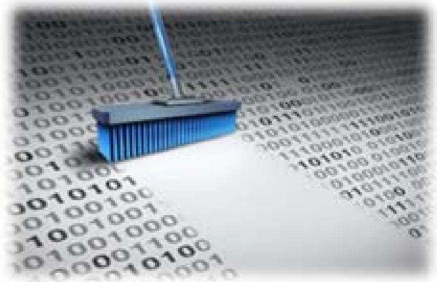


Research Skill Enhancement Webinar co-hosted by RDA CODATA Summer School and CODATA Connect group

presented by

Simisani Ndaba

Importance of Data Cleaning

05 August 2021

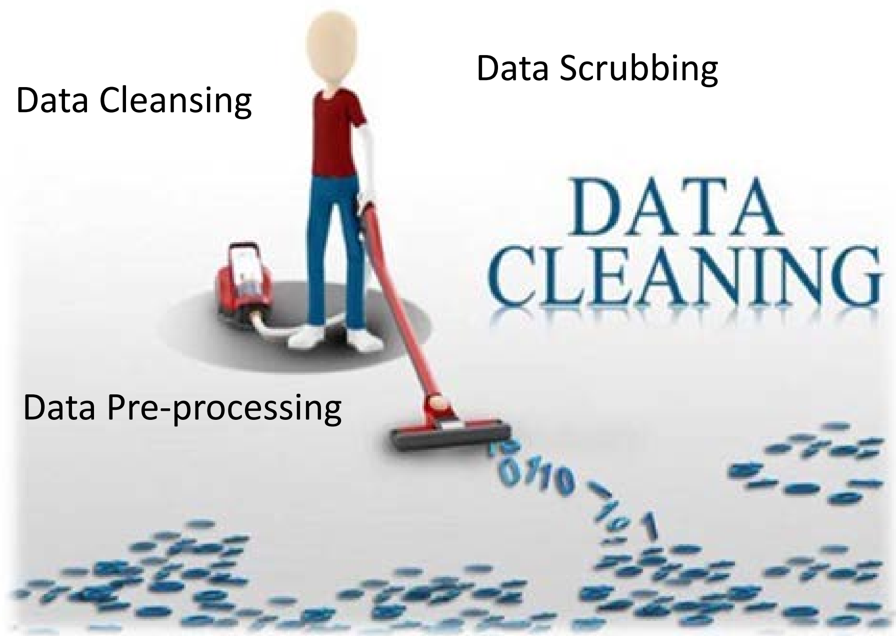
Overview

- Meaning of Data Cleaning
- Need for Data Cleaning
- Data Cleaning Methods
- Data Cleaning Steps
- Best Practices
- Data Quality Attributes
- How Data Cleaning is used in a Dataset
- Overall Benefits of Data Cleaning

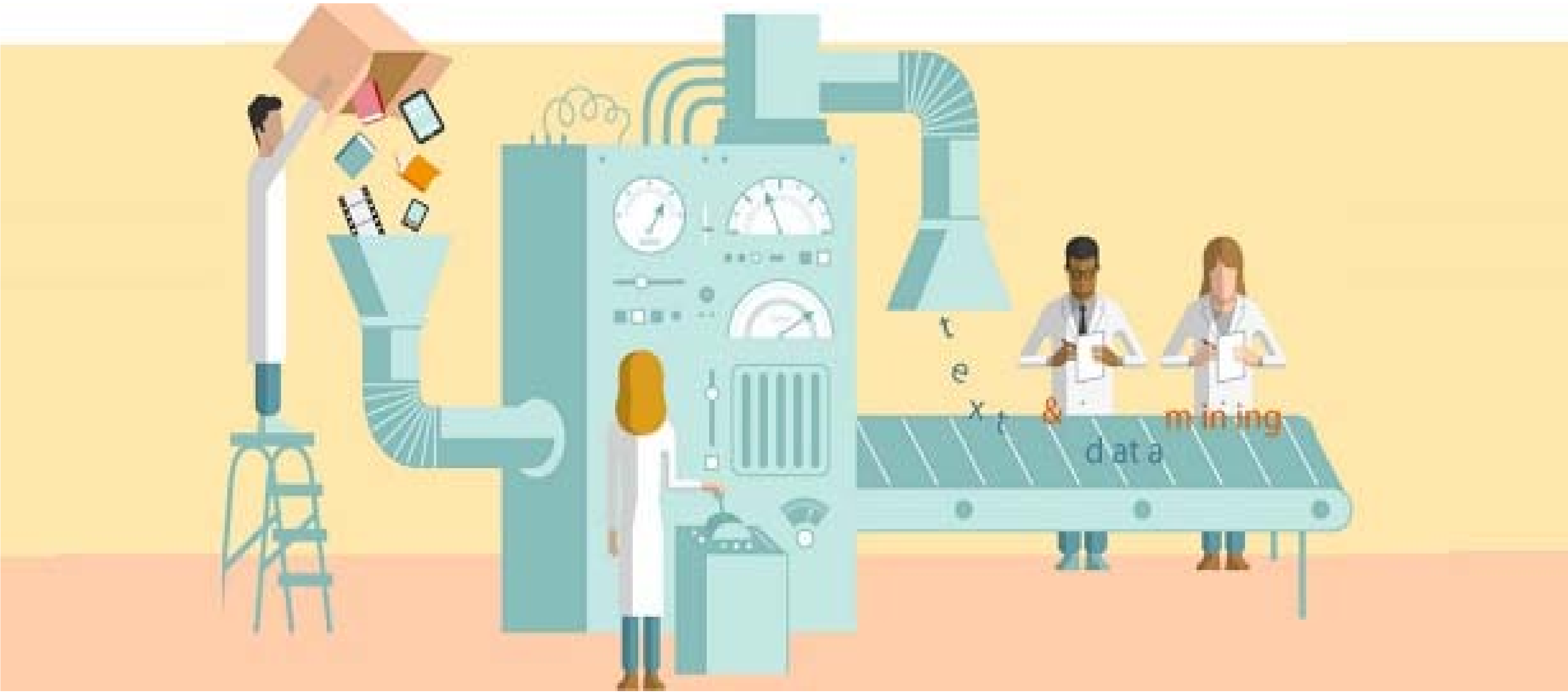
What is Data Cleaning?

