Temasek POLYTECHNIC | School of Informatics & IT

Practical 2
Data Standards and Metadata
Data Quality Dimensions & Data Cleaning
Data Integration + Intro to Tableau

Practical 2

CLASS SCHEDULE

Before Class	**Concepts**	Data Attributes Data Quality Dimension
During Class	**Hands-on**	Recap Unpivoting Data Profiling Data Cleaning Data Standards and Meta Data Data Quality Dimension
After Class	**Hands-on**	*Revise today's class with LMS: Apr/Oct – Week X (…)* *Go through Additional Resources slides* - *Explore MCAR, MAR, MNAR* - *Watch video on "What is Data Profiling"* - *Explore further in Data Profiling and Data Cleaning* - *Read up further on Data Standards*

How To Do Data Visualisation?

Identify Business Problem → Data Preparation → Data Exploration → Data Visualisation

Brainstorming

Integration
Profiling
Cleaning

Making Sense
Discovering Insights

Communicating

Temasek POLYTECHNIC | School of Informatics & IT

What and Why of Data Prep

Music from https://pixabay.com/music/introoutro-beautiful-piano-logo-161975/

Why Data Preparation and What is it?

Raw data may potentially contain lots of untidy data due to data entry or measurement errors because of missing data, inconsistencies, fragmented data in many different tables

Data preparation allows us to shape the data to the desired quality.

Steps Taken to shape the data to the desired quality

How To Do Data Preparation?

	Integration	Profiling	Cleaning
What			
Why	Integrate data tables Transform to right shape	Get an overview of data Assess data quality	Improve data quality
How	Using Join Using Relationship Unpivoting	Structure of data Quality of data (anomalies, missing values, Inconsistencies)	Filter / Replace Split column Remove columns Convert data type Add new columns etc.

Temasek POLYTECHNIC | School of Informatics & IT

Why the Need to Unpivot Your Data

Music from https://pixabay.com/music/introoutro-beautiful-piano-logo-161975/

Which is easier to read?

Table 1

ID	Training	2014	2015	2016	2017	2018
A1	Team	27	25	31	42	20
A1	Safety	10	4	8	12	5
A1	Soft	43	55	62	31	47
A1	Tech	20	50	43	28	33
A2	Team	31	27	40	33	35
A2	Safety	8	9	2	3	7
A2	Soft	28	53	31	40	33
A2	Tech	72	46	56	62	84

Table 2

Year	ID	Training	Hours
2014	A1	Team	27
2015	A1	Team	25
2016	A1	Team	31
2017	A1	Team	42
2018	A1	Team	20
2014	A1	Safety	10
2015	A1	Safety	4
2016	A1	Safety	8
2017	A1	Safety	12
2018	A1	Safety	5
2014	A1	Soft	43
2015	A1	Soft	55
2016	A1	Soft	62
2017	A1	Soft	31
2018	A1	Soft	47
2014	A1	Tech	20
2015	A1	Tech	50
2016	A1	Tech	43
2017	A1	Tech	28
2018	A1	Tech	33
2014	A2	Team	31
2015	A2	Team	27
2016	A2	Team	40
2017	A2	Team	33
2018	A2	Team	35

Tidy Data Structure

- Each variable you measure should be in one column.
- Each observation should be in a different row.
- There should be one table for each "kind" of variable.
- If you have multiple tables, they should include a column in the table that allows them to be linked.

Wickham, H. . (2014). Tidy Data. *Journal of Statistical Software*, *59*(10), 1–23. https://doi.org/10.18637/jss.v059.i10

For developers - Easier to develop tools for data analysis
For users - Easier to manipulate, model and visualise

