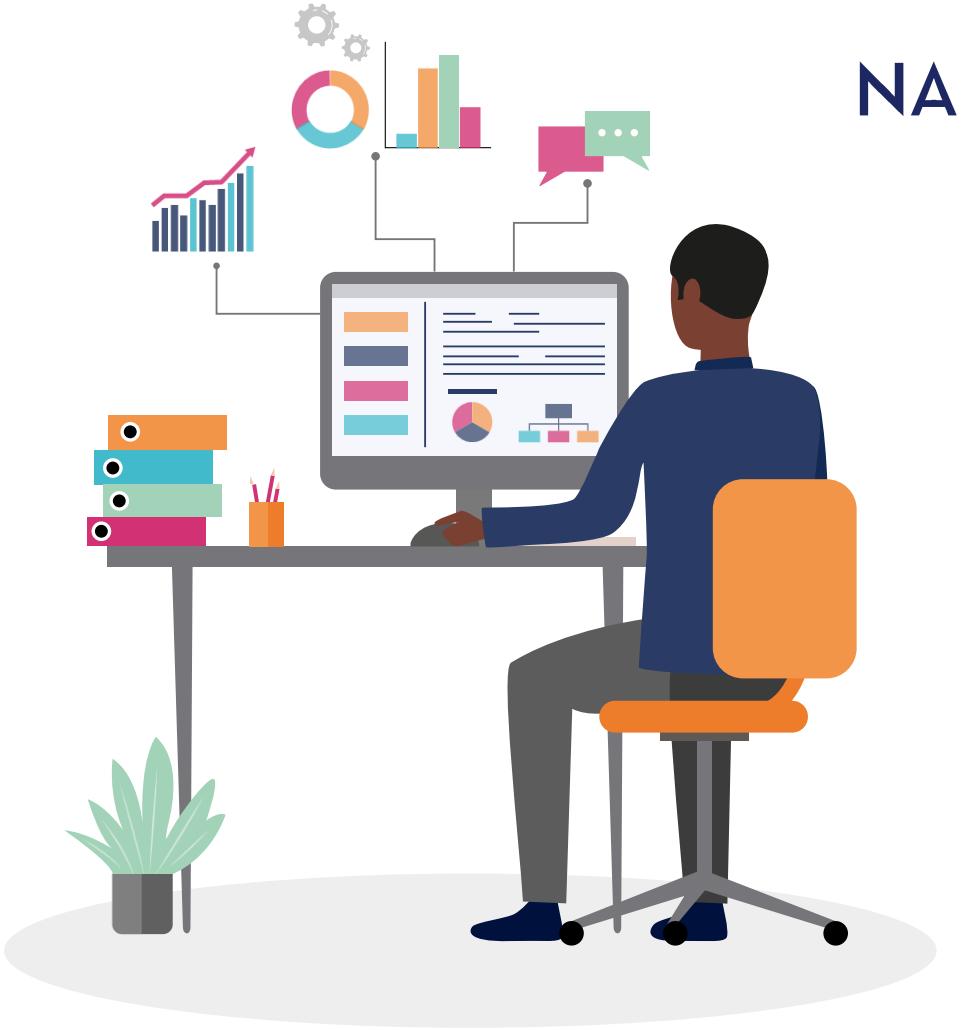


NASCEE Learning Event: Data Cleaning

04 April 2023

Data Artistry.

NASCEE
National Association of Social Change Entities in Education

Capability programme by Data Innovators

What is data vs information?

DATA

Data is raw, unorganised facts that need to be processed. Data can be something simple and seemingly random and useless until it is organised.

INFORMATION

When data is processed, organised, structured or presented in a given context so as to make it useful, it is called information,

Raw data alone is insufficient to make decisions, but information is sufficient to make a decision.

Source: https://byjus.com/biology/difference-between-data-and-information/

Data Artistry.

What is data vs information?

Data

Information

https://www.tsl.texas.gov/slrm/blog/2022/05/the-relationships-between-data-information-and-records/

Data Artistry.

