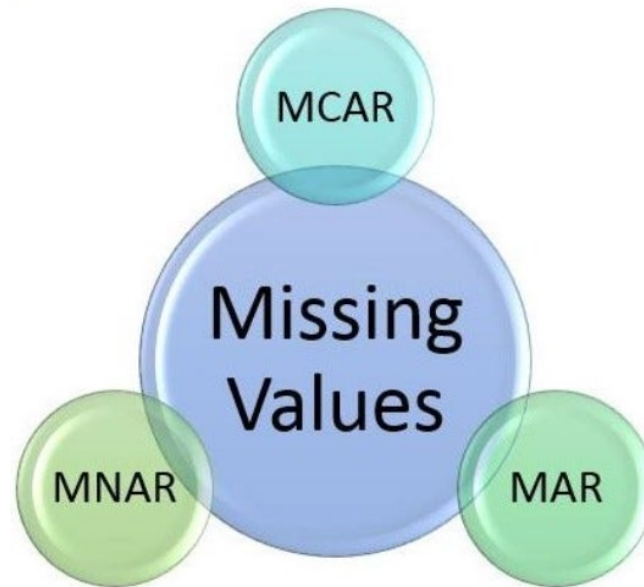


Practical 2

After Class Activity

MCAR, MAR, MNAR - How do I tell them apart?



Explore these 3 websites to discover the difference between MCAR, MAR and MNAR:

<https://www.missingdata.nl/missing-data/missing-data-mechanisms/mar/>

<https://www.missingdata.nl/missing-data/missing-data-mechanisms/mcar/>

<https://www.missingdata.nl/missing-data/missing-data-mechanisms/mnar/>



Type of Missingness : Question 1

Match the following :

Missing At Random (MAR)


Missing data is unrelated to other variable(s) and unrelated to the variable with missing values itself.

Missing Completely At Random (MCAR)

Missing data is related to the variable with missing values itself.

Missing Not At Random (MNAR)

Missing data is related to other variable(s).



Type of Missingness : Question 2

The left table has missing values, while the right table is what the complete data would probably look like based on domain knowledge.

With missing values

ID	Gender	Rating
1	Male	6
2	Male	
3	Female	1
4	Male	4
5	Female	5
6	Female	9
7	Male	
8	Female	4
9	Female	7
10	Male	

Complete Data

ID	Gender	Rating
1	Male	6
2	Male	2
3	Female	1
4	Male	4
5	Female	5
6	Female	9
7	Male	3
8	Female	4
9	Female	7
10	Male	8

What type of missing data is this?

- a. Missing Completely At Random (MCAR)
- b. Missing At Random (MAR)
- c. Missing Not At Random (MNAR)



Multiple Choice

Type of Missingness : Question 3

The left table has missing values, while the right table is what the complete data would probably look like based on domain knowledge.

With missing values

ID	Gender	Rating
1	Male	6
2	Male	2
3	Female	1
4	Male	4
5	Female	5
6	Female	
7	Male	3
8	Female	4
9	Female	7
10	Male	

Complete Data

ID	Gender	Rating
1	Male	6
2	Male	2
3	Female	1
4	Male	4
5	Female	5
6	Female	9
7	Male	3
8	Female	4
9	Female	7
10	Male	8

What type of missing data is this?

- a. Missing Completely At Random (MCAR)
- b. Missing At Random (MAR)
- c. Missing Not At Random (MNAR)



Multiple Choice

<https://www.inwt-statistics.com/blog/understanding-and-handling-missing-data>

Type of Missingness : Question 4

The left table has missing values, while the right table is what the complete data would probably look like based on domain knowledge.

With missing values

ID	Gender	Rating
1	Male	6
2	Male	2
3	Female	1
4	Male	
5	Female	5
6	Female	9
7	Male	3
8	Female	4
9	Female	
10	Male	8

Complete Data

ID	Gender	Rating
1	Male	6
2	Male	2
3	Female	1
4	Male	4
5	Female	5
6	Female	9
7	Male	3
8	Female	4
9	Female	7
10	Male	8

What type of missing data is this?

