Introduction to Data Wrangling

Faculty of Information Technology
Monash University

FIT5196 week 1

Outline

1. Motivations
2. Introduction to FIT5196 Data Wrangling
   - Unit structure
   - Assessments
   - Unit management
3. Introduction to Data Wrangling
   - Data Quality Problems
   - Characteristics of Tidy Data
   - Major Tasks in Data Wrangling
   - Programming Environment
4. Demonstration: Wrangling Air Crashes data with Data Wrangler
5. Summary

What is Data Wrangling?

[1] http://vis.stanford.edu/files/2011-DataWrangling-IVJ.pdf

What is Data Wrangling?

- From Trifacta's Data Wrangling practitioners,[1]

> "We define such data wrangling as a process of iterative data exploration and transformation that enables analysis. One goal is to make data usable — to put them in a form that can be parsed and manipulated by analysis tools. . . . In other words, data wrangling is the process of making data useful. Ideally, the outcome of wrangling is not simply data; it is an editable and auditable transcript of transformations coupled with a nuanced understanding of data organization and data quality issues.

[1] http://vis.stanford.edu/files/2011-DataWrangling-IVJ.pdf

What is Data Wrangling?

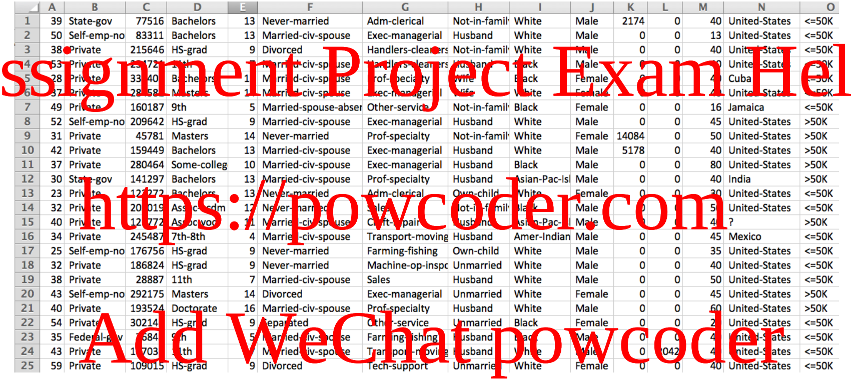
- From Trifacta's Data Wrangling practitioners,[1]

> "We define such data wrangling as a process of iterative data exploration and transformation that enables analysis. One goal is to make data usable — to put them in a form that can be parsed and manipulated by analysis tools. . . . In other words, data wrangling is the process of making data useful. Ideally, the outcome of wrangling is not simply data; it is an editable and auditable transcript of transformations coupled with a nuanced understanding of data organization and data quality issues.

- Two key points
  - Clean and useful data that can be used in the downstream data analysis.
  - Documentation of all data manipulation performed.

[1] http://vis.stanford.edu/files/2011-DataWrangling-IVJ.pdf

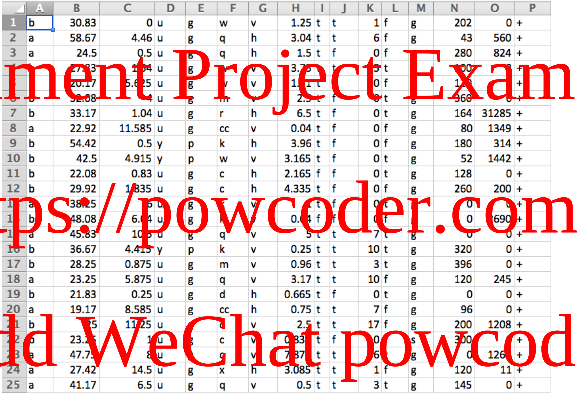
What do analysts wish the data looked like?



	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
17	25	Self-emp-no	176756	HS-grad	9	Never-married	Farming-fishing	Own-child	White	Male	0	0	35	United-States	<=50K
18	32	Private	186824	HS-grad	9	Never-married	Machine-op-inspc	Unmarried	White	Male	0	0	40	United-States	<=50K
19	38	Private	28887	11th	7	Married-civ-spouse	Sales	Husband	White	Male	0	0	50	United-States	<=50K
20	43	Self-emp-no	292175	Masters	14	Divorced	Exec-managerial	Unmarried	White	Female	0	0	45	United-States	>50K
21	40	Private	193524	Doctorate	16	Married-civ-spouse	Prof-specialty	Husband	White	Male	0	0	60	United-States	>50K
22	54	Private	302146	HS-grad	9	Separated	Other-service	Unmarried	Black	Female	0	0	20	United-States	<=50K
23	35	Federal-gov	76845	9th	5	Married-civ-spouse	Farming-fishing	Husband	Black	Male	0	0	40	United-States	<=50K
24	43	Private	117037	11th	7	Married-civ-spouse	Transport-moving	Husband	White	Male	0	2042	40	United-States	<=50K
25	59	Private	109015	HS-grad	9	Divorced	Tech-support	Unmarried	White	Female	0	0	40	United-States	<=50K

- The "census income" data set from UCI machine learning data repository.
- Data analysis task: Predict whether income exceeds $50K/yr based on age, education, marital status, native-country, etc.
- Algorithms: C4.5 (Decision Tree), Naive-Bayes, Nearest Neighbours, etc.

What do analysts wish the data looked like?