Data cleaning is a process in which you go through all of the data in a data set and either;

- remove or update information that is incomplete,
- incorrect,
- improperly formatted,
- duplicated, or
- irrelevant.



Michael Walker (2021) Python Data Cleaning Cookbook


text
&
data
mining

Raw Data vs Clean Data

- Raw data is the data that is collected directly from the data source,

RAW DATA NP I		Number of colonies					
Treatment	Concentration	dish 1	dish 2	dish 3	dish 4	dish 5	dish 6
Positive Control	100µM	0	0	0	0	0	0
Control	0	122	132	120	134	123	154
Solvent Control	0.04%	152	139	132	118	148	142
I	1	145	134	144	149	138	129
I	5	137	133	143	155	141	135
I	10	129	124	135	138	146	143
I	12.5	146	113	131	138	130	145
I	15	72	75	75	82	96	101
I	20	55	28	17	77	41	10
I	25	0	0	0	0	0	0

Template showing an example of raw data: the number of colonies per treatment condition and controls, Ponti et al (2014)

Raw Data vs Clean Data

- "Dirty Data" is raw data full of irrelevances, errors, and corrupt information
- Clean data is in analyzable format

(a) Dirty Data

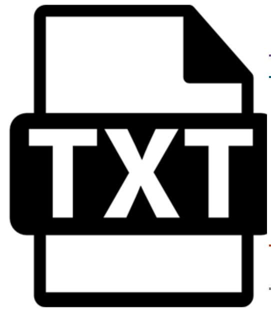
id	title	pub_year	citation _count
t_1	CrowdDB	11	18
t_2	TinyDB	2005	1569
t_3	YFilter	Feb, 2002	298
t_4	Aqua		106
t_5	DataSpace	2008	107
t_6	CrowdER	2012	1
t_7	Online Aggr.	1997	687
...	...	...	...
t_{10000}	YFilter - ICDE	2002	298

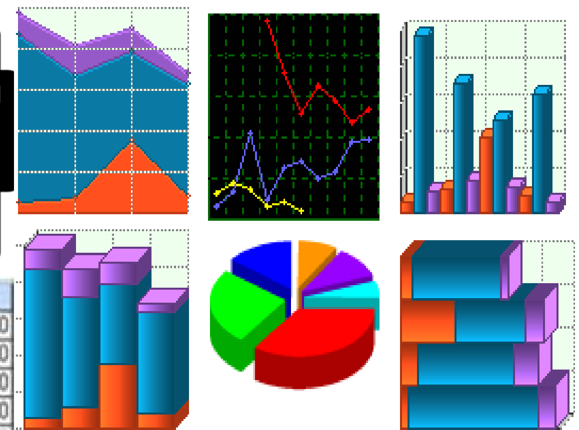
(b) Cleaned Sample

id	title	pub_year	citation _count	#dup
t_1	CrowdDB	**2011**	**144**	**2**
t_2	TinyDB	2005	1569	1
t_3	YFilter	**2002**	298	**2**
t_4	Aqua	**1999**	106	1
t_5	DataSpace	2008	107	1
t_6	CrowdER	2012	**34**	1
t_7	Online Aggr.	1997	687	**3**

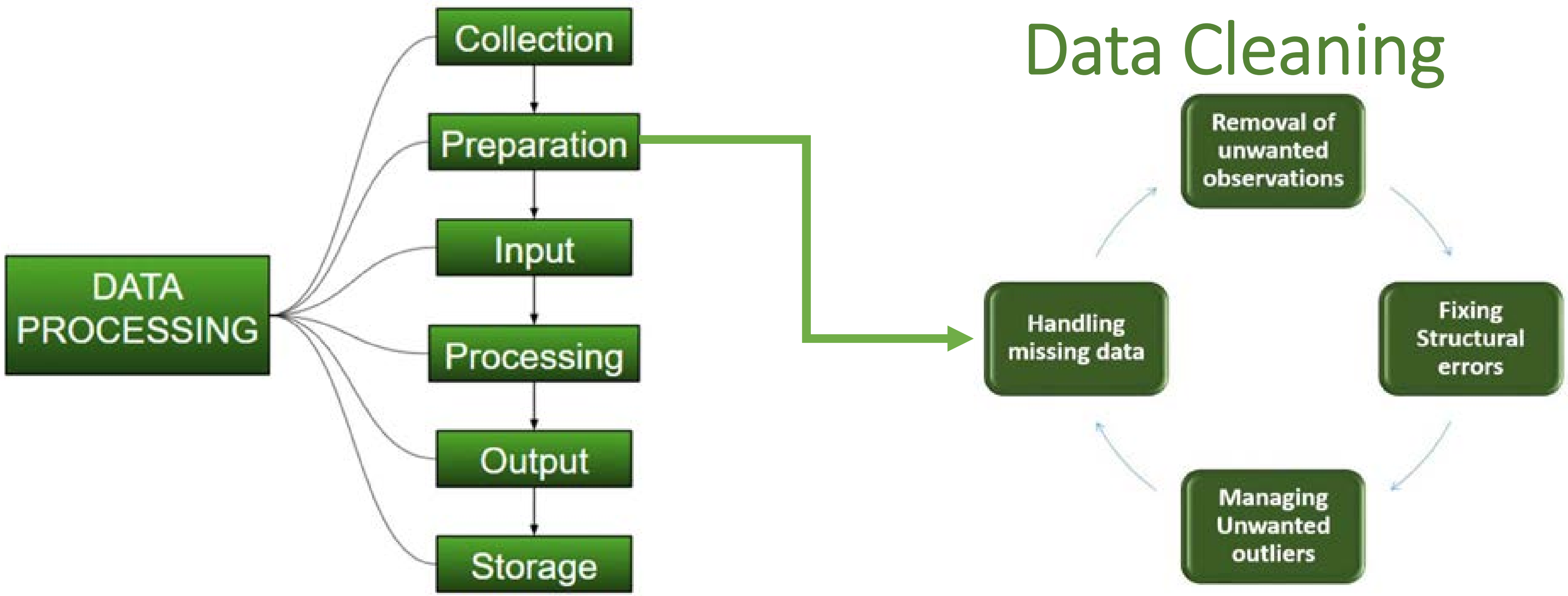
An example of dirty data and cleaned sample (Shaded cells denote dirty values, and their cleaned values are in bold font), Krishnan et al(2014)

