FIT1043 Lecture 4
Introduction to Data Science

Mahsa Salehi

Faculty of Information Technology, Monash University

Semester 2, 2022

Student Feedback Survey

- Hope you enjoyed the unit so far!
- Spend a few mins now to fill in these two anonymous surveys
  - Lecture survey
  - Tutorial survey

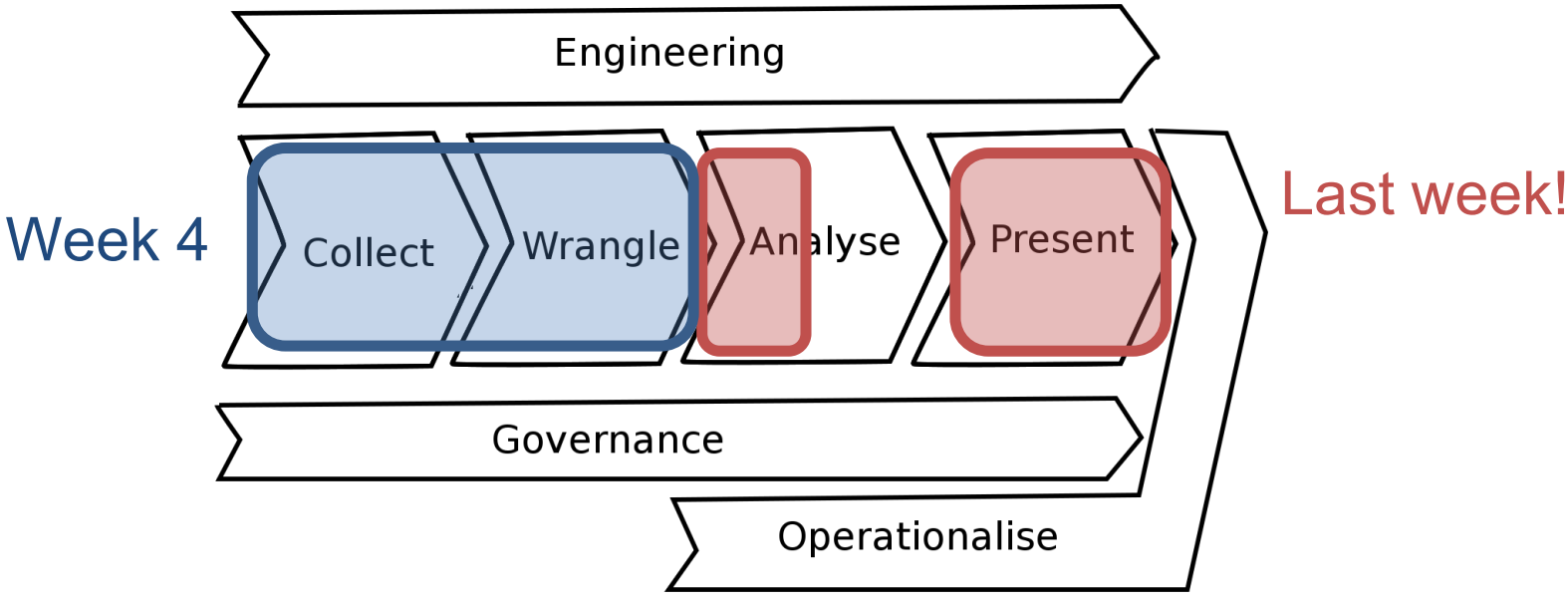
Assignment 1

- Due 29 August 11:55pm
- Dataset: Monthly smartcard replacements dataset in Queensland
- Any questions:
  - Post to forum (Ed discussion)
  - Email: fit1043.clayton-x@monash.edu
  - Email your tutors
  - Attend consultations: click here for the times and locations. We have additional zoom consultations Weeks 4 and 5 for assignment 1.

