc>14</nbemptydoc>
            <uploaddate>1453986006</uploaddate>
            <coordinates human_address="{&amp;quot;address&amp;quot;:&amp;qu
```

Extensible Mark-up Language (XML)

MONASH University

Inconsistent and faulty data

Data quality problems

- Mistyped data
- Inconsistent entry
- Extraneous data
- Etc.

Mark Johnson, 31, 21/Aug/1985, 180, M, 0433010010, Melbourne VIC

Mr. Christian, Peter, 34, 21-09-1982, , M, 0433010118, Sydney NSW

Ethan Steedman, 32, 01/01/1982, 170, M, 0433210019, Sydney NSW

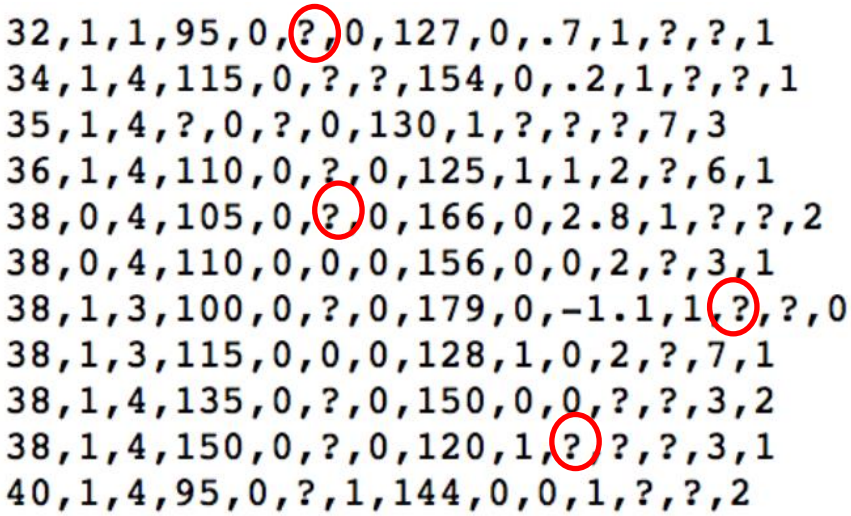
MONASH University

Missing values

Data quality problems

- Data values that should be presented in a dataset but that are absent for many reasons.

Missing values in the Switzerland heart disease data set are indicated by "?".

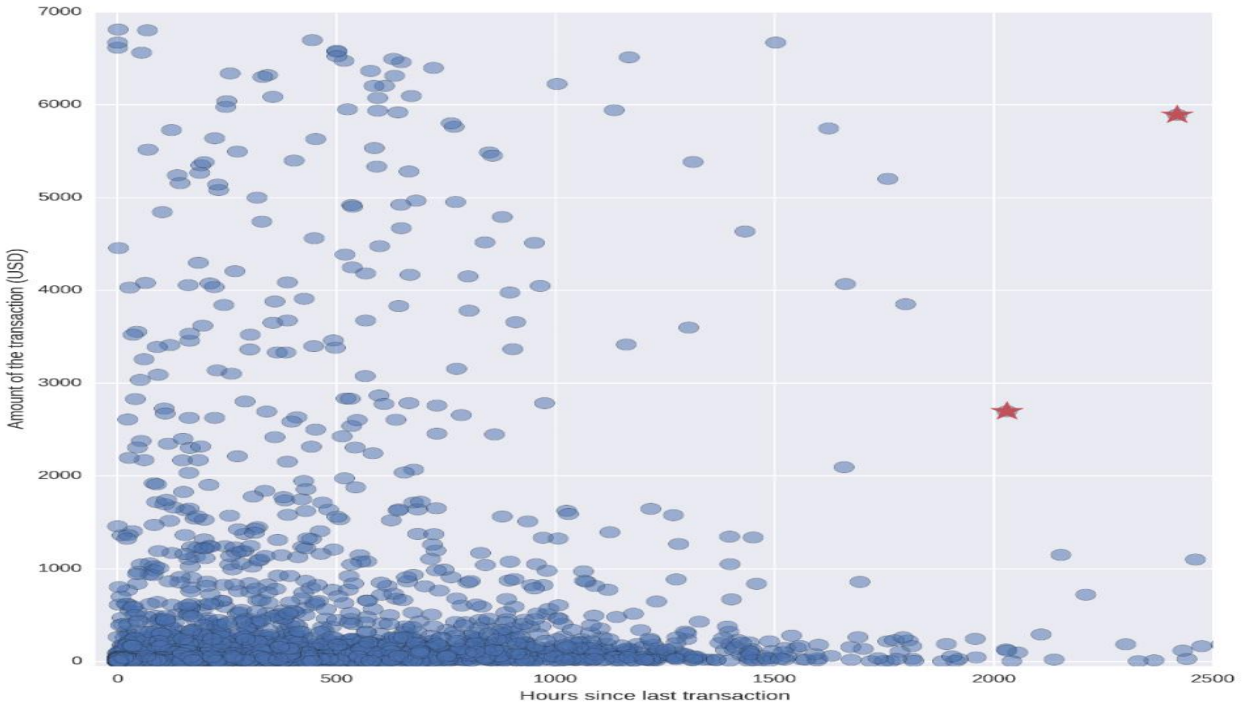


MONASH University

Outliers

Data quality problems

- An observation that lies in an abnormal distance from the majority of the other observations in the dataset.



MONASH University

Duplicates

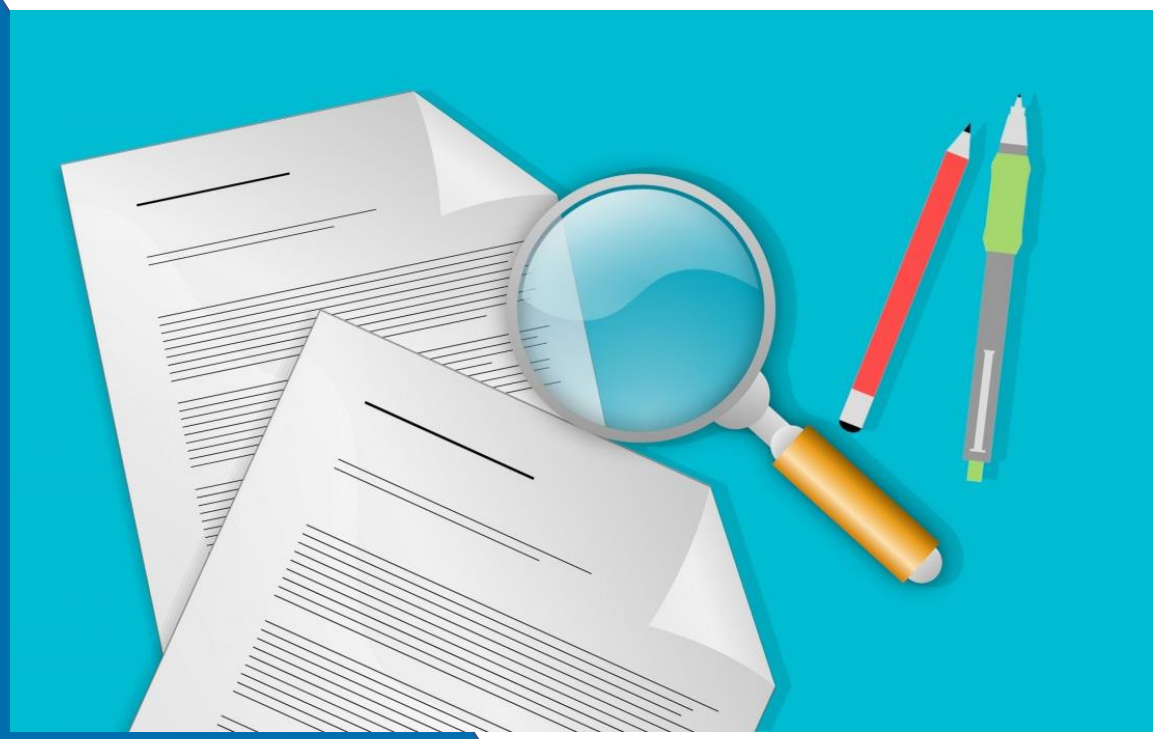
Data quality problems

- Multiple data entries that correspond to the same piece of information

Christoph Cleveland, 20, 10-10-1996, 50, M, 0433550210, Hobart TAS

Chris. Cleveland, 20, 10-10-1996, 176, M, 0433550210, Hobart TAS

Data Auditing



MONASH University

Initial Data Auditing

Given a data set, what are the common initial auditing steps you would conduct?

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
1	survived	pclass	sex	age	sibsp	parch	fare	embarked	class	who	adult_male	deck	embark_tow	alive	alone	name
2	0	3	male	22	1	0	7.25	S	Third	man	TRUE		Southampto	no	FALSE	Braund, Mr. Owe
3	1	1	female	38	1	0	71.2833	C	First	woman	FALSE	C	Cherbourg	yes	FALSE	Cumings, Mrs. J