Year	ID	Training	Hours
2014	A1	Team	27
2015	A1	Team	25
2016	A1	Team	31
2017	A1	Team	42
2018	A1	Team	20
2014	A1	Safety	10
2015	A1	Safety	4
2016	A1	Safety	8
2017	A1	Safety	12
2018	A1	Safety	5
2014	A1	Soft	43
2015	A1	Soft	55
2016	A1	Soft	62
2017	A1	Soft	31
2018	A1	Soft	47
2014	A1	Tech	20
2015	A1	Tech	50
2016	A1	Tech	43
2017	A1	Tech	28
2018	A1	Tech	33
2014	A2	Team	31
2015	A2	Team	27
2016	A2	Team	40
2017	A2	Team	33
2018	A2	Team	35
2014	A2	Safety	8
2015	A2	Safety	9
2016	A2	Safety	2
2017	A2	Safety	3

Melt Data (Un-Pivoting)

If data is originally Short and Wide

Need to transform to the Tall and Narrow format

ID	Training	2014	2015	2016	2017	2018
A1	Team	27	25	31	42	20
A1	Safety	10	4	8	12	5
A1	Soft	43	55	62	31	47
A1	Tech	20	50	43	28	33
A2	Team	31	27	40	33	35
A2	Safety	8	9	2	3	7
A2	Soft	28	53	31	40	33
A2	Tech	72	46	56	62	84

Year	ID	Training	Hours
2014	A1	Team	27
2015	A1	Team	25
2016	A1	Team	31
2017	A1	Team	42
2018	A1	Team	20
2014	A1	Safety	10
2015	A1	Safety	4
2016	A1	Safety	8
2017	A1	Safety	12
2018	A1	Safety	5
2014	A1	Soft	43
2015	A1	Soft	55
2016	A1	Soft	62
2017	A1	Soft	31
2018	A1	Soft	47
2014	A1	Tech	20
2015	A1	Tech	50
2016	A1	Tech	43
2017	A1	Tech	28
2018	A1	Tech	33
2014	A2	Team	31
2015	A2	Team	27

Aggregation (Group By / Pivot)

Once data is in Tidy Data Structure

You can do Aggregation

Year	ID	Training	Hours
2014	A1	Team	27
2015	A1	Team	25
2016	A1	Team	31
2017	A1	Team	42
2018	A1	Team	20
2014	A1	Safety	10
2015	A1	Safety	4
2016	A1	Safety	8
2017	A1	Safety	12
2018	A1	Safety	5
2014	A1	Soft	43
2015	A1	Soft	55
2016	A1	Soft	62
2017	A1	Soft	31
2018	A1	Soft	47
2014	A1	Tech	20
2015	A1	Tech	50
2016	A1	Tech	43
2017	A1	Tech	28
2018	A1	Tech	33
2014	A2	Team	31
2015	A2	Team	27

Year	Sum
2014	239
2015	269
2016	273
2017	251
2018	264

Training	Sum
Safety	68
Soft	423
Team	311
Tech	494

Training	Year	Sum
Safety	2014	18
Safety	2015	13
Safety	2016	10
Safety	2017	15
Safety	2018	12
Soft	2014	71
Soft	2015	108
Soft	2016	93
Soft	2017	71
Soft	2018	80
Team	2014	58
Team	2015	52
Team	2016	71
Team	2017	75
Team	2018	55
Tech	2014	92
Tech	2015	96
Tech	2016	99
Tech	2017	90
Tech	2018	117

P02a - Unpivot – Marriage: Hands-on

- **On LMS, head to "Activity 1: Unpivot" > "Practical Questions". Download practical sheets and datasets.**
- The dataset captures info on the marriages between brides and grooms of various ethnic groups across different years. We will learn to do:
  - Data Preparation : Unpivot data