RAW Data → Processed Data





Work Item	Vendor	Labor	Equipment	Materials	Subcontr.	Subtotal	Markup %	Markup	Total
Permits/Fees	City of Los Angeles				$1,500.00	$1,500.00		$0.00	$1,500.00
Excavation		$6,000.00	$8,000.00	$500.00		$14,500.00	15.00%	$2,175.00	$16,675.00
Utilities		$3,500.00	$2,500.00	$2,750.00	$1,000.00	$9,750.00	15.00%	$1,462.50	$11,212.50
Water Well						$0.00		$0.00	$0.00
Septic Tank						$0.00		$0.00	$0.00
Foundation	Connie's Concrete				$3,500.00	$3,500.00	5.00%	$175.00	$3,675.00
Concrete Flatwork	Connie's Concrete				$1,900.00	$1,900.00	5.00%	$95.00	$1,995.00
Framing		$3,500.00	$1,500.00	$9,000.00		$14,000.00	15.00%	$2,100.00	$16,100.00
Roofing	Robert's Roofing				$3,500.00	$3,500.00	5.00%	$175.00	$3,675.00
Windows/Ext.Doors	Wally's Windows				$8,000.00	$8,000.00	5.00%	$400.00	$8,400.00
Garage Door	Gary's Garage Doors				$2,250.00	$2,250.00	5.00%	$112.50	$2,362.50
Siding						$0.00		$0.00	$0.00
Electrical	Ernie's Electric				$18,500.00	$18,500.00	5.00%	$925.00	$19,425.00
Plumbing	Mac's Mechanical				$16,500.00	$16,500.00	5.00%	$825.00	$17,325.00
HVAC	Mac's Mechanical				$23,000.00	$23,000.00	5.00%	$1,150.00	$24,150.00
Insulation		$3,500.00		$1,000.00		$4,500.00		$0.00	$4,500.00
Masonry	Mason's Masonry				$14,500.00	$14,500.00	5.00%	$725.00	$15,225.00
Drywall	Doug's Drywall				$12,500.00	$12,500.00	5.00%	$625.00	$13,125.00
Interior Trim	Doug's Drywall				$9,000.00	$9,000.00	5.00%	$450.00	$9,450.00
Painting	Paul's Painting				$13,500.00	$13,500.00	5.00%	$675.00	$14,175.00


DATA PROCESSING
Collection
Preparation
Input
Processing
Output
Storage
Data Cleaning
Removal of unwanted observations
Fixing Structural errors
Managing Unwanted outliers
Handling missing data

#	Id	Name	Birthday	Gender	IsTeacher?	#Students	Country	City
1	111	John	31/12/1990	M	0	0	Ireland	Dublin
2	222	Mery	15/10/1978	F	1	15	Iceland	
3	333	Alice	19/04/2000	F	0	0	Spain	Madrid
4	444	Mark	01/11/1997	M	0	0	France	Paris
5	555	Alex	15/03/2000	A	1	23	Germany	Berlin
6	555	Peter	1983-12-01	M	1	10	Italy	Rome
7	777	Calvin	05/05/1995	M	0	0	Italy	Italy
8	888	Roxane	03/08/1948	F	0	0	Portugal	Lisbon
9	999	Anne	05/09/1992	F	0	5	Switzerland	Geneva
10	101010	Paul	14/11/1992	M	1	26	Ytali	Rome