Unit Schedule

Week	Activities	Assignments
1	Overview of data science	Weekly Lecture/tutorial active participation assessment
2	Introduction to Python for data science	
3	Data visualisation and descriptive statistics	
4	Data sources and data wrangling	
5	Data analysis theory	Assignment 1
6	Regression analysis	
7	Classification and clustering	
8	Introduction to R for data science	
9	Characterising data and "big" data	Assignment 2
10	Big data processing	
11	Issues in data management	
12	Industry guest lecture	Assignment 3

Our Standard Value Chain



Outline

- Data resources
  - Open data
  - API
- Data Wrangling
  - Motivation
  - Data quality problems
  - Data auditing in Python
  - Techniques to handle the data quality problems

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

Introduction to Resources

- Where to find and how to use data sources
- Open data

  Machine readable and publicly available
- Data Wrangling

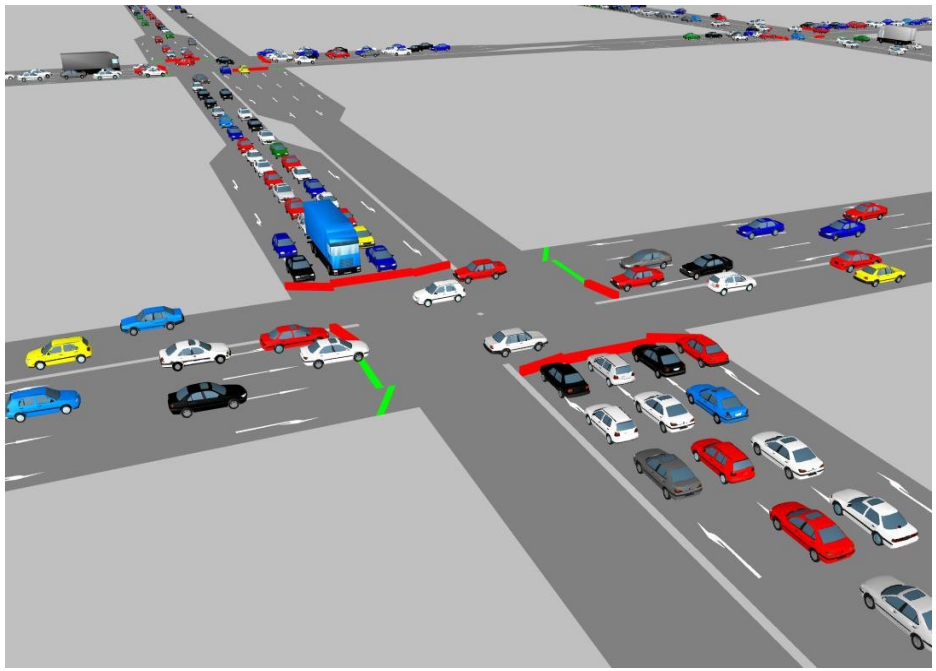
  Data manipulation and preparation for data analysis

Introduction to Resources:
Finding and using data

access to new data sources or clever and creative use of existing multiple data sources are very important in a data science project

Where to find and how to use data sources

Task: forecasting traffic: blockages, clearing, surprising situations, alternate routes



- Critical data:
  - GPS data on traffic flow
  - Maps
  - incidents and events
  - weather
- Challenge:
  - collect different sources of data

image src: math.tu-berlin.de

FLUX Question

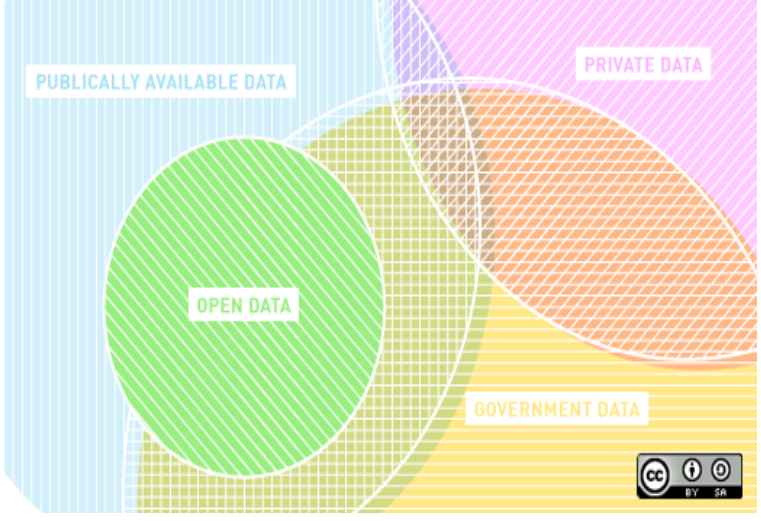
Consider data sources in ”traffic forecasting” task, name a website to access to one of the data sources.