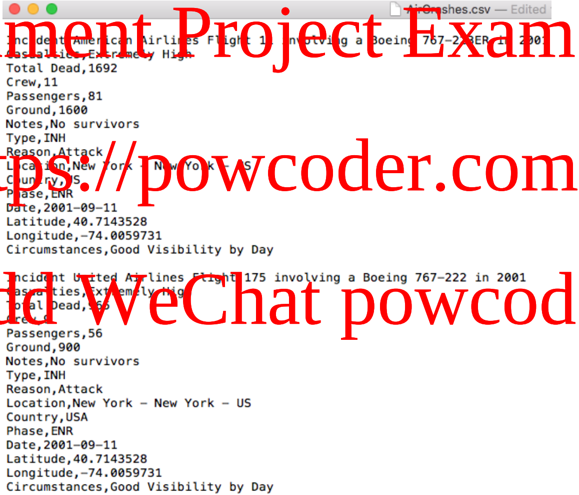


	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
1	b	30.83	0	u	g	w	v	1.25	t	t	1	f	g	202	0	+
2	a	58.67	4.46	u	g	q	h	3.04	t	t	6	f	g	43	560	+
3	a	24.5	0.5	u	g	q	h	1.5	t	f	0	f	g	280	824	+
4	b	27.83	1.54	u	g	w	v	3.75	t	t	5	t	g	100	3	+
5	b	20.17	5.625	u	g	w	v	1.71	t	f	0	f	s	120	0	+
6	b	32.08	4	u	g	m	v	2.5	t	f	0	t	g	360	0	+
7	b	33.17	1.04	u	g	r	h	6.5	t	f	0	t	g	164	31285	+
8	a	22.92	11.585	u	g	cc	v	0.04	t	f	0	f	g	80	1349	+
9	b	54.42	0.5	y	p	k	h	3.96	t	f	0	f	g	180	314	+
10	b	42.5	4.915	y	p	w	v	3.165	t	f	0	t	g	52	1442	+
11	b	22.08	0.83	u	g	c	h	2.165	f	f	0	t	g	128	0	+
12	b	29.92	1.835	u	g	c	h	4.335	t	f	0	f	g	260	200	+
13	a	38.25	6	u	g	k	v	1	t	f	0	t	g	0	0	+
14	b	48.08	6.04	u	g	k	v	0.04	f	f	0	f	g	0	2690	+
15	a	45.83	10.5	u	g	q	v	5	t	t	7	t	g	0	0	+
16	b	36.67	4.415	y	p	k	v	0.25	t	t	10	t	g	320	0	+
17	b	28.25	0.875	u	g	m	v	0.96	t	t	3	t	g	396	0	+
18	a	23.25	5.875	u	g	q	v	3.17	t	t	10	f	g	120	245	+
19	b	21.83	0.25	u	g	d	h	0.665	t	f	0	t	g	0	0	+
20	a	19.17	8.585	u	g	cc	h	0.75	t	t	7	f	g	96	0	+
21	b	25	11.25	u	g	c	v	2.5	t	t	17	f	g	200	1208	+
22	b	23.25	1	u	g	c	v	0.835	t	f	0	f	s	300	0	+
23	a	47.75	8	u	g	c	v	7.875	t	t	6	t	g	0	1260	+
24	a	27.42	14.5	u	g	x	h	3.085	t	t	1	f	g	120	11	+
25	a	41.17	6.5	u	g	q	v	0.5	t	t	3	t	g	145	0	+

- The "credit approval" data set from UCI machine learning data repository.
- Data analysis: Predict whether or not a credit card application should be approved.
- Algorithms: Decision Tree.

What does data really look like?

Airline Crash dataset from Wikipedia:

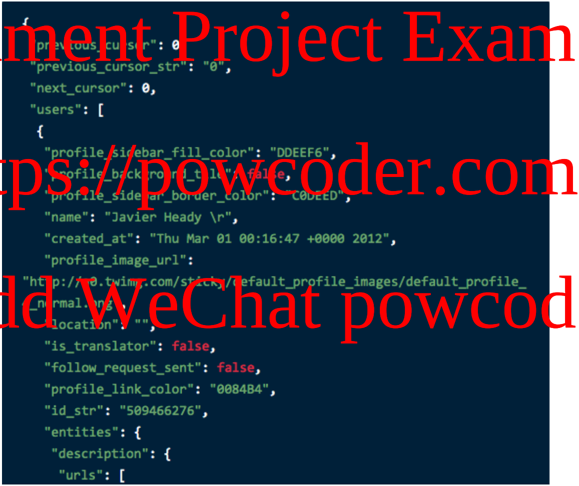


```
Incident American Airlines Flight 11 involving a Boeing 767-223ER in 2001
Casualties,Extremely High
Total Dead,1692
Crew,11
Passengers,81
Ground,1600
Notes,No survivors
Type,INH
Reason,Attack
Location,New York - New York - US
Country,USA
Phase,ENR
Date,2001-09-11
Latitude,40.7143528
Longitude,-74.0059731
Circumstances,Good Visibility by Day

Incident United Airlines Flight 175 involving a Boeing 767-222 in 2001
Casualties,Extremely High
Total Dead,965
Crew,9
Passengers,56
Ground,900
Notes,No survivors
Type,INH
Reason,Attack
Location,New York - New York - US
Country,USA
Phase,ENR
Date,2001-09-11
Latitude,40.7143528
Longitude,-74.0059731
Circumstances,Good Visibility by Day
```

What does data really look like?

Twitter data [2]

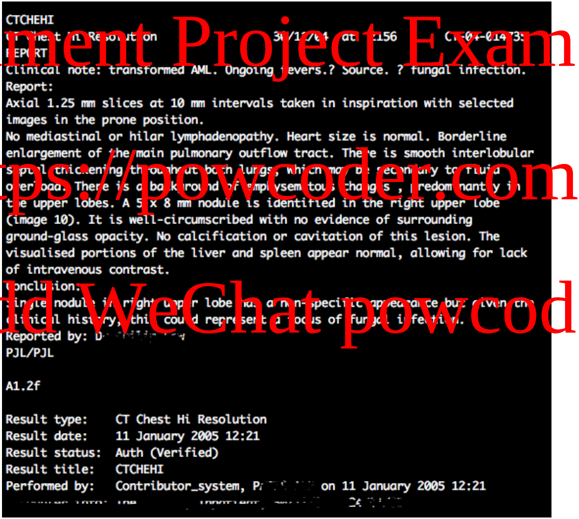


```
{
  "previous_cursor": 0,
  "previous_cursor_str": "0",
  "next_cursor": 0,
  "users": [
   {
    "profile_sidebar_fill_color": "DDEEF6",
    "profile_background_tile": false,
    "profile_sidebar_border_color": "C0DEED",
    "name": "Javier Heady \r",
    "created_at": "Thu Mar 01 00:16:47 +0000 2012",
    "profile_image_url":
"http://a0.twimg.com/sticky/default_profile_images/default_profile_
_normal.png",
    "location": "",
    "is_translator": false,
    "follow_request_sent": false,
    "profile_link_color": "0084B4",
    "id_str": "509466276",
    "entities": {
     "description": {
      "urls": [
```

[2] https://dev.twitter.com/rest/reference/get/blocks/list

What does data really look like?