Missing values

Invalid values

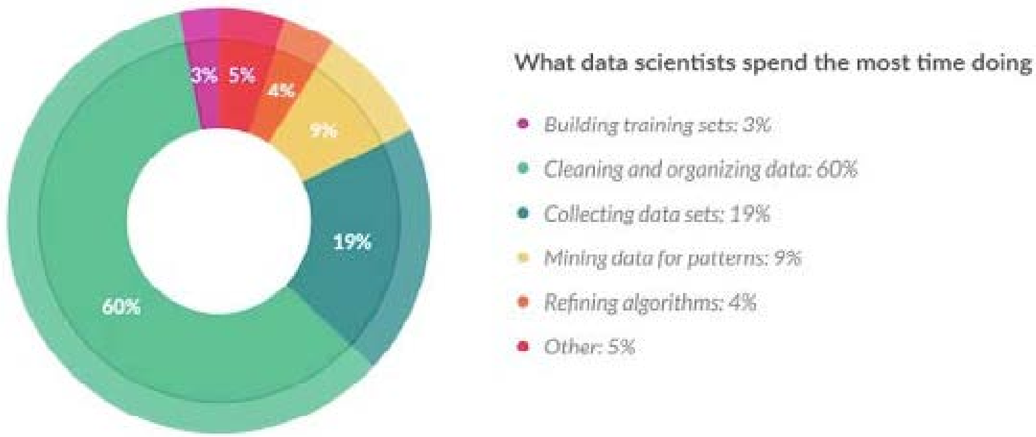
Misfielded values

Misspellings

Uniqueness

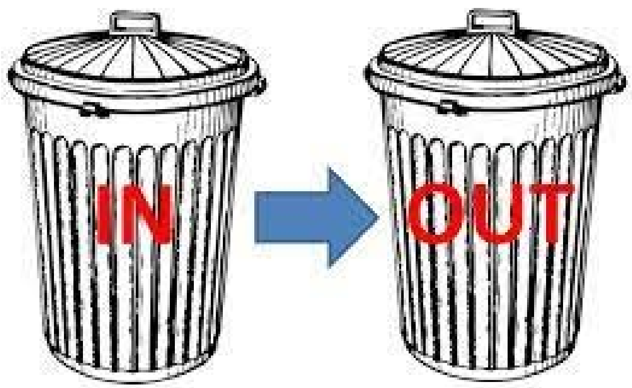
Formats

Attribute dependencies



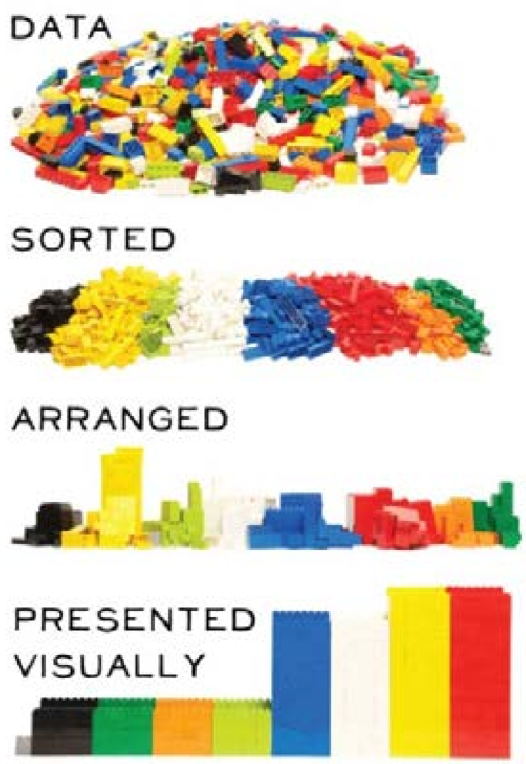
Source: CrowdFlower 2016 to 2018

The Need for Data Cleaning



- Having data that is clean can help in performing the analysis faster, saving precious time.
- Improving the quality of data to make them “fit for use” by users
- Improving users documentation and presentation.
- False conclusions because of incorrect or “dirty” data can inform poor decision-making.
- False conclusions can lead to moments in reporting when you realize your data doesn’t stand up to scrutiny.
- It is important to create a culture of quality data in your research work.

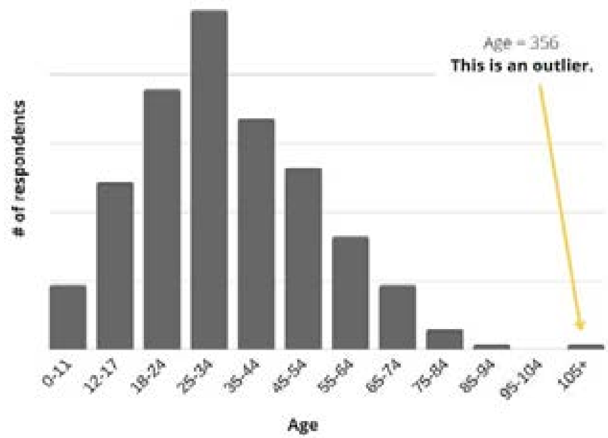
The Need for Data Cleaning contd..



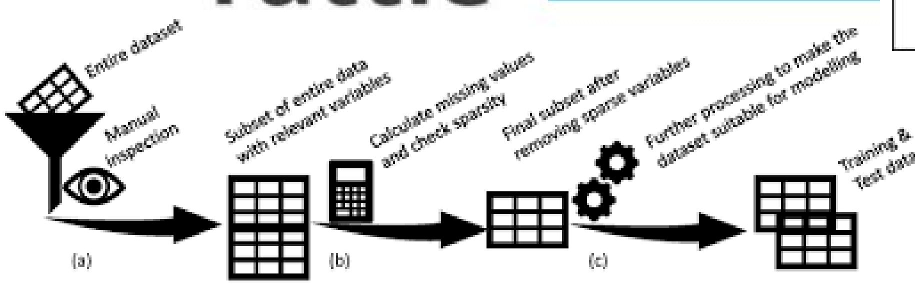
- Combining multiple data sources creates synchronisation issues
- If data is incorrect, outcomes are unreliable
- data cleaning processes will vary from dataset to dataset.
- establish a template for your data cleaning process

Data Cleaning Methods

- Histograms
- Conversion Tables
- Tools
- Algorithms
- Manually




```
Algorithm DataCleaning (LogFile: Web log file; LogFile: Web log file)
    Begin
        While not eof (LogFile) Do
            LogRecord = Read (LogFile)
            If ((LogRecord.Cs-url-stem <> gif,jpeg,jpg,css,js)) AND (LogRecord.Cs-
            method= 'GET') AND (LogRecord.Sc-status = (200))
            Then Write (LogFile, LogRecord)
            End If
        End While
    End
```



How do you Clean Data?