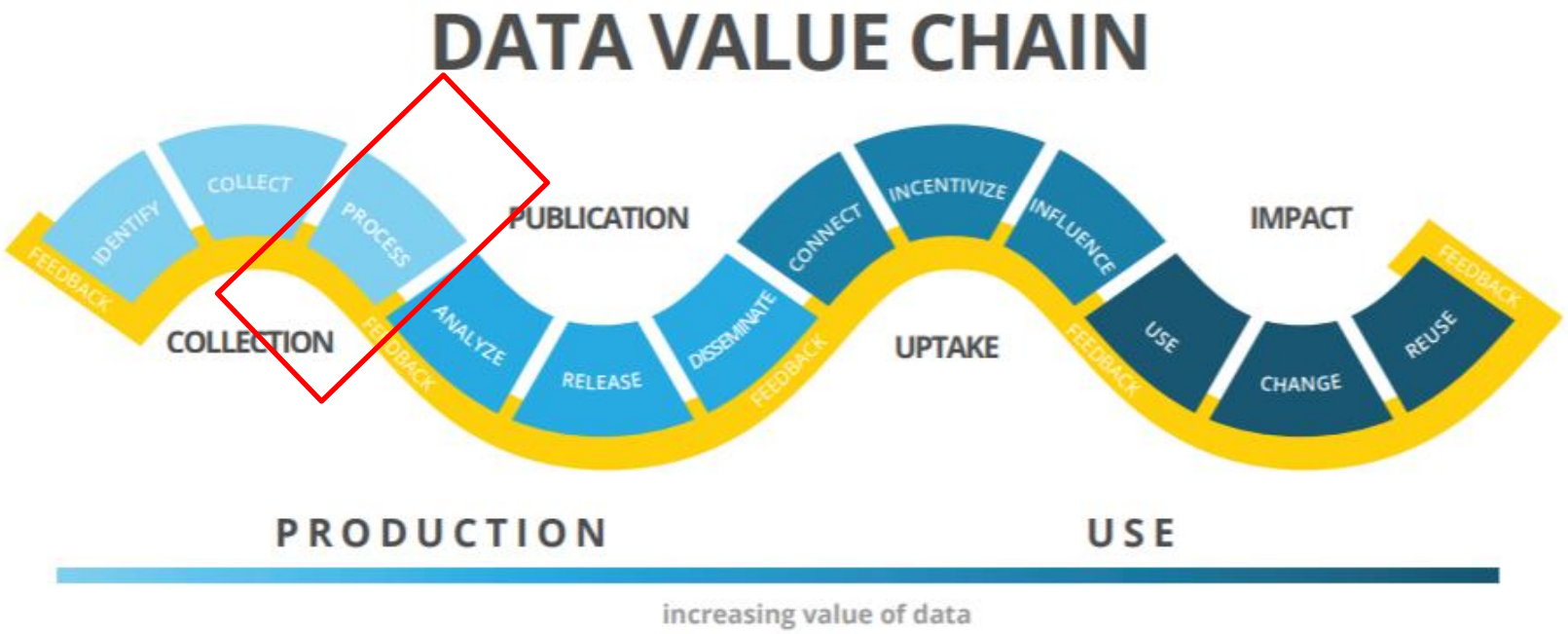
What is data vs information?

This are phone numbers of people to contact for interviews

100239
100240
100241
100242
100243
100244
100245
100246
100247
100248
100249

Data Artistry.

https://www.tsl.texas.gov/slrm/blog/2022/05/the-relationships-between-data-information-and-records/



Source: The Data Value Chain: Moving from Production to Impact Prepared for Data2X by Open Data Watch

Data Artistry.

Why data cleaning/Where data cleaning fits in…..

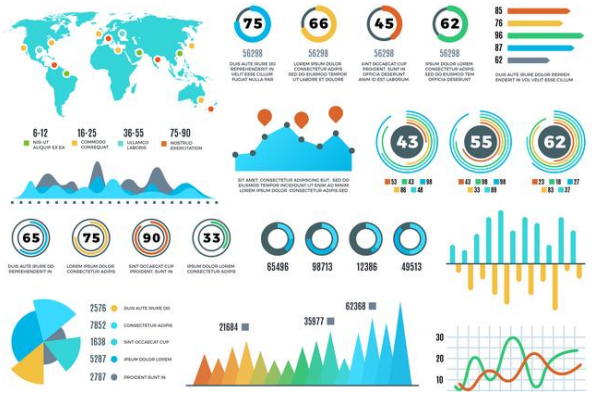
From chaos

Segment>>	Consumer				Consumer Total
Ship Mode>>	First Class	Same Day	Second Class	Standard Class	
Order ID					
CA-2011-100293					
CA-2011-100706			129.44		129.44
CA-2011-100895				605.47	605.47
CA-2011-100916					
CA-2011-101266			13.36		13.36
CA-2011-101560					
CA-2011-101770					
CA-2011-102274					
CA-2011-102673					
CA-2011-102988					
CA-2011-103317					
CA-2011-103366	149.95				149.95
CA-2011-103807					
CA-2011-103989					
CA-2011-104283				616.14	616.14