Fungal disease CT report.



```
CTCHEHI
CT Chest Hi Resolution                30/12/04  at 1156    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[redacted]
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[redacted] on 11 January 2005 12:21
[redacted]
```

Our goal

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

Data + Wrangling + Analysis = Data Product (or Knowledge)

Outline

MONASH University

1. Motivations

2. Introduction to FIT5196 Data Wrangling
   - Unit structure
   - Assessments
   - Unit management

3. Introduction to Data Wrangling

4. Demonstration: Wrangling Air Crashes data with Data Wrangler

5. Summary

Unit Objectives

- What the course is trying to achieve:
  - parse data in the required format;
  - assess the quality of data for problem identification;
  - resolve data quality issues ready for the data analysis process;
  - integrate data sources for data enrichment;
  - document the wrangling process for professional reporting;
  - write program scripts for data wrangling processes
- What it is not trying to achieve:
  - Introduction to Python programming, e.g., how to program in Python.

You MUST be very familiar with Python and the usage of Python packages!

Unit outline

Week	topic
1	Introduction to Data Wrangling
2	Introduction to Regular Expressions
3	Parsing Raw Data in Different Formats
4	Text Data Preprocessing
5	Text Data Preprocessing
6	Data Cleansing
7	Data Cleansing
8	Data Cleansing
9	Data Integration
10	Data Integration and reshaping
11	Data Enrichment, Transformation, normalization, etc.
12	Summary

Assessments

- Summary

	Value	Due Date	Type
Assessment 1	35%	Week 6 Sunday, 2 September 2018	Individual
Assessment 2	35%	Week 10, Wednesday 3 October 2018	Individual
Assessment 3	30%	Week 12, Sunday 21 October 2018	Individual

- General criteria for marking
  - **The submitted code must work without any errors and must give the correct results.**
  - The code should be well structured and properly commented.
  - The notebook should be structured in a logical way so that it clearly shows how students finish the tasks in the assessment.

Assessments: 1

- Assessment 1: Parsing Data + Text Preprocessing
- Brief description: Data extracted from different sources is often stored in different formats. In this assessment, you are required to write Python (either Python 2 or Python 3) script to
  1. extract data from an XML, HTML, and PFD files,
  2. convert data stored in an XML and HTML files to a JSON,
  3. and generate sparse representation for the pdf files file.
- Due date: Week 6 Sunday, 2 September 2018

Assessment details: 2

- Assessment 2: Parsing and Cleansing Raw Data
- Brief description: This assessment addresses one the most important steps in data wrangling, i.e., cleansing data. Students are required to
  - inspect, audit and then identify problems existing in the parsed data; and propose appropriate methods to fix these problems;
  - Different generic and major data problems could be found in the data might include:
    - Lexical errors, e.g., typos and spelling mistakes
    - Irregularities, e.g., abnormal data values and data formats
    - Violations of the Integrity constraint.
    - Outliers
    - Duplications
    - Missing values
    - Inconsistency, e.g., inhomogeneity in values and types in representing the same data
- Hurdle: The outcome of the interview, where students will need to communicate their processes, justify their approaches, and answer some questions.
- Due date: Week 10, Wednesday 3 October 2018

Assessment details: 3

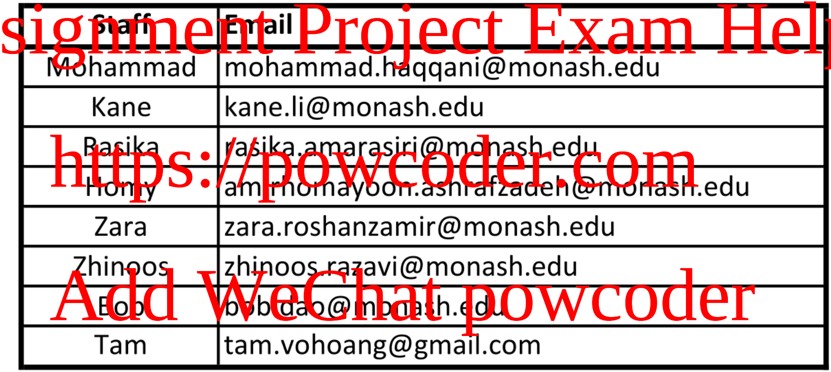
- Assessment task 3: Data Integration and Reshaping
- Brief description: This assessment focuses on data integration and reshaping. The students are required to integrate data that might be collected from different sources. The students need to resolve different levels of conflicts in integration according to what we will be discussed in the lectures. The output of this assessment should be an integrated dataset and a Jupyter Notebook containing information about your designed global schema and all the Python scripts used in integrating the data. Moreover, the students will need to apply various normalization/transformation methods to the data, and analyze how they affect the distribution of the data.
- Due date: Week 12, Sunday 21 October 2018

Unit management: CE & Lecturers

- CE: Dr. Lan Du
  - Contact: Lan.Du@monash.edu
  - Office: R142, Building 63 - 25 Exhibition Walk, Clayton Campus
- Lecturer: Mr. Mohammad Haqqani
  - Contact: Mohammad.Haqqani@monash.edu
  - Office: TBA
- Lecturers' consultation by appointment

Unit management: Tutors

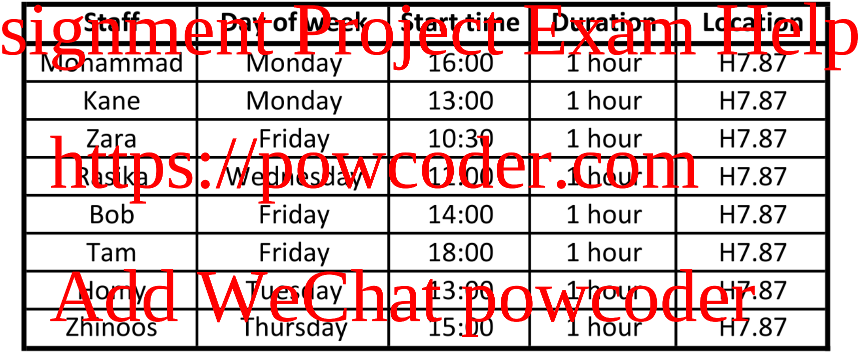
- Tutors



Staff	Email
Mohammad	mohammad.haqqani@monash.edu
Kane	kane.li@monash.edu
Rasika	rasika.amarasiri@monash.edu
Homy	amirhomayoon.ashrafzadeh@monash.edu
Zara	zara.roshanzamir@monash.edu
Zhinoos	zhinoos.razavi@monash.edu
Bob	bob.dao@monash.edu
Tam	tam.vohoang@gmail.com