4	1	3	female	26	0	0	7.925	S	Third	woman	FALSE		Southampto	yes	TRUE	Heikkinen, Miss
5	1	1	female	35	1	0	53.1	S	First	woman	FALSE	C	Southampto	yes	FALSE	Futrelle, Mrs.
6	0	3	male	35	0	0	8.05	S	Third	man	TRUE		Southampto	no	TRUE	Allen, Mr. Will
7	0	3	male		0	0	8.4583	Q	Third	man	TRUE		Queenstown	no	TRUE	Moran, Mr. Jame
8	0	1	male	54	0	0	51.8625	S	First	man	TRUE	E	Southampto	no	TRUE	McCarthy, Mr. T
9	0	3	male	2	3	1	21.075	S	Third	child	FALSE		Southampto	no	FALSE	Palsson, Master
10	1	3	female	27	0	2	11.1333	S	Third	woman	FALSE		Southampto	yes	FALSE	Johnson, Mrs. O
11	1	2	female	14	1	0	30.0708	C	Second	child	FALSE		Cherbourg	yes	FALSE	Nasser, Mrs. Ni
12	1	3	female	4	1	1	16.7	S	Third	child	FALSE	G	Southampto	yes	FALSE	Sandstrom, Miss
13	1	1	female	58	0	0	26.55	S	First	woman	FALSE	C	Southampto	yes	TRUE	Bonnell, Miss.
14	0	3	male	20	0	0	8.05	S	Third	man	TRUE		Southampto	no	TRUE	Saundercock, Mr
15	0	3	male	39	1	5	31.275	S	Third	man	TRUE		Southampto	no	FALSE	Andersson, Mr.
16	0	3	female	14	0	0	7.8542	S	Third	child	FALSE		Southamtpo	no	TRUE	Vestrom, Miss.
17	1	2	female	55	0	0	16	S	Second	woman	FALSE		Southampto	yes	TRUE	Hewlett, Mrs. (
18	0	3	male	2	4	1	29.125	Q	Third	child	FALSE		Queenstown	no	FALSE	Rice, Master. E
19	1	2	male		0	0	13	S	Second	man	TRUE		Southampto	yes	TRUE	Williams, Mr. C
20	0	3	female	31	1	0	18	S	Third	woman	FALSE		Southampto	no	FALSE	Vander Planke,
21	1	3	female		0	0	7.225	C	Third	woman	FALSE		Cherbourge	yes	TRUE	Masselmani, Mrs

MONASH University

Data Auditing

Some general steps to perform data auditing (assuming you're given or have read the data into a DataFrame `df`)

- Dimension of the data: `df.shape()`
  - Number of rows
  - Number of columns