Introduction to Resources:
Open data

organizations provide machine readable to support data science

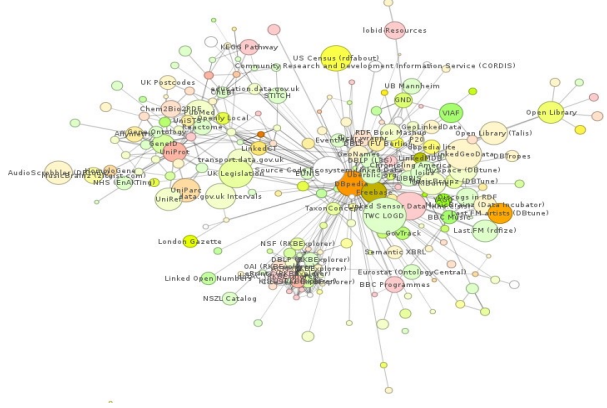
Open Data

- Publicly available
  - government and IT departments building data and infrastructure to allow sharing
  - e.g., Data.GOV has 230k datasets, and Data.GOV.AU has 30k
- Machine readable
- But..
  - it is not always usable
  - people need the right skills



Open Data..

- A common format for open data is "Linked Open Data (LOD)"
  - Triples: subject, verb and object
  - Example: DBpedia – a dataset containing extracted data from Wikipedia; it contains about 3.4 million concepts described by 1 billion triples. e.g., DBPedia page for "Albert Einstein"
  - Enables data from different sources to be connected and queried.



API

API: **A**pplication **P**rogrammer **I**nterface

Routines providing programatic access to an application.

- API is like a user interface but it is designed for computers to access to the functionality of a software (google maps)

- Computers talk to each other

- API consumers vs API providers

FLUX Question

Name a popular data/information API.

Example APIs

Many companies are exposing their data **and their website functionality** as APIs for others to make use of:

- *Facebook API*
- *Twitter API*
- *LinkedIn API*
- *Google Maps API*
- *Youtube API*
- *Amazon Advertising API*
- *TripAdvisor API*

For list of APIs check here

Twitter

Twitter is the most famous microblogging platform

- with big corporate use
- contains lots of metadata: information about users, their follower network, locations, hashtags, emojis+emoticons, ...

Sample Twitter XML Data

```
<?xml version="1.0" encoding="UTF-8" ?>
- <statuses type="array">
  - <status>
      <created_at>Wed Jun 10 00:57:28 +0000 2009</created_at>
      <id>2097065233</id>
      <text>sitting in vegas @ airport, kid in stroller, with dvd player in lap. First ever for me. HELLO!</text>
      <source>web</source>
      <truncated>false</truncated>
      <in_reply_to_status_id />
      <in_reply_to_user_id />
      <favorited>false</favorited>
      <in_reply_to_screen_name />
    - <user>
        <id>5189091</id>
        <name>kristin bednarz</name>
        <screen_name>kristinbednarz</screen_name>
        <location>iPhone: 33.447393,-101.821675</location>
        <description>photographer in WEST TEXAS</description>
        <profile_image_url>http://s3.amazonaws.com/twitter_production/profile_images/80432676/BIO_norr
        <url>http://www.yourlifemypassion.com</url>
        <protected>false</protected>
        <followers_count>245</followers_count>
        <profile_background_color>352726</profile_background_color>
        <profile_text_color>3E4415</profile_text_color>
        <profile_link_color>D02B55</profile_link_color>
        <profile_sidebar_fill_color>99CC33</profile_sidebar_fill_color>
        <profile_sidebar_border_color>829D5E</profile_sidebar_border_color>
        <friends_count>90</friends_count>
        <created_at>Thu Apr 19 04:54:45 +0000 2007</created_at>
        <favourites_count>3</favourites_count>
        <utc_offset>-21600</utc_offset>
        <time_zone>Central Time (US & Canada)</time_zone>
```

Twitter Developer API

See *Twitter's developer platform*

- library interfaces for Java, C++, Javascript, Python, Perl, PHP, ...
- allows other applications to manage Twitter data for users
- extensive developer policy

Introduction to Resources:
Data Wrangling

manipulating data to make it directly usable for analysis

What do analysts wish the data looks like?