Unit management: Consultations

- Consultation time



Staff	Day of week	Start time	Duration	Location
Mohammad	Monday	16:00	1 hour	H7.87
Kane	Monday	13:00	1 hour	H7.87
Zara	Friday	10:30	1 hour	H7.87
Rasika	Wednesday	11:00	1 hour	H7.87
Bob	Friday	14:00	1 hour	H7.87
Tam	Friday	18:00	1 hour	H7.87
Henry	Tuesday	13:00	1 hour	H7.87
Zhinoos	Thursday	15:00	1 hour	H7.87

Unit management: Lectures & Tutorial Classes

	Day of week	Start time	Weeks	Duration	Location	Staff
Lecture	Thu	12:00	26/7-20/9, 4/10-18/10	2 hours	CA_B/B214	Mohammad Haqqani
	Mon	18:00	23/7-17/9, 1/10-15/10	2 hours	CA_B/B215	Mohammad Haqqani
Laboratory	Tue	18:00	24/7-18/9, 2/10-16/10	2 hours	CA_B/B348B	Hoang Tam Vo
	Fri	18:00	27/7-21/9, 5/10-19/10	2 hours	CA_B/B342	Zhinoos Razavi Hesabi
	Thu	20:00	26/7-20/9, 4/10-18/10	2 hours	CA_B/B342	Hoang Tam Vo
	Wed	18:00	25/7-19/9, 3/10-17/10	2 hours	CA_K/K107	Amir Homayoon Ashrafzadeh
	Mon	16:00	23/7-17/9, 1/10-15/10	2 hours	CA_K/K108	Kane (Mingzhao) Li
	Mon	14:00	23/7-17/9, 1/10-15/10	2 hours	CA_T/T134	Kane (Mingzhao) Li
	Thu	14:00	26/7-20/9, 4/10-18/10	2 hours	CA_B/B342	Zahra Roshan Zamir
	Thu	16:00	26/7-20/9, 4/10-18/10	2 hours	CA_B/B344	Zhinoos Razavi Hesabi
	Wed	10:00	25/7-19/9, 3/10-17/10	2 hours	CA_K/K108	Rasika Amarasiri
	Mon	10:00	23/7-17/9, 1/10-15/10	2 hours	CA_B/B344	Rasika Amarasiri
	Tue	10:00	24/7-18/9, 2/10-16/10	2 hours	CA_B/B348B	Bob Dao
	Tue	14:00	24/7-18/9, 2/10-16/10	2 hours	CA_B/B350	Bob Dao
	Wed	16:00	25/7-19/9, 3/10-17/10	2 hours	CA_K/K108	Amir Homayoon Ashrafzadeh

Outline

1. Motivations
2. Introduction to FIT5196 Data Wrangling
3. Introduction to Data Wrangling
   - Data Quality Problems
   - Characteristics of Tidy Data
   - Major Tasks in Data Wrangling
   - Programming Environment
4. Demonstration: Wrangling Air Crashes data with Data Wrangler
5. Summary

Data Wrangling: the No.1 challenge in data analysis

- Challenges:
  - A massive amount of data has become available and can be collected from various sources, like mobile devices, web pages, email, ATMs, social media, corporate databases, etc
    - Freebase dump: 250G
  - Data from different sources often comes in different formats (e.g., JSON, XML, CSV, Excel and PDF)
  - Data possesses a variety of data quality issues (e.g., inconsistent values, duplicates, missing values, and outliers)
  - It is impossible to directly run any existing data analysis algorithms over raw data.
  - The whole process of data wrangling can account up to 80% of the time in the analysis cycle.
- Opportunities:
  - One prediction for big data analytics: Automating and simplifying complex data wrangling process with machine learning technologies.
    - Enable enterprises, like banks and finance institutions, gain better insights and derive greater business values from their data.
  - Demands from various domains: business, finance, health informatics, government agents. etc.

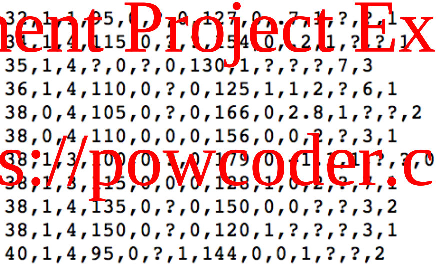
Data glitches — data quality problems

The real-world data is almost always incomplete, dirty and inconsistent, which attributes to the necessity of data wrangling.

- Where the data problems come from? For example,
    - Manual entry errors
    - Malfunction of measurement devices
    - Data sources follow different conventions, formats, or data models.
- Data Quality problems:
    - Interpretability issue
    - Data format issues
    - Inconsistent and faulty data
    - Missing values
    - Outliers
    - Duplicates

Data quality problems: Interpretability issue

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
```

Figure: The Switzerland heart disease data set from UCI machine learning repository

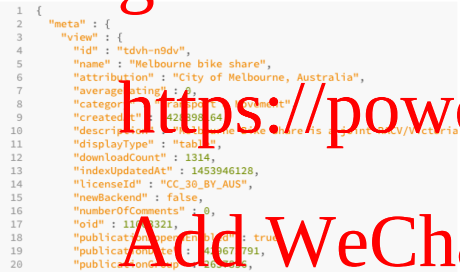
- Attributes in columns in order: age, sex, chest pain type, resting blood pressure, serum cholesterol, fasting blood sugar, etc. [3]
- A data dictionary is needed.

[3] See http://archive.ics.uci.edu/ml/datasets/Heart+Disease for more details.

Data quality problems: Data format issues

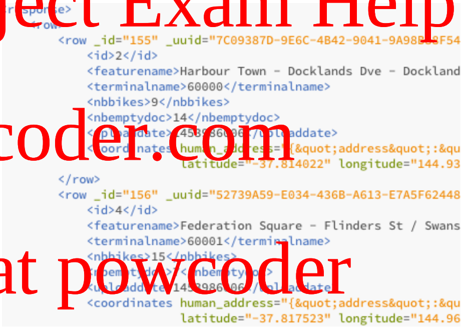
In which format is your data stored?

JavaScript Object Notation (JSON):



```
{
  "meta" : {
    "view" : {
      "id" : "tdvh-n9dv",
      "name" : "Melbourne bike share",
      "attribution" : "City of Melbourne, Australia",
      "averageRating" : 0,
      "category" : "Transport / Movement",
      "createdAt" : 1428398164,
      "description" : "Melbourne Bike Share is a joint RACV/Victoria
      "displayType" : "table",
      "downloadCount" : 1314,
      "indexUpdatedAt" : 1453946128,
      "licenseId" : "CC_30_BY_AUS",
      "newBackend" : false,
      "numberOfComments" : 0,
      "oid" : 11003321,
      "publicationAppendEnabled" : true,
      "publicationDate" : 1429677791,
      "publicationGroup" : 2637836,
```

Extensible Markup Language (XML)