```
(8500, 10)
```

Data Auditing

Head and tail rows: `df.head(); df.tail()`

	survived	pclass	sex	age	sibsp	parch	fare	embarked	class	who	adult_male	deck	embark_town	alive	alone	name
0	0	3	male	22.0	1	0	7.2500	S	Third	man	True	NaN	Southampton	no	False	Braund, Mr. Owen Harris
1	1	1	female	38.0	1	0	71.2833	C	First	woman	False	C	Cherbourg	yes	False	Cumings, Mrs. John Bradley (Florence Briggs Th...
2	1	3	female	26.0	0	0	7.9250	S	Third	woman	False	NaN	Southampton	yes	True	Heikkinen, Miss. Laina
3	1	1	female	35.0	1	0	53.1000	S	First	woman	False	C	Southampton	yes	False	Futrelle, Mrs. Jacques Heath (Lily May Peel)
4	0	3	male	35.0	0	0	8.0500	S	Third	man	True	NaN	Southampton	no	True	Allen, Mr. William Henry

Data Auditing

What are the numerical and categorical columns?

Check basic information about the dataframe

- Number of records,
- Whether there are null values,
- Etc

datatype: `df.info()`

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 892 entries, 0 to 891
Data columns (total 16 columns):
survived       892 non-null int64
pclass         892 non-null int64
sex            892 non-null object
age            715 non-null float64
sibsp          892 non-null int64
parch          892 non-null int64
fare           892 non-null float64
embarked       890 non-null object
class          892 non-null object
who            892 non-null object
adult_male     892 non-null bool
deck           204 non-null object
embark_town    890 non-null object
alive          892 non-null object
alone          892 non-null bool
name           892 non-null object
dtypes: bool(2), float64(2), int64(4), object(8)
memory usage: 99.4+ KB
```