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1	39	State-gov	77516	Bachelors	13	Never-married	Adm-clerical	Not-in-famil	White	Male	2174	0	40	United-States	<=50K
2	50	Self-emp-no	83311	Bachelors	13	Married-civ-spouse	Exec-managerial	Husband	White	Male	0	0	13	United-States	<=50K
3	38	Private	215646	HS-grad	9	Divorced	Handlers-cleaners	Not-in-famil	White	Male	0	0	40	United-States	<=50K
4	53	Private	234721	11th	7	Married-civ-spouse	Handlers-cleaners	Husband	Black	Male	0	0	40	United-States	<=50K
5	28	Private	338409	Bachelors	13	Married-civ-spouse	Prof-specialty	Wife	Black	Female	0	0	40	Cuba	<=50K
6	37	Private	284582	Masters	14	Married-civ-spouse	Exec-managerial	Wife	White	Female	0	0	40	United-States	<=50K
7	49	Private	160187	9th	5	Married-spouse-abser	Other-service	Not-in-famil	Black	Female	0	0	16	Jamaica	<=50K
8	52	Self-emp-no	209642	HS-grad	9	Married-civ-spouse	Exec-managerial	Husband	White	Male	0	0	45	United-States	>50K
9	31	Private	45781	Masters	14	Never-married	Prof-specialty	Not-in-famil	White	Female	14084	0	50	United-States	>50K
10	42	Private	159449	Bachelors	13	Married-civ-spouse	Exec-managerial	Husband	White	Male	5178	0	40	United-States	>50K
11	37	Private	280464	Some-colleg	10	Married-civ-spouse	Exec-managerial	Husband	Black	Male	0	0	80	United-States	>50K
12	30	State-gov	141297	Bachelors	13	Married-civ-spouse	Prof-specialty	Husband	Asian-Pac-Isl	Male	0	0	40	India	>50K
13	23	Private	122272	Bachelors	13	Never-married	Adm-clerical	Own-child	White	Female	0	0	30	United-States	<=50K
14	32	Private	205019	Assoc-acdm	12	Never-married	Sales	Not-in-famil	Black	Male	0	0	50	United-States	<=50K
15	40	Private	121772	Assoc-voc	11	Married-civ-spouse	Craft-repair	Husband	Asian-Pac-Isl	Male	0	0	40	?	>50K
16	34	Private	245487	7th-8th	4	Married-civ-spouse	Transport-moving	Husband	Amer-Indian	Male	0	0	45	Mexico	<=50K

What does data really look like?

CTCHEHI
CT Chest Hi Resolution 30/11/04 at 2156 CT-04-014735
REPORT:
Clinical note: transformed AML. Ongoing fevers.? Source. ? fungal infection.
Report:
Axial 1.25 mm slices at 10 mm intervals taken in inspiration with selected images in the prone position.
No mediastinal or hilar lymphadenopathy. Heart size is normal. Borderline enlargement of the main pulmonary outflow tract. There is smooth interlobular septal thickening throughout both lungs, which may be secondary to fluid overload. There is a background of emphysematous changes , predominantly in 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 lack of intravenous contrast.
Conclusion:
Single nodule in right upper lobe has a non-specific appearance but given the 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
Performed by: Contributor_system, P on 11 January 2005 12:21

Why Wrangling?

- 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 data

Our goal

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

Data + Wrangling + Analysis = Data Product

What is Data Wrangling?