```
<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="{&quot;address&quot;:&qu
                   latitude="-37.814022" longitude="144.93
    </row>
    <row _id="156" _uuid="52739A59-E034-436B-A613-E7A5F62448
      <id>4</id>
      <featurename>Federation Square - Flinders St / Swans
      <terminalname>60001</terminalname>
      <nbbikes>15</nbbikes>
      <nbemptydoc>7</nbemptydoc>
      <uploaddate>1453986006</uploaddate>
      <coordinates human_address="{&quot;address&quot;:&qu
                   latitude="-37.817523" longitude="144.96
```

- other formats:
  - CSV, Excel and PDF

Data quality problems: Inconsistent and faulty data

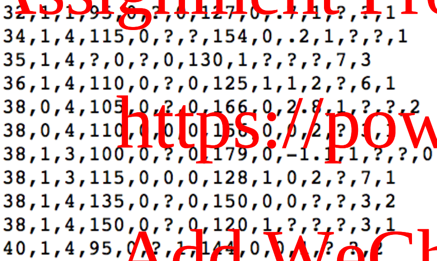
Does you data contain mis-typed, non-standard, inconsistent entries, etc.?

Mr. Mark John	33	21-08-1985	180	M	0433010010	Mel,VIC
Mr. Chris, Peter	34	21-Sep-1982	?	Fale	0000000000	Syd, NSW
Ethan Steedman	36	01/01/82	17o	M	0388886789	Mel,VIC

- Inconsistant date format
- Age not matching date of birth
- Different name formats

Data quality problems: Missing values

In your data set, are any data values that should be presented but absent for some reasons?