Import data

Merge Data set

ID	var1	var2	var3
588	2	d	1
654	1	y	1
527	1	o	0
955	2	c	0
954	1	t	0

ID	var1	var2	var3
1280	1	p	1
1917	2	t	0
1854	2	x	1
1701	2	e	0
1928	1	q	1

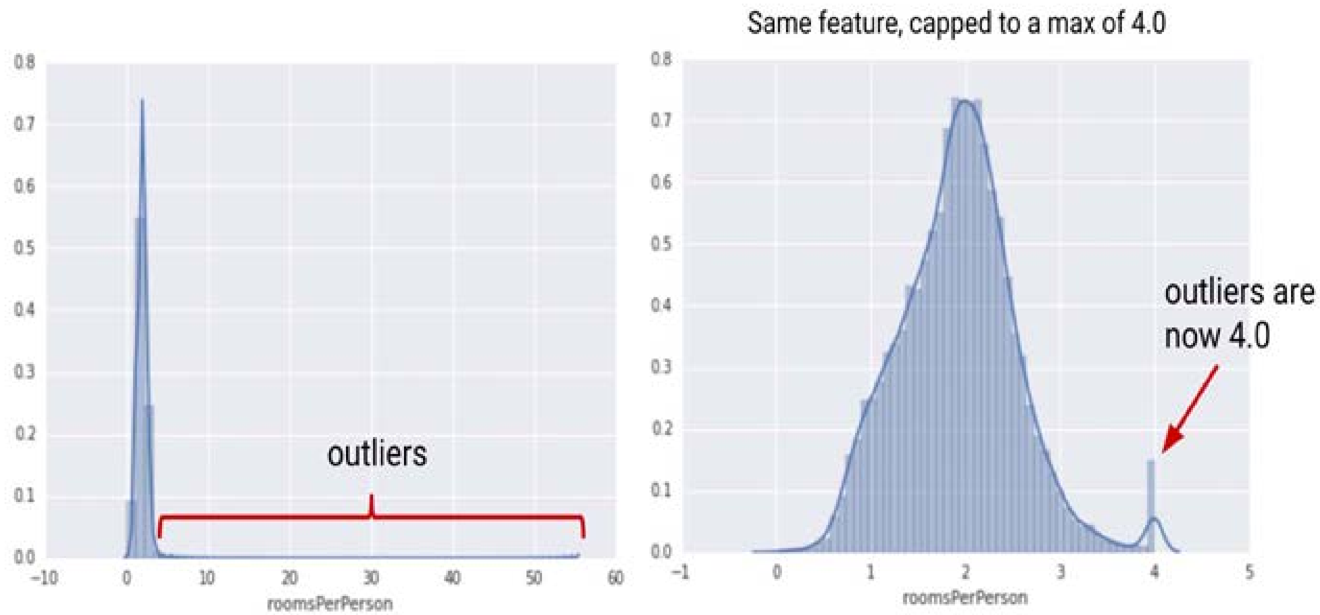
ID	var1	var2	var3
588	2	d	1
654	1	y	1
527	1	o	0
955	2	c	0
954	1	t	0
1280	1	p	1
1917	2	t	0
1854	2	x	1
1701	2	e	0
1928	1	q	1

Rebuilding Missing Data

Mobile ID	Mobile Package	Download Speed	Data Limit Usage
1	Fast+	157	80%
2	N/A	99	70%
3	Fast+	167	10%
4	Fast+	90	80%
5	Lite	76	70%
6	N/A	155	10%
7	Fast+	200	95%
8	Lite	76	77%
9	N/A	180	95%

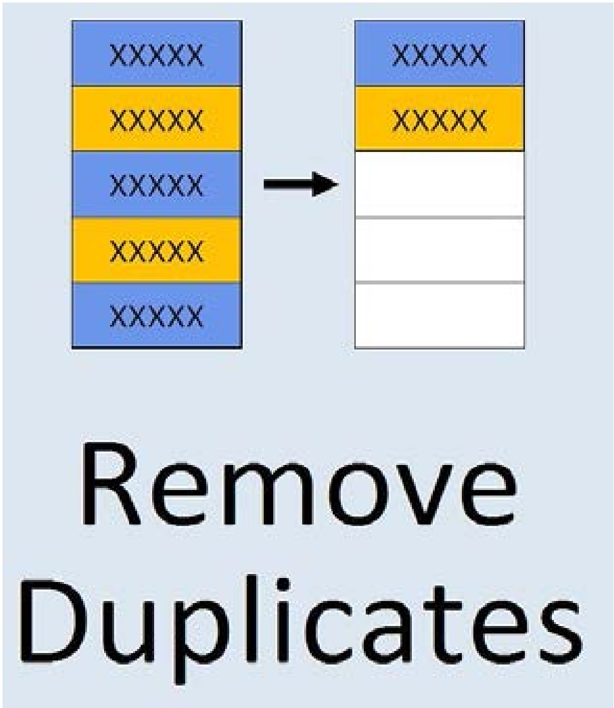
Mobile ID	Mobile Package	Download Speed	Data Limit Usage
1	Fast+	157	80%
2	Fast+	99	70%
3	Fast+	167	10%
4	Fast+	90	80%
5	Lite	76	70%
6	Fast+	155	10%
7	Fast+	200	95%
8	Lite	76	77%
9	Fast+	180	95%

Standardisation



Normalisation

Raw	Normalized
2moro 2mrrw 2morrow 2mrw tomrw	tomorrow
b4	before
otw	on the way
:) :-) ;-)	smile



Verification and Enrichment

Export Data

TIBCO®

WinPure™

talend

tableau®

cloudingo

validity

TRIFACTA

SAP®

Data Ladder
Get the most out of your data

Informatica
Data Quality

IBM InfoSphere. software

snowflake

in Data Cleaning

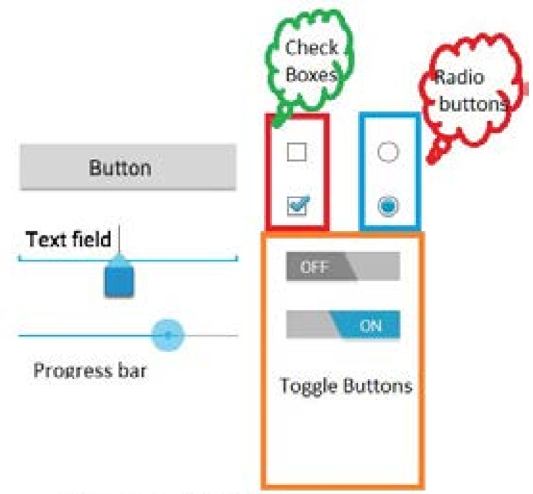
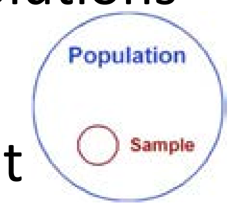


Fig : Input Controls

- Consider your data in the most holistic way possible
- Increased controls on database inputs
- Choose the right software solutions
- Limit your sample size
- Spot check errors throughout
- Leverage free online courses



Data Quality

Validity

The degree to which your data conforms to defined rules or constraints.