Process of transforming “raw” data into data that can be analyzed to generate valid actionable results and insights

Data Wrangling

- Data pre-processing
- Data preparation
- Data cleansing
- Data transformation
- etc

Sources of Data Quality Issues

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

Data quality problems: Interpretability issue

- 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 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 your data set 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
```

Data quality problems: Data format issue

- 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)

Data quality problems: Inconsistent and faulty data

- mistyped data
- inconsistent entry
- extraneous (irrelevant or unrelated) 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

Data quality problems: Missing values

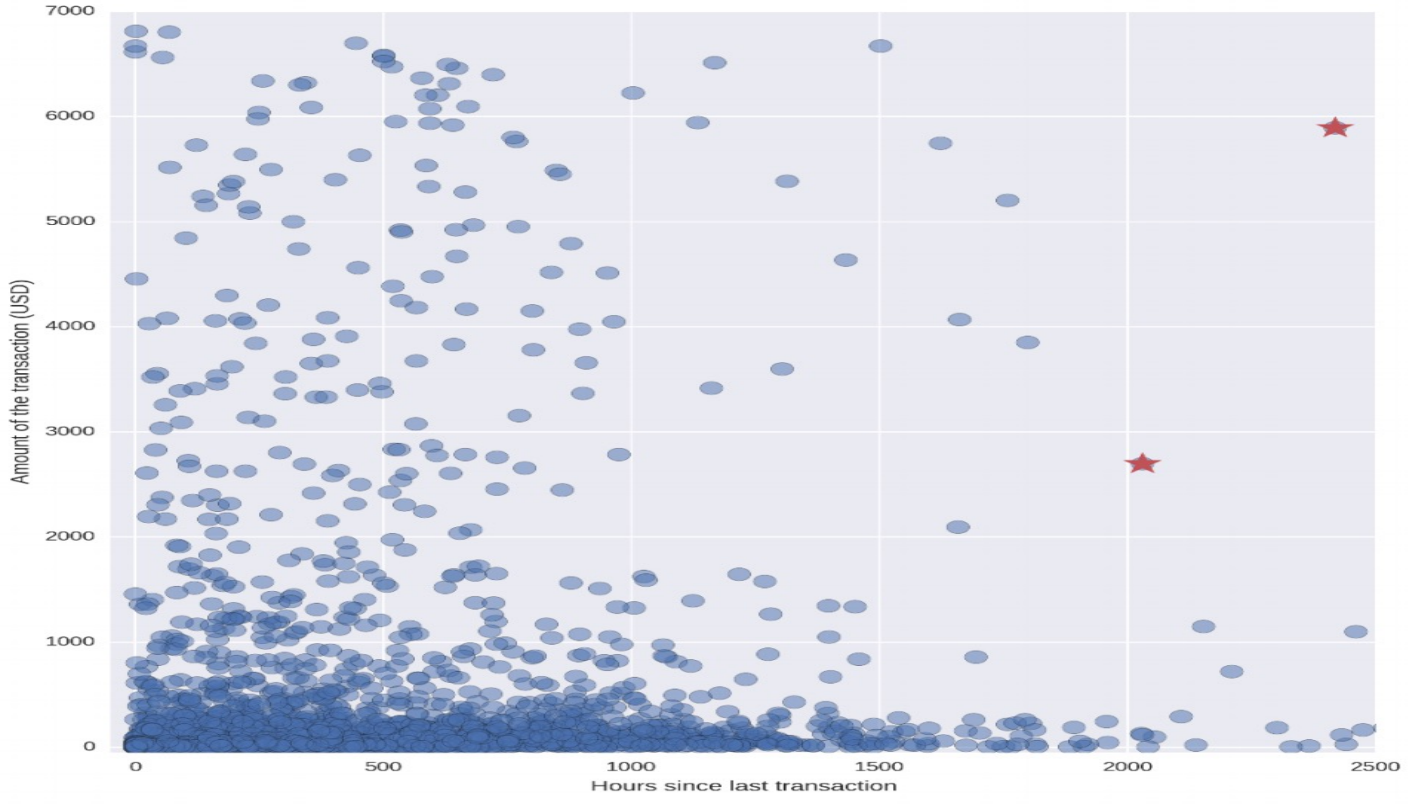
- Data values that should be presented in a dataset but that are absent for many reasons.

Figure: Missing values in the Switzerland heart disease data set are indicated by “?”.