```
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
```

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

- Big issue of finding missing values: missing values are defaulted to a valid value of the variable itself.
    - Represent missing values by zero
- Missing data could result in serious bias in the analyses.

Data quality problems: outliers

Is there any observation that lies an abnormal distance from the majority of the other observations in the dataset?[4]

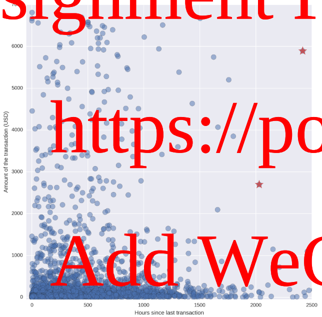


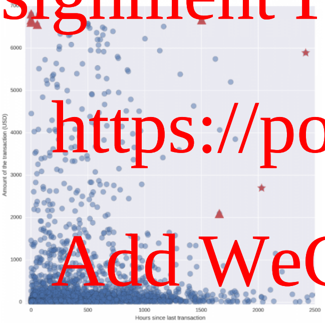
Figure: Bank transaction data, where x: hours since last transaction, y: transaction amount

- Outliers can be either bad or interesting:
  - Finance institutions might be interested in identifying transactions that do not behave in a normal way.
    - High value transactions are occurring on inactive accounts

[4] Figure is from http://blog.easysol.net/advanced-outlier-detection/

Data quality problems: outliers

Is there any observation that lies an abnormal distance from the majority of the other observations in the dataset?[4]



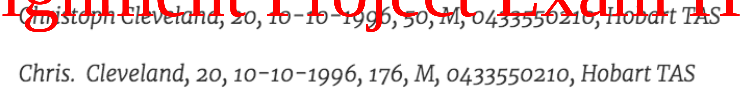
- Outliers can be either bad or interesting:
  - Finance institutions might be interested in identifying transactions that do not behave in a normal way.
    - High value transactions are occurring on inactive accounts

Figure: Outliers identifies by the Local Outlier Factor (LOF) method.

[4] Figure is from http://blog.easysol.net/advanced-outlier-detection/

Data quality problems: Duplicates

In your data, are there multiple entries that actually corresponds 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

- For example, in a database where DOB and mobile phone number can uniquely identify an individual and the attribute of interest is height, the two entries above are duplicates.

Data quality problems: Consequences

"Data quality issues can seriously skew the results of data mining and analysis, with consequences that can potentially cost billions; corporations could make erroneous decisions on misleading results, and machinery could be incorrectly calibrated leading to disastrous failures." — by Dasu. [5]

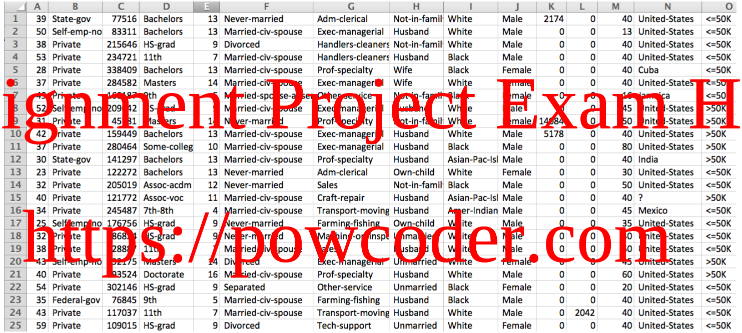
Example

A credit card company is interested in predicting whether an individual will default on his or her credit payment.

- The company will pay for a large price for misclassifying defaulters as non-defaulters due to the data quality problems.

[5] "Data Glitches: Monsters in Your Data" in "Handbook of Data Quality" 2013

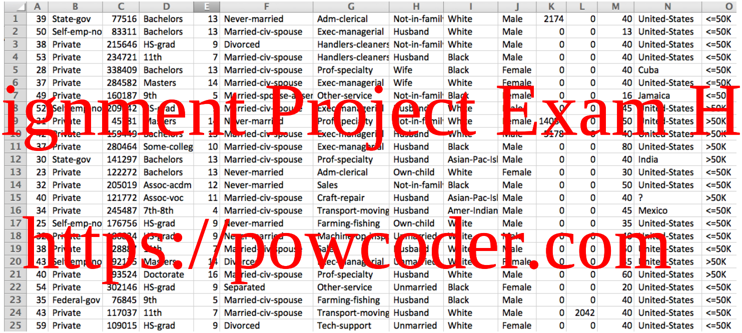
Tidy data



	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-absen	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
17	25	Self-emp-no	176756	HS-grad	9	Never-married	Farming-fishing	Own-child	White	Male	0	0	35	United-States	<=50K
18	32	Private	186824	HS-grad	9	Never-married	Machine-op-inspc	Unmarried	White	Male	0	0	40	United-States	<=50K
19	38	Private	28887	11th	7	Married-civ-spouse	Sales	Husband	White	Male	0	0	50	United-States	<=50K
20	43	Self-emp-no	292175	Masters	14	Divorced	Exec-managerial	Unmarried	White	Female	0	0	45	United-States	>50K
21	40	Private	193524	Doctorate	16	Married-civ-spouse	Prof-specialty	Husband	White	Male	0	0	60	United-States	>50K
22	54	Private	302146	HS-grad	9	Separated	Other-service	Unmarried	Black	Female	0	0	20	United-States	<=50K
23	35	Federal-gov	76845	9th	5	Married-civ-spouse	Farming-fishing	Husband	Black	Male	0	0	40	United-States	<=50K
24	43	Private	117037	11th	7	Married-civ-spouse	Transport-moving	Husband	White	Male	0	2042	40	United-States	<=50K
25	59	Private	109015	HS-grad	9	Divorced	Tech-support	Unmarried	White	Female	0	0	40	United-States	<=50K

- Data Structure: most statistical datasets are rectangular tables made up of rows and columns.
- Data semantics
  - A dataset is a collection of Values.
  - A variable contains all values that measure the same underlying attribute.
  - An observation contains all values measured on the same unit (like a patient)

Tidy data

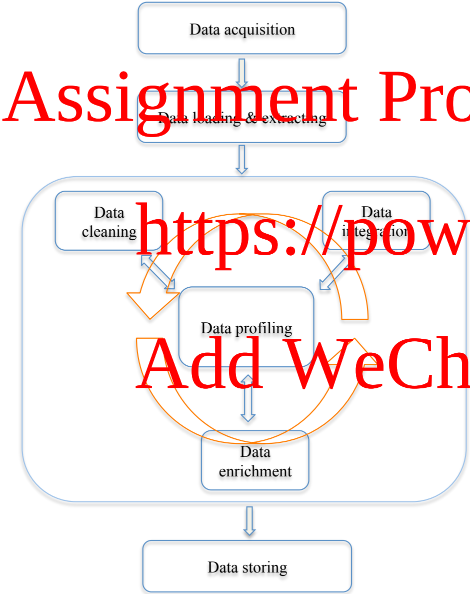


	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1	39	State-gov	77516	Bachelors	13	Never-married	Adm-clerical	Not-in-famil	White	Male	2174	0	40	United-States	<=50K
2	50	Self-emp-no	83311	Bachelors	13	Married-civ-spouse	Exec-managerial	Husband	White	Male	0	0	13	United-States	<=50K
3	38	Private	215646	HS-grad	9	Divorced	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
17	25	Self-emp-no	176756	HS-grad	9	Never-married	Farming-fishing	Own-child	White	Male	0	0	35	United-States	<=50K
18	32	Private	186824	HS-grad	9	Never-married	Machine-op-insp	Unmarried	White	Male	0	0	40	United-States	<=50K
19	38	Private	28887	11th	7	Married-civ-spouse	Sales	Husband	White	Male	0	0	50	United-States	<=50K
20	43	Self-emp-no	292175	Masters	14	Divorced	Exec-managerial	Unmarried	White	Female	0	0	45	United-States	>50K
21	40	Private	193524	Doctorate	16	Married-civ-spouse	Prof-specialty	Husband	White	Male	0	0	60	United-States	>50K
22	54	Private	302146	HS-grad	9	Separated	Other-service	Unmarried	Black	Female	0	0	20	United-States	<=50K