- Data-Type Constraints: values must be of a particular datatype, e.g., boolean, numeric, date, etc.
- Range Constraints: numbers or dates should fall within a certain range.
- Mandatory Constraints: certain columns cannot be empty.
- Unique Constraints: a field, or a combination of fields, must be unique across a dataset.

Accuracy

- Ensure your data is close to the true values.
- Defining possible valid values allows invalid values to be easily spotted, it does not mean that they are accurate.
- Difference between accuracy and precision.
  - Accuracy refers to how close a measurement is to the true or accepted value. Precision refers to how close measurements of the same item are to each other. **Precision is independent of accuracy**.

Completeness

- The degree to which all required data is known.
- Missing data is going to happen for various reasons.
- One can mitigate this problem by questioning the original source if possible, say re-interviewing the subject.
- Chances are, the subject is either going to give a different answer or will be hard to reach again.

Consistency

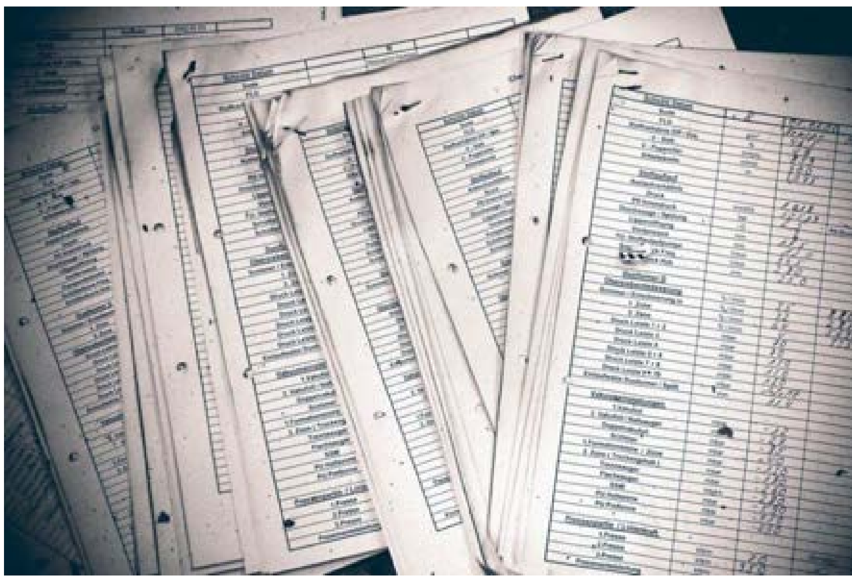
- Ensure your data is consistent within the same dataset and/or across multiple data sets.
- Inconsistency occurs when two values in the data set contradict each other.
- A valid age, say 10, mightn’t match with the marital status, say divorced. A customer is recorded in two different tables with two different addresses. Which one is true?.

Uniqueness

- A measure of unwanted duplication existing within or across systems for a particular field, record, or data set.

Timeliness

- The extent to which age of the data is appropriated for the task at hand.



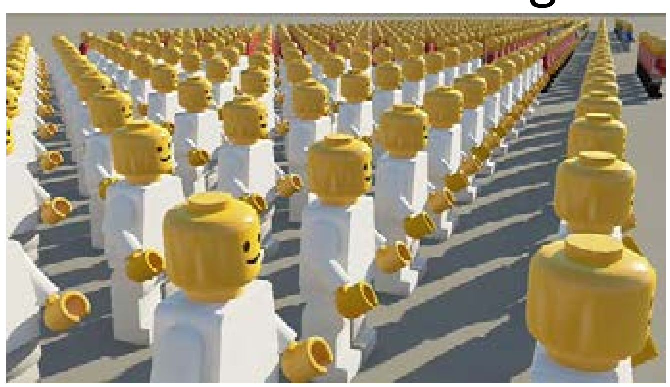

Other Dimensions

Integrity

- The quality, reliability, trustworthiness, and completeness of a data set – providing accuracy, consistency and context.
- This criteria looks as whether a dataset follows the rules and standards set. Are there any values missing that can harm the efficacy of the data or keep analysts from discerning important relationships or patterns?

Uniformity

- The degree to which the data is specified using the same unit of measure.
- The weight may be recorded either in pounds or kilos. The date might follow the USA format or European format. The currency is sometimes in USD and sometimes in YEN.
- And so data must be converted to a single measure unit.



Example using Data cleansing in stages

- The following example a data set containing company registration numbers, e-mails, addresses, etc. consisting on using;
  - Importing dataset
  - Data validation and Removing Irrelevant data
  - Formatting data to a common value (standardization / consistency)
  - Cleaning up duplicates
  - Filling missing data vs. erasing incomplete data

Example using Data cleansing…..
Data cleansing Step 1: Importing Data set

List of tax numbers of Polish companies (Transparent Data,2021)

Data Validation of company TAX numbers (raw data)

Example using Data cleansing…..
Data cleansing Step 2: Data Validation

- In this dataset, the last digit of each tax identification number
- this is called a 'check digit' which is validated by an algorithm
- Check digit Validation =

5422698451

  - multiplying each of the **first nine digits** of the tax number (542269845)by weights (in sequence: 6, 5, 7, 2, 3, 4, 5, 6, 7)

(5*6)+(4*5)+(2*7)+(2*2)+(6*3)+(9*4)+(8*5)+(4*6)+(5*7)=221

  - summing the results of this multiplication, and then dividing checksum by 11.