- a. Missing Completely At Random (MCAR)
- b. Missing At Random (MAR)
- c. Missing Not At Random (MNAR)



Multiple Choice

<https://www.inwt-statistics.com/blog/understanding-and-handling-missing-data>

Data Profiling



<https://www.youtube.com/watch?v=HtaYjVwW-Mo&t=4s>



P02a - Data Management: Hands-on Additional Qns

- On LMS, head to “Activity 1: Superstore Dataset” > “Practical Questions”. Download practical sheets and datasets.
- We will learn to do :
 - Data Profiling - look at data types and how to inspect data for quality issues
 - Data Cleaning - how to handle the data quality issues found earlier when profiling



Data Profiling #1

What do you observe for these columns?

Column	Data Type	Range/Values	Any Observations?
Ship mode			
Segment	nominal/string	Consumer; Corporate; Home Office; Hm Off (4)	inconsistency
Country	geographical	United States; US; Blanks (3)	Missing values
City			
State			
Postal code			
Region			
Category			
Sub-category			



Data Profiling #2

What do you observe for these columns?

Column	Data Type	Range/Values	Any Observations?
Sales			
Cost			
Quantity			
Discount			



Data Profiling #3

What do you observe for these columns?

Column	Data Type	Range/Values	Any Observations?
Order Date			
Ship Date			

Column	Data Type	Range/Values	Any Observations?
Order ID	nominal/string	e.g. CA-2011-100006 (5009)	Need to split State-Year-Order ID
Customer ID			
Customer Name			
Product ID			
Product Name			

DIY Cleaning



1. There are some inconsistencies in **State** (TX vs Texas, CA vs California). Rename the original column as **State_Old**, then create a new column called **State**. What is the missing line in this pseudo-code?

```
If [State_Old] = "CA" Then "California"  
...  
Else [State_Old]  
End
```

- | | |
|---|---------------------|
| a. Else If [State_Old] equals "TX" Then "Texas" | b. Else [State_Old] |
| c. Else If [State_Old] equals "Texas" Then "TX" | d. Else [Others] |

DIY Cleaning



2. Perform a split on the **Order ID** Column. How many **unique** years were split from the Order ID column?

a. 3

b. 4

c. 5

d. 2

3. Which is the **longest duration** between **Order Date** and **Ship Date**?
Add a column and use a formula to find out.

a. 1 day

b. 3 days

c. 5 days

d. 7 days

What else can we do for Data Cleaning?

Technically speaking, when trying to detect anomalies, we briefly mentioned in an earlier video about some charts that can be used for categorical vs numeric variables. This part belongs is an overlapping zone between Data Preparation and Data Exploration. We will delve more into usage of charts to detect anomalies in the next section, together with other Data Exploration techniques.

Data Cleaning : Handling Anomalies

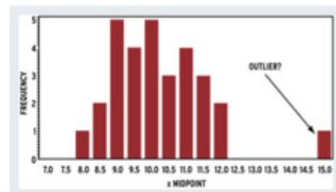
How To Detect Them

- For Categorical Data - Frequency Table or Bar Chart

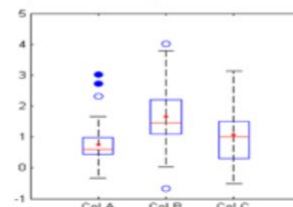
REGION	SALES
North	\$29034
South	\$56728
Soth	\$12000
East	\$27000
West	\$89092



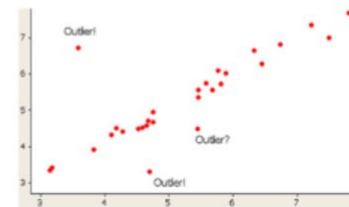
- For Numeric Data - Histogram, Box Plot, Scatter Plot



<https://cxl.com/blog/outliers/>



<https://wiki.eigenvector.com/index.php?title=Boxplot>



<https://apandre.files.wordpress.com/2011/08/outlier2a.jpg>

Metadata and Data Standards and Infocomm Media Development Authority (IMDA) Singapore

Metadata and Data Standards



<https://www.youtube.com/watch?v=Xmp8vNeFस्क>

<https://www.statcan.gc.ca/en/wtc/data-literacy/catalogue/892000062021006>



Data Standards

What : Rules used to standardize the way data are described, represented and structured (e.g. a controlled list that user can select.)

Why : Because what you do with your data can make it easier to work with downstream.