23	35	Federal-gov	76845	9th	5	Married-civ-spouse	Farming-fishing	Husband	Black	Male	0	0	40	United-States	<=50K
24	43	Private	117037	11th	7	Married-civ-spouse	Transport-moving	Husband	White	Male	0	2042	40	United-States	<=50K
25	59	Private	109015	HS-grad	9	Divorced	Tech-support	Unmarried	White	Female	0	0	40	United-States	<=50K

- A dataset is messy or tidy depending on how rows, columns and types are matched up with observations, variables and tables.[6]
  - Each variable forms a column
  - Each observation forms a row
  - Each type of observational unit forms a table
  - If you have multiple tables, they should include a column in the table that allows them to be linked.

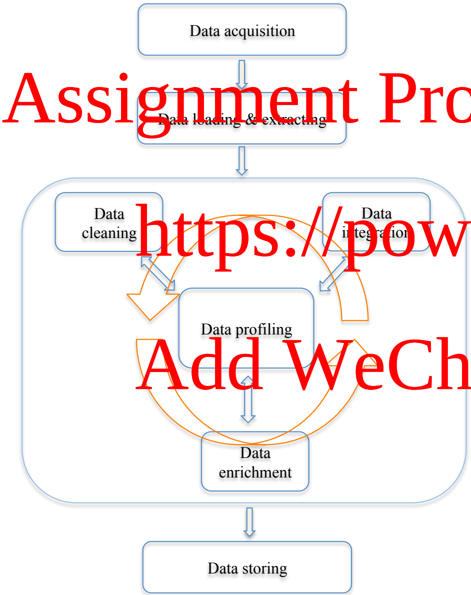
[6] See "Tidy Data" By Hadley Wickham, published in Journal of Statistical Software, 2014

Major Task in Data Wrangling



- Data acquisition: Gather data from different resources, e.g., the web, sensors, and conventional databases via API requests (e.g., Twitter's API and Google API), web scraping (acquiring data from the internet through many ways other than API access), etc. Tools used include various python package, pandas, R, etc.

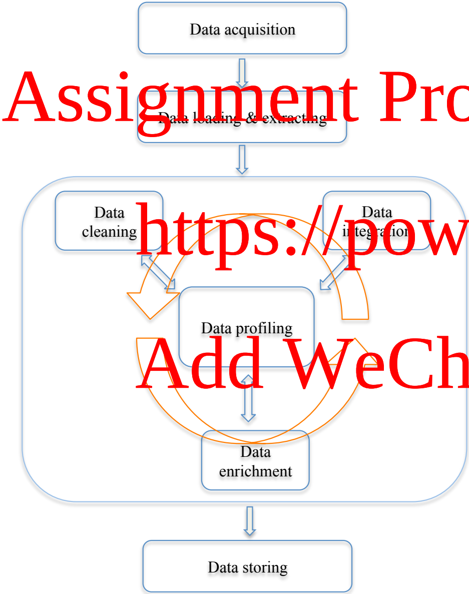
Major Task in Data Wrangling



- Data loading & extracting: Load and parse data stored in many different formats like XML, JSON, CSV, natural language text, etc. Tools used include, for instance, BeautifulSoup (one of many python packages for parsing XML/HTML), regular expressions, NLTK (a python package for natural language processing).

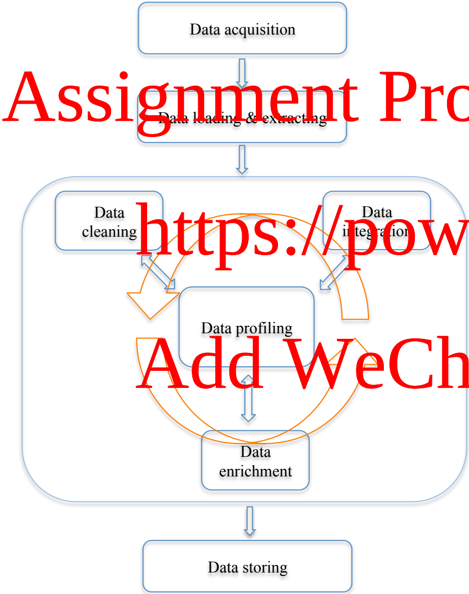
Major Task in Data Wrangling

MONASH University



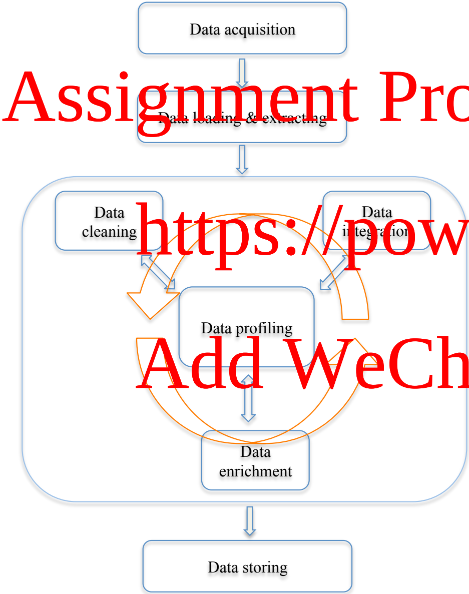
- Data cleaning: Diagnose and handle various data quality problems. Performing data cleaning we need a set of operations that impute missing values, resolve inconsistencies, identify/remove outliers, unify data formats and other problems discussed in previous slides.

Major Task in Data Wrangling



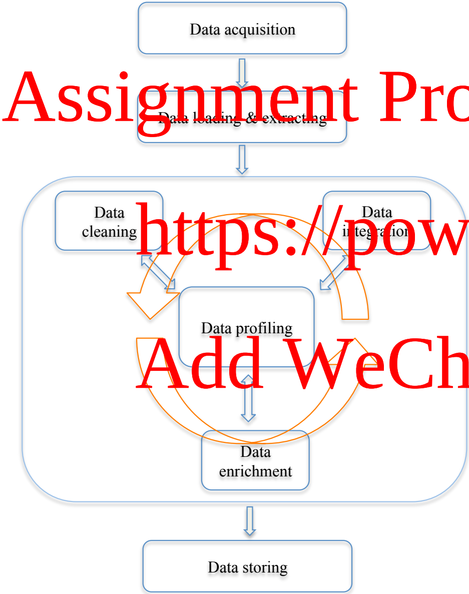
- Data integration: Merge data from different resources to create a rich and complete data set. It involves a set of operations that resolve related issues, such as data duplication, entity matching, and schema matching.

Major Task in Data Wrangling



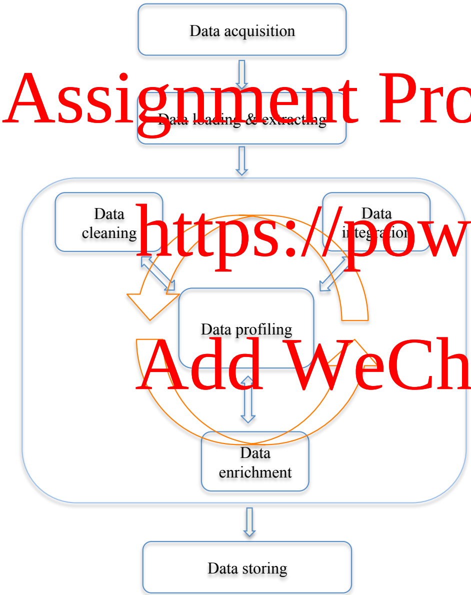
- Data profiling: Utilises different kinds of descriptive statistics and visualisation tools to improve data quality. The data profiling process might uncover more data quality problems, and suggests more operations for data cleaning and data integration.

Major Task in Data Wrangling



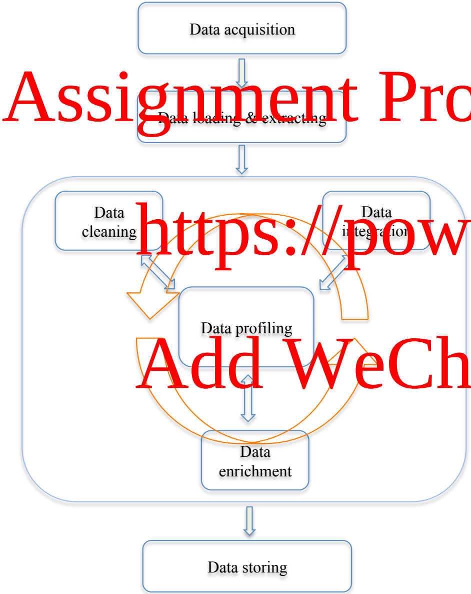
- Data enrichment: Enrich existing data by feature generation, data transformation, data aggregation and data reduction, etc.

Major Task in Data Wrangling



- Data storing: Finally store the clean data in various formats, which are easily accessible by downstream analysis tools.

Major Task in Data Wrangling



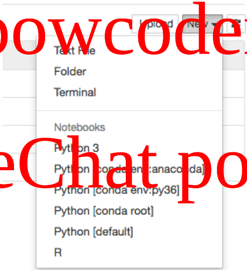
- Documenting the process: We should also keep a detailed description of all data manipulations applied in the above tasks and generate a proper code book that describes each variable and its values in the clean data.
  - Why documentation: collaboration
  - Collaboration tools: Jupyter notebook, Github, etc.

Programming language & environment: Python + Jupyter Notebook

- Programming language: Python 2.7 or 3.6
  - A scripting language that is easy to get started with and it also comes with a large number of libraries that can be used in data wrangling tasks
  - Major libraries used in this units include (but not limited to)
    - Pandas: a library that provides high-level data structures and manipulation tools that are designed to make data processing fast and easy in Python
    - NLTK: a platform for building Python programs to work with human language data
    - BeautifulSoup: a simple and efficient library for navigating, searching, and modifying HTML and XML documents.
    - Scipy: a fundamental library for scientific computing.
    - scikit-learn: an efficient Python library for data mining and data analysis.
- Programming environment: Jupyter notebook
  - The Jupyter Notebook is a web application that allows you to create and share documents that contain live code, equations, visualisations and explanatory text.

Programming language & environment: Python + Jupyter Notebook

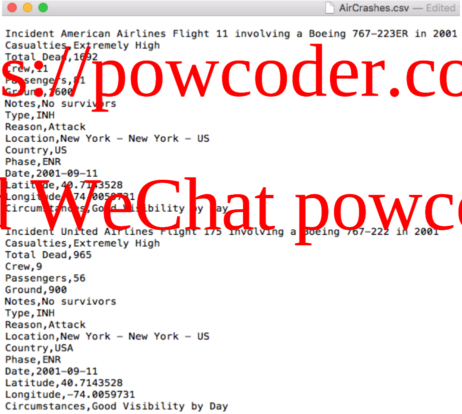
- Dual Python environments
  - Conda Managing environments: https://conda.io/docs/user-guide/tasks/manage-environments.html
    1. conda create -n python3 python=3.6
    2. conda install nb_conda



- Most of the notebooks will run in both versions.

Wrangling Air Crashes data with Data Wrangler

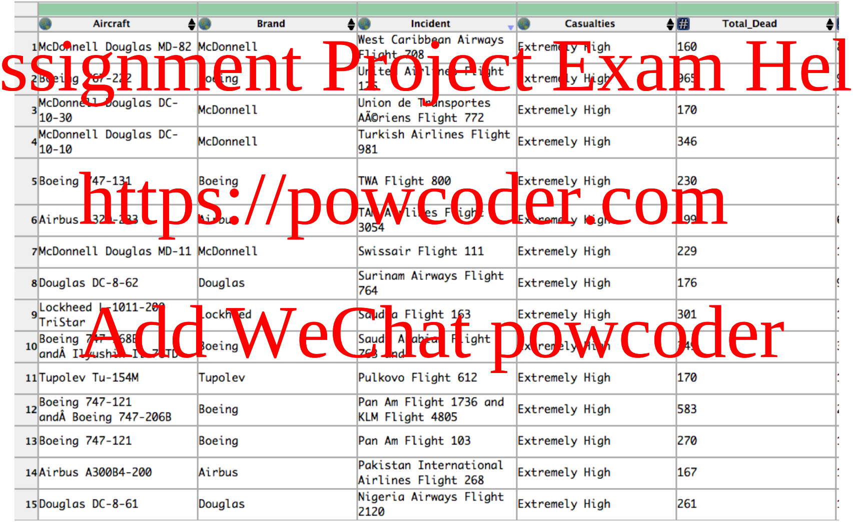
- Data set: Air Crashes data downloaded from Wikipedia
- Application: Data Wrangler from http://vis.stanford.edu/wrangler/
- Goal: wrangle the data so that each row corresponds to one crash disaster, each column is one variable.