221%11=0.9

  - The remainder of the division should be identical to the last digit in the tax number, that is, from the list 542269845(1)
  - 0.9 (rounding off)= 1

Example using Data cleansing…..
Data cleansing Step 2: Data Validation and (Removed Irrelevant data)

- Data Validation of company TAX numbers (data after validation)

Example using Data cleansing…..
Data cleansing Step 3: Formatting data to a common form

- The next step is to Normalize the data.
- Some tax numbers were written with dashes, spaces or the prefix "PL" which stands for Poland.

PL6331813071
5422449707
583-101-48-98

- How do we format all company tax numbers to a common form. How?
  - Omit the prefix with the country code.
  - write all numbers without any special characters separating the digits.

Example using Data cleansing…..
Data cleansing Step 3: Formatting data to a common form

after formatting data

Example using Data cleansing.....
Data cleansing Step 4: Cleaning up duplicates

- The next step in data cleaning is to check for duplicates

after removing duplicates

Example using Data cleansing…..
Data cleansing Step 5: Filling missing data and erasing incomplete data

- The next step is preventing the possession of incomplete data.
- Voivodeship or district can be easily completed based on the name of the city or postal code

VOIVODESHIP	DISTRICT	POST CODE	CITY	STREET
warmińsko-mazurskie	ostróda	14-100	ostróda	hurtowa
pomorskie	pruszcz gdański	83-031	łęgowo	ul. marco polo
lubuskie	m. gorzów wielkopolski	66-400	gorzów wielkopolski	wybickiego
warmińsko-mazurskie		10-228	olsztyn	ul. zamenhofa
wielkopolskie	poznań-stare miasto	61-859	poznań	grobla
warmińsko-mazurskie	działdowo	14-120	dąbrówno	agrestowa
warmińsko-mazurskie		10-508	olsztyn	mickiewicza
wielkopolskie	olsztyn	10-450	olsztyn	ul. piłsudskiego
warmińsko-mazurskie	lubomino	11-135	lubomino	kopernika
wielkopolskie	grodzisk wielkopolski	62-065	grodzisk wielkopolski	ul. fabryczna
lubuskie	lubsko	68-300	lubsko	ul. włókniarzy
	m. lublin	20-327	lublin	wrońska
podkarpackie	świlcza	36-072	świlcza	ul. —
podkarpackie	m. rzeszów	35-510	rzeszów	ślusarczyka
śląskie	dąbrowa górnicza	41-300	dąbrowa górnicza	ul. wiejska (baza pec)
pomorskie	m. gdynia	81-212	gdynia	hutnicza
wielkopolskie	poznań-stare miasto	61-623	poznań	ul. wilczak
	poznań-grunwald	60-318	poznań	ul. władysława węgorka
warmińsko-mazurskie	kętrzyn	11-400	kętrzyn	ul. rynkowa
	m. szczecin	70-556	szczecin	null

addresses data set: Filling missing data and erasing incomplete data

Example using Data cleansing…..
Data cleansing Step 5: Filling missing data and erasing incomplete data

VOIVODESHIP	DISTRICT	POST CODE	CITY	STREET
warmińsko-mazurskie	ostróda	14-100	ostróda	hurtowa
pomorskie	pruszcz gdański	83-031	łęgowo	ul. marco polo
lubuskie	m. gorzów wielkopolski	66-400	gorzów wielkopolski	wybickiego
warmińsko-mazurskie	olsztyn	10-228	olsztyn	ul. zamenhofa
wielkopolskie	poznań-stare miasto	61-859	poznań	grobla
warmińsko-mazurskie	działdowo	14-120	dąbrówno	agrestowa
warmińsko-mazurskie	olsztyn	10-508	olsztyn	mickiewicza
wielkopolskie	olsztyn	10-450	olsztyn	ul. piłsudskiego
warmińsko-mazurskie	lubomino	11-135	lubomino	kopernika
wielkopolskie	grodzisk wielkopolski	62-065	grodzisk wielkopolski	ul. fabryczna
lubuskie	lubsko	68-300	lubsko	ul. włókniarzy
lubelskie	m. lublin	20-327	lublin	wrońska
podkarpackie	m. rzeszów	35-510	rzeszów	ślusarczyka
śląskie	dąbrowa górnicza	41-300	dąbrowa górnicza	ul. wiejska (baza pec)
pomorskie	m. gdynia	81-212	gdynia	hutnicza
wielkopolskie	poznań-stare miasto	61-623	poznań	ul. wilczak
wielkopolskie	poznan-grunwald	60-318	poznań	ul. władysława węgorka
warmińsko-mazurskie	kętrzyn	11-400	kętrzyn	ul. rynkowa