Types of Data Standards

- Data format standards

- Examples: dates, negative numbers, currency, 2-letter Canadian province codes

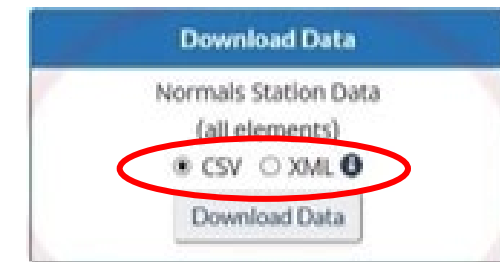
- Data file format standards

- Example: Comma Separated Variable (CSV) files

- Variable standards

- Examples: standardized categories and structures for employment status, age groups, industry, occupation, products

	Jan	Feb	Mar
Temperature			
Daily Average (°C)	-13.7	-10.6	-3.9
Standard Deviation	4.5	3.9	3.0
Daily Maximum (°C)	-9.0	-5.8	0.9
Daily Minimum (°C)	-18.3	-15.4	-8.6
Extreme Maximum (°C)	11.0	9.5	19.5
Date (yyyy/dd)	1980/23	1988/26	1993/23



Example
Research A uses 4 age categories.
Research B uses 5 age categories.
Is that standardized categories?

<https://www.statcan.gc.ca/en/wtc/data-literacy/catalogue/892000062021006>




Metadata

What : Extra info provided about the data

- Source (what, who, why, when, where) of the data
- Data quality
- Methods used to process the data
- **Any Data Standards applied or followed**
- How data are grouped for more efficient search (via hashtags, key words)

Why : So that it makes it easier to find, interpret, trust and use the data, especially when shared across multiple systems.

- 
- Previously, Metadata is simply defined as : Data about data, Data dictionary
 - **Now, Metadata also includes business terms, explanation and usage.**
This is to better control and manage business info.

Example of Reference Metadata

- Reference metadata
 - Examples: who collected it, when, for what purpose; methods; quality
- Descriptive metadata
 - Examples: titles, footnotes, labels on data visualizations
- Structural metadata
 - Examples: variables, classifications, identifiers, valid values, code lists

Canadian Climate Normals 1981-2010 Station Data

Temperature and Precipitation Graph

Normals Data

Station / Element Metadata

The minimum number of years used to calculate these Normals is indicated by a code for each element. A "+" beside an extreme date indicates that this date is the first occurrence of the extreme value. Values and dates in bold indicate all-time extremes for the location.

Data used in the calculation of these Normals may be subject to further quality assurance checks. This may result in minor changes to some values presented here.

SASKATOON WATER TP *

SASKATCHEWAN

Latitude: 52°07'00.000" N **Longitude:** 106°41'00.000" W **Elevation:** 483.10 m

Climate ID: 4057202

WMO ID:

TC ID:

* This station meets WMO standards for temperature and precipitation.

Related Data

[Calculation Information](#)

[Station / Element Metadata](#)

[1971-2000 Climate Normals](#)

Additional Search Options

[Nearby Stations with Data](#)

Download Data

Normals Station Data
(all elements)

☒ CSV ☐ XML 

[Download Data](#)