```
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
```

Data quality problems: outliers

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



Data quality problems: Duplicates

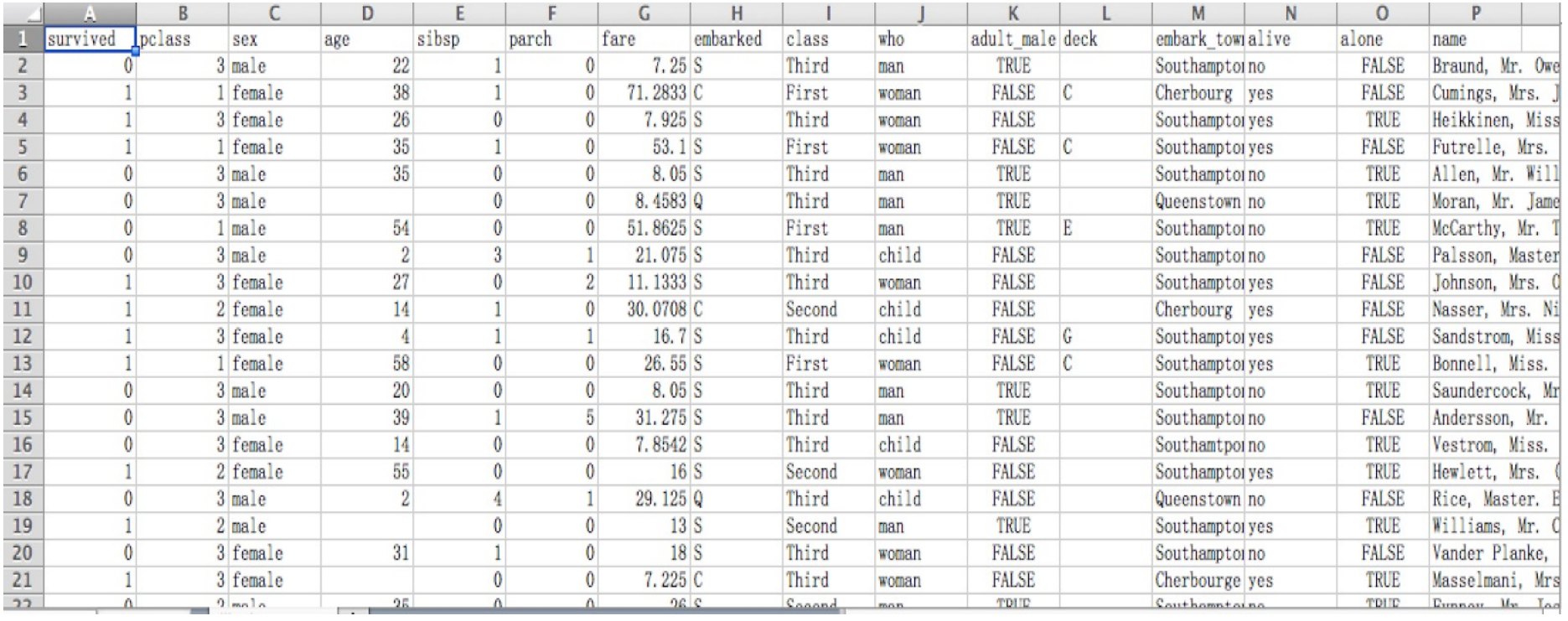
- Multiple data entries that correspond to the same piece of information.

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

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

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. C
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

Data Auditing Tools

- Some general steps to perform data auditing (assuming you're given a dataframe `df`)
- dimension: `df.shape()`
  - number of rows
  - number of columns

```
(8500, 10)
```

Data Auditing Tools

- 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 Tools

- Check basic information about the dataframe - number of records, whether there are null values, datatype: `df.info()`

- What are the numerical and categorical columns?

```
<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 Tools

- Check some basic statistics about columns:
  - numerical columns: `df.describe()`
  - object columns: `df.describe(include = ['O'])`

	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	0.000000	0.000000	14.454200
75%	1.000000	3.000000	NaN	1.000000	0.000000	31.000000
max	1.000000	3.000000	80.000000	8.000000	6.000000	512.329200

	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

Data Auditing Tools

- 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

Remember *Pearson correlation* from last week

[Discussion 1] Misspelling and inconsistency

- **Discuss about:**

What are the problems?