Table after filling missing data and erasing incomplete data

Datasets For Data Cleaning Practice

Common Crawl Corpus

Google Books Ngrams

Hotel Booking Demand

FiveThirtyEight

Taxi Trajectory Data

Trending YouTube Video Statistics

Kaggle

Iris Species

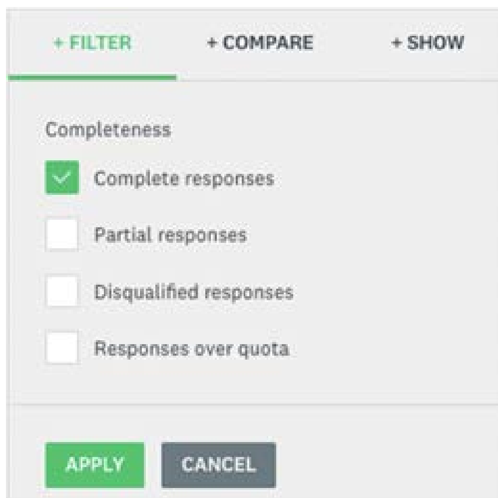
PAN at CLEF

Socrata

Example using Data cleansing in stages

SurveyMonkey.com Data Cleaning

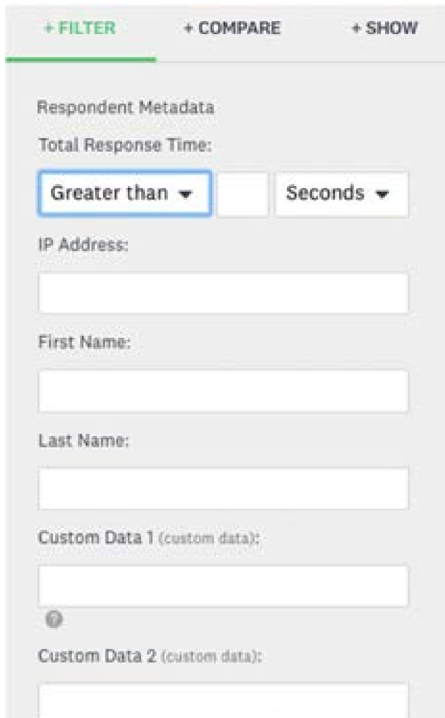
- In qualitative data collection, Survey Data Analysis is used
- Survey data cleaning involves identifying and removing responses
- Going through some common cases in SurveyMonkey.com
  - Respondents who answer a portion of the question



Example using Data cleansing in stages

SurveyMonkey.com Data Cleaning contd…

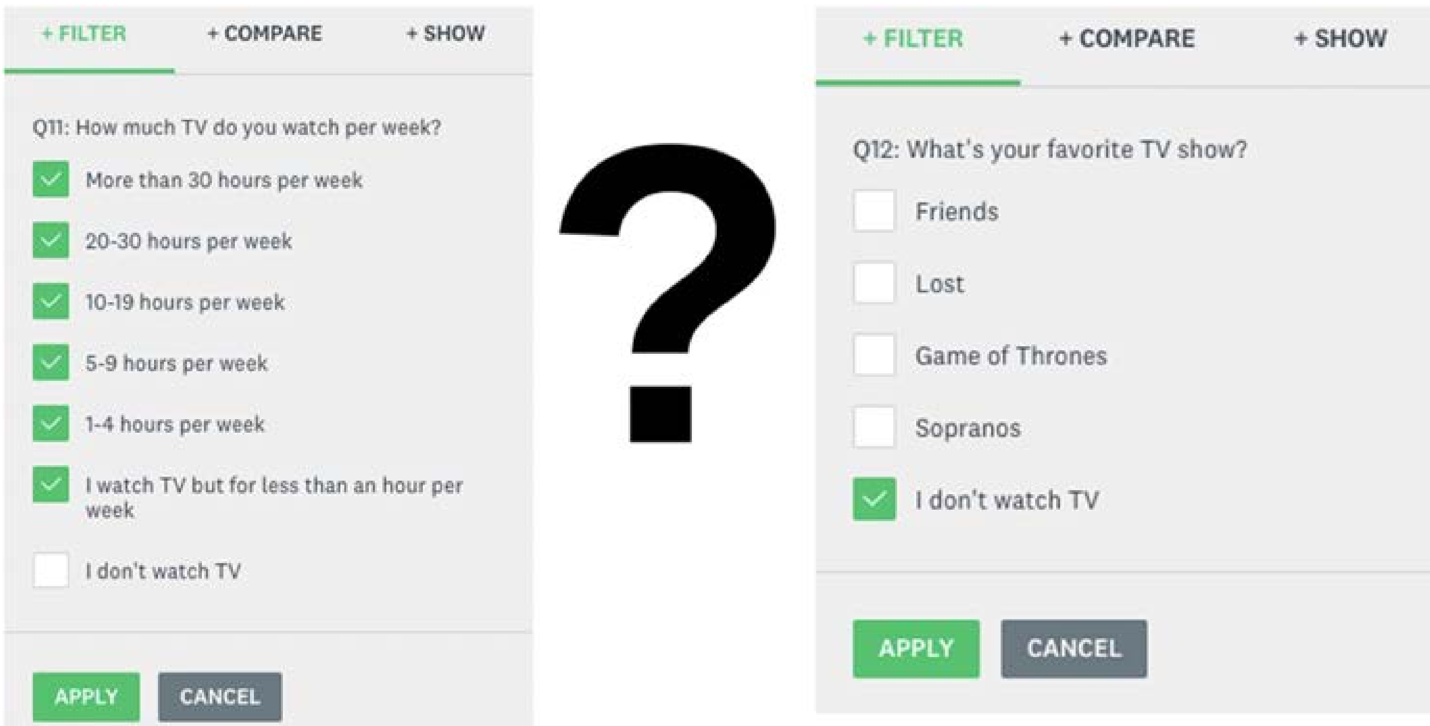
- Respondents who speed through your survey



Example using Data cleansing in stages

SurveyMonkey.com Data Cleaning contd…

- Respondents who give inconsistent responses



Challenges of data cleaning

- Data cleaning solutions can have several problems during the process. You needs to understand the various problems and figure out how to tackle them.
  - Ongoing maintenance can be expensive and time-consuming
  - Limited knowledge about what is causing anomalies, creating ifficulties in creating the right transformations
  - Privacy and Security
  - Data deletion, where a loss of information leads to incomplete data that cannot be accurately ‘filled in’
  - It is difficult to build a data cleansing graph to assist with the process ahead of time

Overall Benefits of Data Cleaning

- Data Quality, that help increase your efficiency and speed up the decision-making process.
- You can better monitor your errors to help you eliminate incorrect, corrupt, or inconsistent data.
- You will make fewer errors overall.
- You can map different functions and what your data should do.
- It’s easy to remove errors across multiple data sources.

In Summary

- One of the most interesting things about data in this era is its ease of accessibility-online through social media, search engines, websites, etc.
- Most of the data is either incorrect or full of irrelevancies. In order to leverage on the easily accessible huge data, we need to take our time to clean it.
- Data cleaning is arguably one of the most important steps towards achieving great results from the data analysis process.
- If the data isn’t cleaned, data analysis will not yield a perfect result.


THANK YOU
for listening!

ndabas@ub.ac.bw

simisani.ndaba013@gmail.com