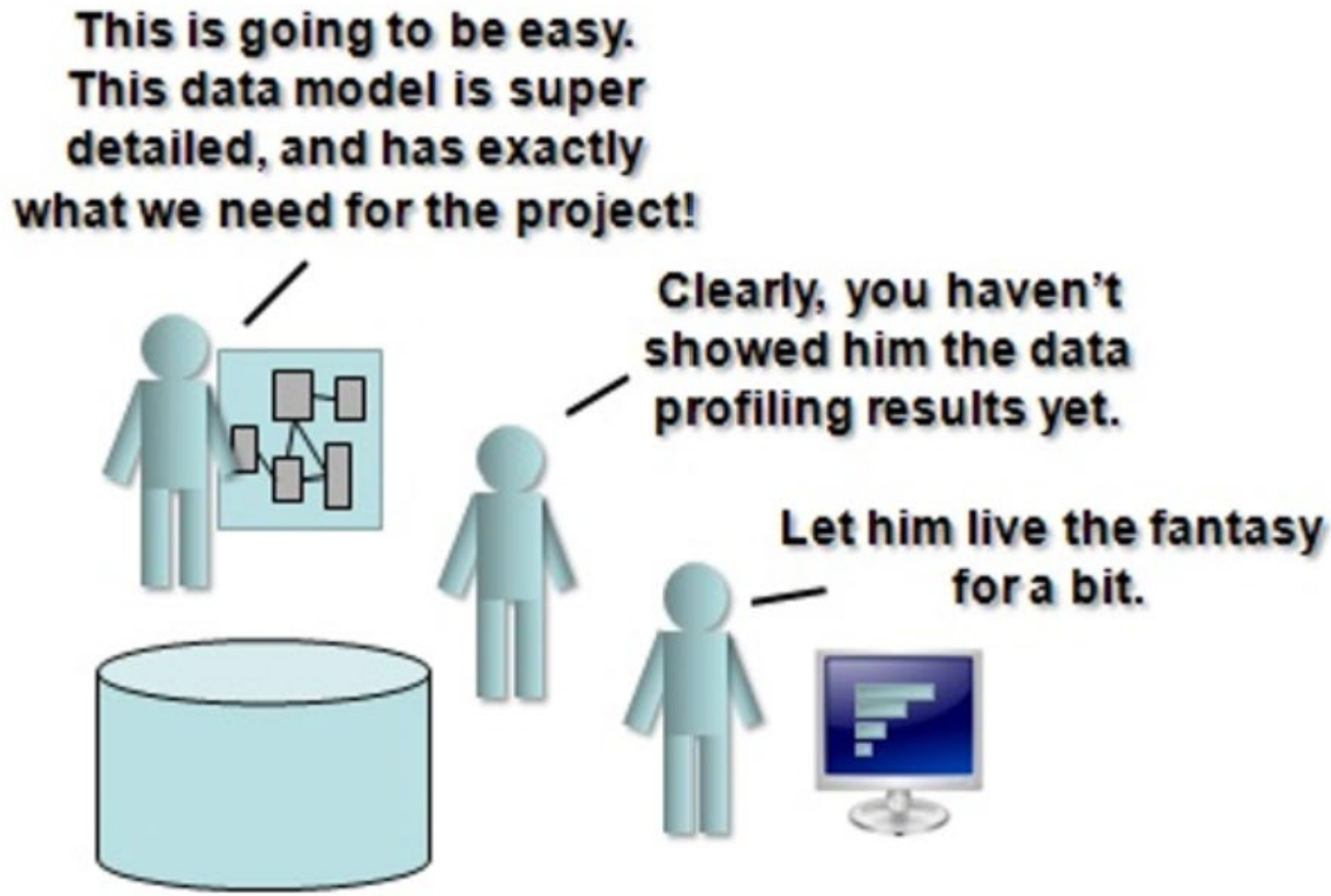
Takeaways

- What is the Tidy Data structure
- What it means to unpivot your data
- Easier to manipulate, model and visualise your data

Temasek POLYTECHNIC | School of Informatics & IT

Data Profiling – Explore
Intro to Superstore Dataset

Data Profiling is a "reality check" of your data. How is that so?



http://www.datamartist.com/what-is-data-profiling-and-when-and-why-should-it-be-done

Data Profiling

With data model, you only know, at a high level, what is ***supposed to be*** in the columns.

But do you know what is ***actually*** in the columns? Are the contents what you expected them to be?

Data Profiling brings you down a path of detective work and allows you to see the reality of your data.