Data Auditing

Check some basic statistics about columns:

- Numerical columns: `df.describe()`
- Object columns: `df.describe(include=[np.object])`

https://pandas.pydata.org/pandas-docs/stable/reference/api/pandas.DataFrame.describe.html

	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	0.805706	49.665589
min	0.000000	1.000000	0.420000	0.000000	0.000000	0.000000
25%	0.000000	2.000000	NaN	0.000000	0.000000	7.917700
50%	0.000000	3.000000	NaN	0.		
75%	1.000000	3.000000	NaN	1.		
max	1.000000	3.000000	80.000000	8.		

	sex	embarked	class	who	deck	embark_town	alive	name
count	892	890	892	892	204	890	892	892
unique	4	3	3	3	7	7	2	891
top	male	S	Third	man	C	Southampton	no	Behr, Mr. Karl Howell
freq	574	644	491	538	60	643	550	2

MONASH University

Data Auditing

Check correlation amongst variables: `df.corr()`

	survived	pclass	age	sibsp	parch	fare	adult_male	alone
survived	1.000000	-0.339932	-0.079229	-0.035960	0.080874	0.257012	-0.555222	-0.204758
pclass	-0.339932	1.000000	-0.367110	0.083789	0.019245	-0.548667	0.092448	0.137065
age	-0.079229	-0.367110	1.000000	-0.301632	-0.185857	0.096148	0.278219	0.196753
sibsp	-0.035960	0.083789	-0.301632	1.000000	0.414985	0.159654	-0.253892	-0.583247
parch	0.080874	0.019245	-0.185857	0.414985	1.000000	0.216221	-0.350200	-0.582176
fare	0.257012	-0.548667	0.096148	0.159654	0.216221	1.000000	-0.181997	-0.271540
adult_male	-0.555222	0.092448	0.278219	-0.253892	-0.350200	-0.181997	1.000000	0.403131
alone	-0.204758	0.137065	0.196753	-0.583247	-0.582176	-0.271540	0.403131	1.000000

Examples

MONASH University

[1] Misspelling and Inconsistency

.....	**Suburbs**
.....	burwood.
.....	springvale
....	Burwood
....	Springvae
....	East Melbourne
.....	E. Melbourne
.....	.....

- *What are the problems?*
- *How can you detect these data problems?*
- *How can you resolve them?*

[1] Misspelling and Inconsistency

- Inconsistency
  - common cases:
    - upper vs. lower case
    - inconsistency in domain value representation, e.g., 0 vs. No, 1 vs. Yes
  - detecting and fixing
    - investigate unique domain values (`unique()`)
    - make the representation consistent, e.g., replace
- Misspelling
  - investigate unique domain values (`unique()`)
  - string matching
    - calculate domain value frequencies (`value_counts()`)
    - for all values, find matches for the infrequent values
    - replace infrequent values with the best match (if it exists) from the more frequent values.

[2] Irregularities

.....	**Entry Time**
.....	12/13/2010
.....	1/1/2014
....	45/2/2010
....	20/3/2011
....	2/14/2014
.....	25/12/2014
.....	.....

- *What are the problems?*
- *How can you detect these data problems?*
- *How can you resolve them?*

[2] Irregularities

- Common cases:
  - invalid dates
  - domain dependent value, value not valid for a specific domain, e.g., negative value for number of passengers
- Detecting
  - investigate unique domain values (`unique()`)
  - investigate value ranges for the column
  - type casting, e.g., parse date string to datetime object, catch exceptions when it is not a valid date format
- Fixing
  - refer to documentation if it exists, to see whether these values have special meaning
  - replace
  - remove

[3] Integrity Constraint Violation

….	year_built	time_settled
…..	2010	12/13/2010
…..	2010	1/1/2014
….	2010	45/2/2002
….	2010	20/3/2011
….	2021	2/14/2014
…..	…..	…..

- *What are the problems?*
- *How can you detect these data problems?*
- *How can you resolve them?*

[3] Integrity Constraint Violation

- Common case:
  - highly dependent on context, e.g.,
    - sold date vs. advertised date,
    - one field is the sum of the other two,
    - land size must be larger than building size, etc.
- Detecting
  - highly dependent on the domain and problems
- Fixing
  - swap
  - remove, etc.