```
Incident American Airlines Flight 11 involving a Boeing 767-223ER in 2001
Casualties,Extremely High
Total Dead,1692
Crew,11
Passengers,81
Ground,1600
Notes,No survivors
Type,INH
Reason,Attack
Location,New York - New York - US
Country,US
Phase,ENR
Date,2001-09-11
Latitude,40.7143528
Longitude,-74.0059731
Circumstances,Good Visibility by Day

Incident United Airlines Flight 175 involving a Boeing 767-222 in 2001
Casualties,Extremely High
Total Dead,965
Crew,9
Passengers,56
Ground,900
Notes,No survivors
Type,INH
Reason,Attack
Location,New York - New York - US
Country,USA
Phase,ENR
Date,2001-09-11
Latitude,40.7143528
Longitude,-74.0059731
Circumstances,Good Visibility by Day
```

Wrangling Air Crashes data with Data Wrangler



	Aircraft	Brand	Incident	Casualties	Total_Dead
1	McDonnell Douglas MD-82	McDonnell	West Caribbean Airways Flight 708	Extremely High	160
2	Boeing 767-222	Boeing	United Airlines Flight 175	Extremely High	965
3	McDonnell Douglas DC-10-30	McDonnell	Union de Transportes AÃ©riens Flight 772	Extremely High	170
4	McDonnell Douglas DC-10-10	McDonnell	Turkish Airlines Flight 981	Extremely High	346
5	Boeing 747-131	Boeing	TWA Flight 800	Extremely High	230
6	Airbus A320-233	Airbus	TAM Airlines Flight 3054	Extremely High	199
7	McDonnell Douglas MD-11	McDonnell	Swissair Flight 111	Extremely High	229
8	Douglas DC-8-62	Douglas	Surinam Airways Flight 764	Extremely High	176
9	Lockheed L-1011-200 TriStar	Lockheed	Saudia Flight 163	Extremely High	301
10	Boeing 747-168B andÂ Ilyushin Il-76TD	Boeing	Saudi Arabian Flight 763 and	Extremely High	349
11	Tupolev Tu-154M	Tupolev	Pulkovo Flight 612	Extremely High	170
12	Boeing 747-121 andÂ Boeing 747-206B	Boeing	Pan Am Flight 1736 and KLM Flight 4805	Extremely High	583
13	Boeing 747-121	Boeing	Pan Am Flight 103	Extremely High	270
14	Airbus A300B4-200	Airbus	Pakistan International Airlines Flight 268	Extremely High	167
15	Douglas DC-8-61	Douglas	Nigeria Airways Flight 2120	Extremely High	261

Wrangling Air Crashes data with Data Wrangler

Demonstration with Data Wrangler!

Summary: what to do this week

- Please please download and read materials provided in Moodle.
- Set up your programming environment by installing Anaconda Python 2 or Python 3 distribution. (Suggestion!)
- Attend tutorial 1 in **week 2**.
- Last but not least
  - Choose FIT5196 wisely.
  - Use the discussion forum in a proper way and with respect!