How can you detect these data problems?

How can you resolve them?

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

FLUX Question

How many problems can you identify?

[Discussion 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.

[Discussion 2] Irregularities

- **Discuss about:**

What are the problems?

How can you detect these data problems?

How can you resolve them?

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

[Discussion 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 (`pandas.to_datetime`)
- Fixing
  - refer to documentation if it exists, to see whether these values have special meaning
  - replace
  - remove

[Discussion 3] Integrity Constraint Violation

- **Discuss about:**

What are the problems?

How can you detect these data problems?

How can you resolve them?

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

[Discussion 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.

[Discussion 4] Duplications

- **Discuss about:**

what are the duplications in the example and how to detect and resolve these data duplications?

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	.....
.....	.....		.....			.....

[Discussion 4] Duplications

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

[Discussion 5] Missing values

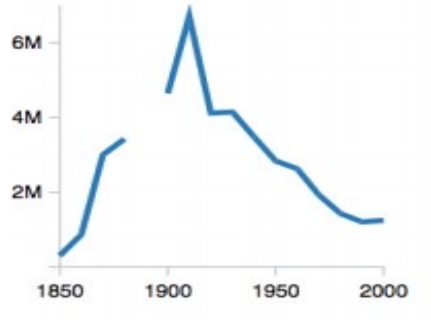
- **Discuss about:**

how to detect these missing value records and how to fix them?

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
…..		…..		…..	…..

[Discussion 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



[Discussion 6] Outliers

- **Discuss about:**

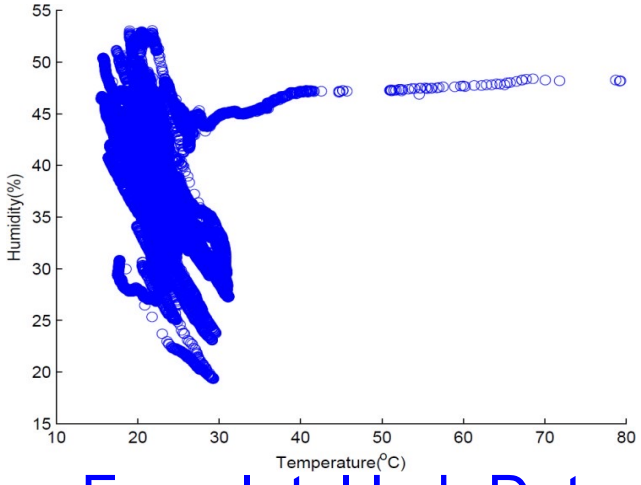
how to detect these outliers? Can we directly remove outliers once found from tools, e.g., boxplot?

Sold Price	Bedrooms	Land (sqm)	Condition	Suburb	……
800000	2	-	Old	CBD	…..
80	2	250	Fair	CBD	…..
1100000	-	550	Fair	Burwood	…..
500000	Male	800	New	Dandenong	….
500000	-	500	Fair	Burwood	…..
…..		…..		…..	…..

[Discussion 6] Outliers

- Common cases: numerical field

- Challenge:
  - not easy to find



From Intel Lab Data

- 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

FLUX Question

How to deal with missing data?

A. Removing the row or column
B. Replace with a special “unknown” value
C. Replace with an average value

Tutorial/Lab week 4

1. Data Wrangling with Python

2. Data Wrangler is a Stanford project for data cleaning and preprocessing

Readings: Data Wrangling Examples

"How we found the worst place to park in New York City" has examples, and a discussion of the complexities of getting data out of New York City:

Danger spots for cycles: *NYPD crash data* obtained by daily download of PDF files followed by (non-trivial) extraction
NB. they now have Excel data to ease the work!

Dirty waterways: *fecal coliform measurements on waterways* from Department of Environmental Protection's website; extracted from Excel sheets per site; each in a different format

Faulty road markings: parking tickets for fire-hydrants by location from *NYC Open Data portal* need to normalize the addresses supplied