Data cleaning

Comes order

Segment	Ship Mode	OrderID	Sales
Consumer	First Class	CA-2011-103366	149.95
Consumer	First Class	CA-2011-109043	243.6
Consumer	First Class	CA-2011-113166	9.568
Consumer	First Class	CA-2011-124023	8.96
Consumer	First Class	CA-2011-130155	34.2
Consumer	First Class	CA-2011-136861	31.984
Consumer	First Class	CA-2011-153927	286.65
Consumer	First Class	CA-2011-157784	514.03
Consumer	First Class	CA-2011-160094	1000.95
Consumer	First Class	CA-2011-164749	9.912
Consumer	First Class	CA-2011-166730	39.128
Consumer	First Class	CA-2012-102722	106.5
Consumer	First Class	CA-2012-102778	18.176
Consumer	First Class	CA-2012-117828	194.32
Consumer	First Class	CA-2012-130218	59.48
Consumer	First Class	CA-2012-132318	182.91
Consumer	First Class	CA-2012-137974	2298.9



https://www.alamy.com/stock-photo-cartoon-of-businessman-with-piles-of-paper-on-his-desk-it-started-79324192.html

https://www.google.com/url?sa=i&url=https%3A%2F%2Fneilpatel.com%2Fblog%2Fdata-visualization%2F&psig=AOvVaw0_TXGsSxzgpZt_Cu_7Fy-A&ust=1674277813154000&source=images&cd=vfe&ved=0CBEQjhxqFwoTCOCZzoSx1fwCFQAAAAAdAAAAABAE

https://foresightbi.com.ng/microsoft-power-bi/dirty-data-samples-to-practice-on/

Data Artistry.

What is data cleaning?

Source: https://excelkid.com/15-ways-to-clean-data-in-excel/

Data cleaning is a process of detecting, correcting, replacing, modifying or removing messy data from a record set, table or database.

Source: Data Cleaning. (2016). ACAPS.

Data Artistry.

Common data errors

Before analysing the data, it is important to ensure that it is accurate and consistent. Data error causes can range from;

Erroneous entry e.g. typing age 34% instead of 34

Extraneous entries but needed data e.g. typing name and title in a name-only field

Inconsistencies across files e.g. mismatch between what is on a learner report and data captured on the system

Data Artistry.

Source: Data Cleaning. (2016). ACAPS.

Data quality

"Data quality refers to the development and implementation of activities that apply quality management techniques to data in order to ensure the data is fit to serve the specific needs of an organization in a particular context. Data that is deemed fit for its intended purpose is considered high quality data."

Source: https://www.heavy.ai/technical-glossary/data-quality

Data Artistry.

Determining Data Quality

Validity

The degree to which the data follows the rules of requirement e.g. Do we only see IDs under the Learner ID column?

Accuracy

Establishing whether data provided is correct e.g. if a Learner ID has 6 digits, are there IDs with less?

Consistency

Is the data provided consistent with the data source e.g. if Grade 1 is not offered Maths, then there shouldn't be maths data for Grade 1 learners

Completeness

The degree to which all required data fields are provided e.g. are there any missing values in your data?

Source: https://www.upgrad.com/blog/data-cleaning-techniques/

Data Artistry.

Implications of working with unclean data

A teacher wants to plan a field trip, they ask the school principal for data on the students that have signed up so that they plan accordingly.