<https://www.statcan.gc.ca/en/wtc/data-literacy/catalogue/892000062021006>

Example of Descriptive Metadata

- Reference metadata

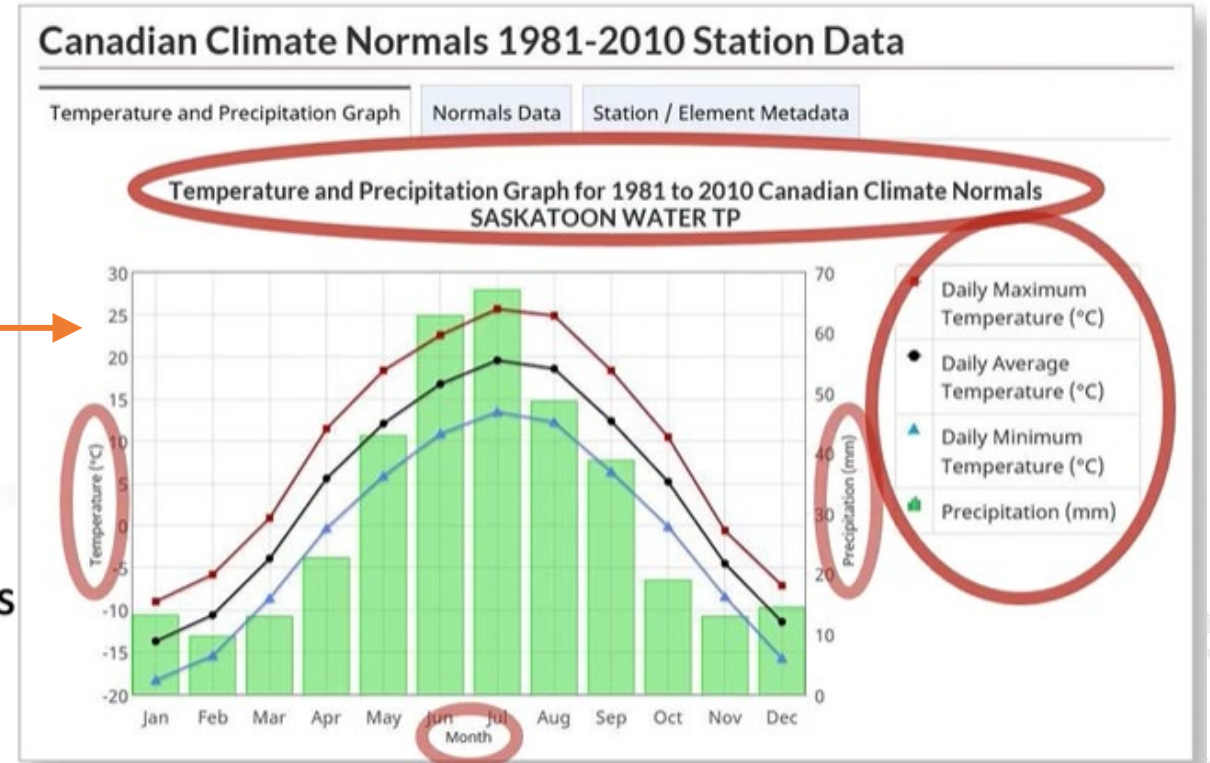
- Examples: who collected it, when, for what purpose; methods; quality

- Descriptive metadata

- Examples: titles, footnotes, labels on data visualizations

- Structural metadata

- Examples: variables, classifications, identifiers valid values, code lists



<https://www.statcan.gc.ca/en/wtc/data-literacy/catalogue/892000062021006>

Example of Structural Metadata

- Reference metadata


- Examples: who collected it, when, for what purpose; methods; quality

- Descriptive metadata

- Examples: titles, footnotes, labels on data visualizations

- Structural metadata

- Examples: variables, classifications, identifiers, valid values, code lists



S/N	Variable Name	Variable Label	Data Type	Codes	Remarks
1	ID	Customer Identification Number	Character		
2	Gen	Gender	Character	F='Females', M='Male'	
3	Occ	Occupation	Character		
4	Inc	Income	Numeric		
5	Age	Age	Numeric		



Why Data Standards and Metadata are Important

Data standards and metadata enable data to be FAIR...

- **F**indable: easily searchable
- **A**ccessible: easy to use
- **I**nteroperable: easily combined
- **R**e-useable: easy to share

Important for sharing data across the entire organisation, multiple systems or government agencies.

<https://www.statcan.gc.ca/en/wtc/data-literacy/catalogue/892000062021006>

IM(ICT&SS) and Importance of Metadata/Data Standards

Singapore Context



Data Acquisition, which includes Data Minimisation where agencies are not to collect data in excess to minimise the risks due to unauthorised use and disclosure; use of WOG Data Platforms to obtain data for their use cases; maintain good Data Quality by ensuring that data is accurate, consistent, timely, relevant and complete; ensure Data Discoverability by maintaining accurate metadata and making these available for search is a key way to make data discoverable; and comply with Data Storage and Retention Requirements by retaining data only for the period necessary for the fulfilment of the purposes.