[4] Duplicates

Index	Name	Gender	D.O.B	Mobile	Address	.....
10	John	N/A	7/9/1985	0412685210	N/A	.....
145	John Walter	M	7/9/1985	0412685210	2 Yale st., burwood	.....
200	John Walter	Male	7/9/1985	0412685210	2 Yale street, burwood	....
268	Walter, John	Male	7/9/1985	0412685210	2 Yale street, burwood	....
450	John Walter	-	1985	-	2 Yale street, burwood	.....
.....	.....		.....			.....

- *What are the duplications in the example and how to detect and resolve these data duplications?*

[4] Duplicates

- Common cases:
    - complete duplication
    - duplicate due to field missing
        - different record have different piece of info
- Detecting
    - identifying keys to check duplicates
        - fixing other quality issues
        - try different keys
- Fixing
    - combine information/merge
    - remove duplicates

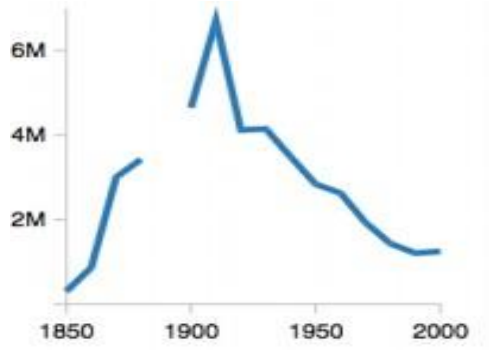
[5] Missing Values

Adv. Price	Bedrooms	Land (sqm)	Condition	Suburb	Source
800000	2	-	Old	CBD	realestate
80	2	250	*	CBD	domain
1100000	-	-	Fair	Burwood	realestate
*	3	800	*	Dandenong	domain
Contact Agent	-	500	Fair	Burwood	realestate
…..		…..		…..	…..

- *How to detect these missing value records and how to fix them?*

[5] Missing Values

- Detecting
  - investigate unique domain values (`unique()`)
  - investigate value range, cautious about extremely small and large values
  - domain analysis
- Fixing
  - imputation
    - mean and mode
    - regression (find variables that are closely related (e.g., `df.corr()`)
    - dummy value
  - removal
  - all depends on the situation and needs justification

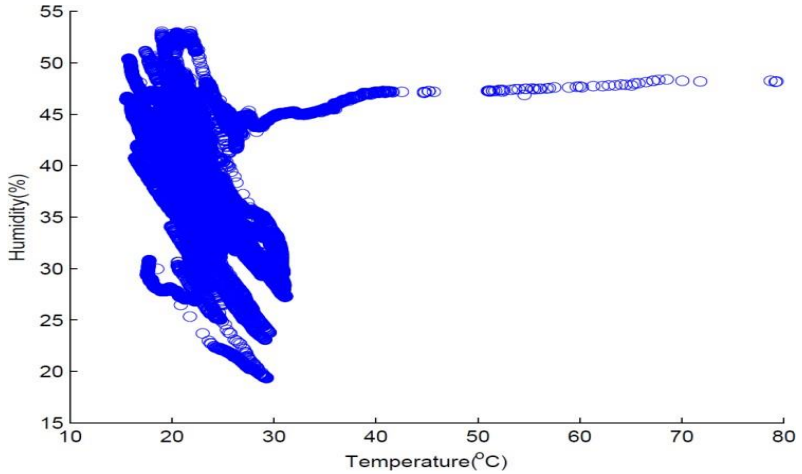


[6] Outliers

Adv. Price	Bedrooms	Land (sqm)	Condition	Suburb	Source
800000	2	-	Old	CBD	realestate
80	2	250	Fair	CBD	domain
1100000	-	550	Fair	Burwood	realestate
500000	8	800	New	Dandenong	domain
500000	-	500	Fair	Burwood	realestate
…..		…..		…..	…..

- *How to detect these outliers? Can we directly remove outliers once found from tools, e.g., boxplot?*

[6] Outliers



- Common cases: numerical field
- Challenge:
  - not easy to find
- Detecting
  - range of values `df.describe()`
  - Graphical tools, e.g., boxplot (default using a IQR rule)
  - 3σ edit rule
  - Good to do some comparison between results found by different identifiers
- Fixing
  - Similar to handling missing values

Recap: Learning Outcomes

Week 4

By the end of this week you should be able to:

- Explain open data and linked open data
- Explain how to access to new data sources through APIs
- Identify how different APIs work
- Inspect data quality problems in datasets and recommend solutions to fix them
- *Use data wrangling operations in Python (in Laboratory class)*