Inconsistency

Gender	Student Name	Payment
Female	Teboho Letuka	R50
Male	John Doe	R500
Male	Harry Potter	R50
Female	Masego Tabane	R50
Female	Noluthando Mqalekane	R50
Male	Tau Mogale	R50
Female	Teboho Letuka	R50
Female	Masego Tabane	R50

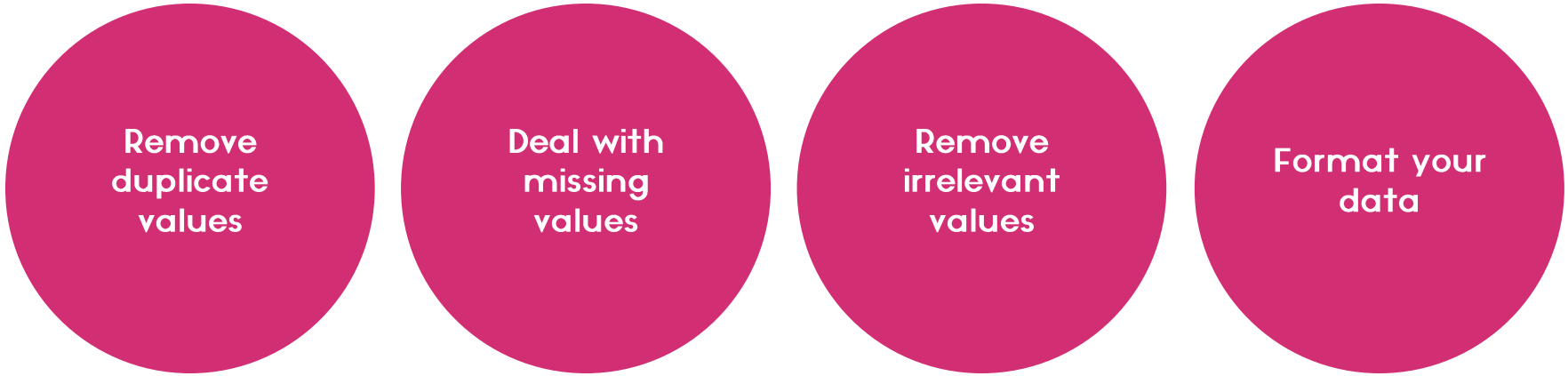
Duplicates

Total students attending	Total Amount
8	R850

Given the data received, the teacher orders 8 lunch packs for the value of R850 for the students that will be attending.

Data Artistry.

Data Cleaning Techniques



Data Artistry.

Source: https://www.upgrad.com/blog/data-cleaning-techniques/

Identifying duplicates

Gender	Student Name	Age	Payment
Female	Teboho Letuka	16	50
Male	John Doe	14	50
Male	Harry Potter	15	50
Female	Masego Tabane	14	50
Female	Noluthando Mqalekane	13	50
Male	Tau Mogale	16	50
Female	Teboho Letuka	13	50
Female	Masego Tabane	14	50

Data Artistry.

Source: https://www.upgrad.com/blog/data-cleaning-techniques/

Dealing with missing values

Gender	Student Name	Age	Payment	Parental Consent
Female	Teboho Letuka	16	50	Yes
Male	John Doe	14	500	Yes
	Harry Potter	15	50	Yes
Female	Masego Tabane	14	50	Yes
Female	Noluthando Mqalekane	13	50	Yes
Male	Tau Mogale	16	50	
Female	Teboho Letuka	13	50	Yes
Female		14		

Data Artistry.

Source: https://www.upgrad.com/blog/data-cleaning-techniques/

Remove irrelevant values

Gender	Student Name	Age	Payment	Parental Consent	Time paid
Female	Teboho Letuka	16	50	Yes	Monday 9am
Male	John Doe	14	50	Yes	Monday 10am
Male	Harry Potter	15	50	Yes	Tuesday 2pm
Female	Masego Tabane	14	50	Yes	Tuesday 4pm
Female	Noluthando Mqalekane	13	50	Yes	Tuesday 4pm
Male	Tau Mogale	16	50	Yes	Friday 8am
Female	Teboho Letuka	13	50	Yes	Friday 2pm

Data Artistry.

Source: https://www.upgrad.com/blog/data-cleaning-techniques/

Formatting data

Gender	Student Name	Age	Payment	Parental Consent
Female	Teboho Letuka	16	50	Yes
Male	John Doe	14	50	Yes
Male	HARRY POTTER	15	50	Yes
Female	Masego Tabane	14	50	Yes
Fmale	Noluthando Mqalekane	13	50	Yes
male	Tau Mogale	16	50	Yes
Female	Teboho Letuka	13	50	Yes

Data Artistry.

Source: https://www.upgrad.com/blog/data-cleaning-techniques/

Thank You!

Data Artistry.

Data Artistry.