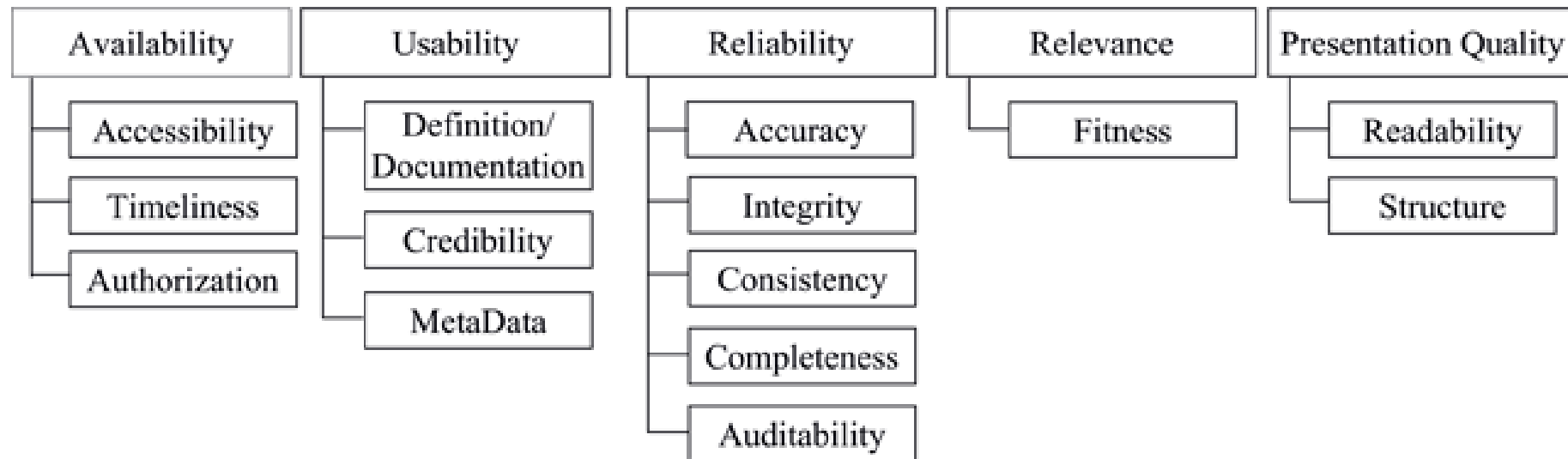
Data Processing and Fusion, which includes minimising errors arising from data processing, such as coding errors, data entry errors, computation errors, and minimising the risk of unauthorised re-identification of individuals or entities through fusion or integration of de-identified datasets.

IM(ICT&SS) : Instruction Manual for Infocomm Technology and Smart Systems (ICT&SS) Management (previously known as IM8)

<https://www.developer.tech.gov.sg/guidelines/standards-and-best-practices/instruction-manual-for-ict-ss-management.html>

Data Quality Dimensions and Infocomm Media Development Authority (IMDA) Singapore

Data Quality Dimensions



<https://datascience.codata.org/articles/10.5334/dsj-2015-002/#T1>

IM (ICT&SS) and the Importance of Data Quality Dimensions

Singapore Context

Accuracy
Timeliness
Completeness
Consistency
Relevance

A Singapore Government Agency Website [How to identify](#) ✓

 **Singapore Government
Developer Portal**

[Our Digital Journey](#) ✓

[Guidelines](#) ✓

[Products](#)

[SG Tech Stack](#)

[Communities](#) ✓

[Documentation](#)

Have feedback? Please [let us know](#).

[HOME](#) / [GUIDELINES](#) / [STANDARDS AND BEST PRACTICES](#)
/ [INSTRUCTION MANUAL FOR ICT SS MANAGEMENT](#)

Instruction Manual for Infocomm Technology and Smart Systems (ICT&SS) Management

Data Acquisition, which includes Data Minimisation where agencies are not to collect data in excess to minimise the risks due to unauthorised use and disclosure; use of WOG Data Platforms to obtain data for their use cases; maintain good Data Quality by ensuring that data is accurate, consistent, timely, relevant and complete; ensure Data Discoverability by maintaining accurate metadata and making these available for search is a key way to make data discoverable; and comply with Data Storage and Retention Requirements by retaining data only for the period necessary for the fulfilment of the purposes.

<https://www.developer.tech.gov.sg/guidelines/standards-and-best-practices/instruction-manual-for-ict-ss-management.html>



Data Quality Dimensions

Accuracy – Is data error-free and reliable?

Consistency – Is the data format standardized?

Completeness – Are all the data elements populated in the system?

Relevance - Does it meet business needs?

Timeliness - Is data available at a time when it is still meaningful?