P02b - Data Management: Hands-on

- **On LMS, head to "Activity 2: Data Management" > "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

Context

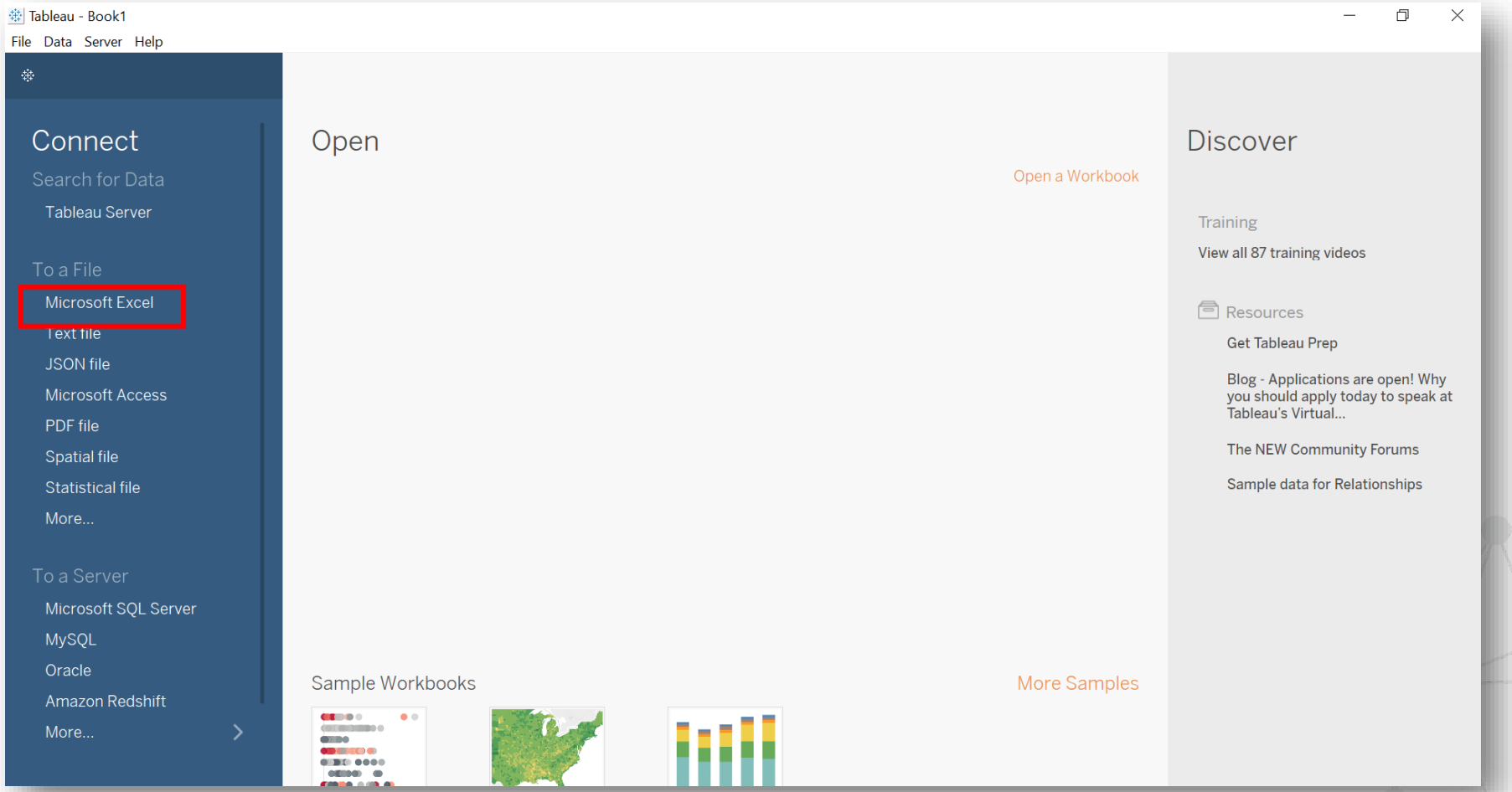
Dataset – **DV_SuperStore_Raw.xlsx**

Data: Sales, cost, product and customers data for stores across different states and region

Target Audience : Management of Superstore

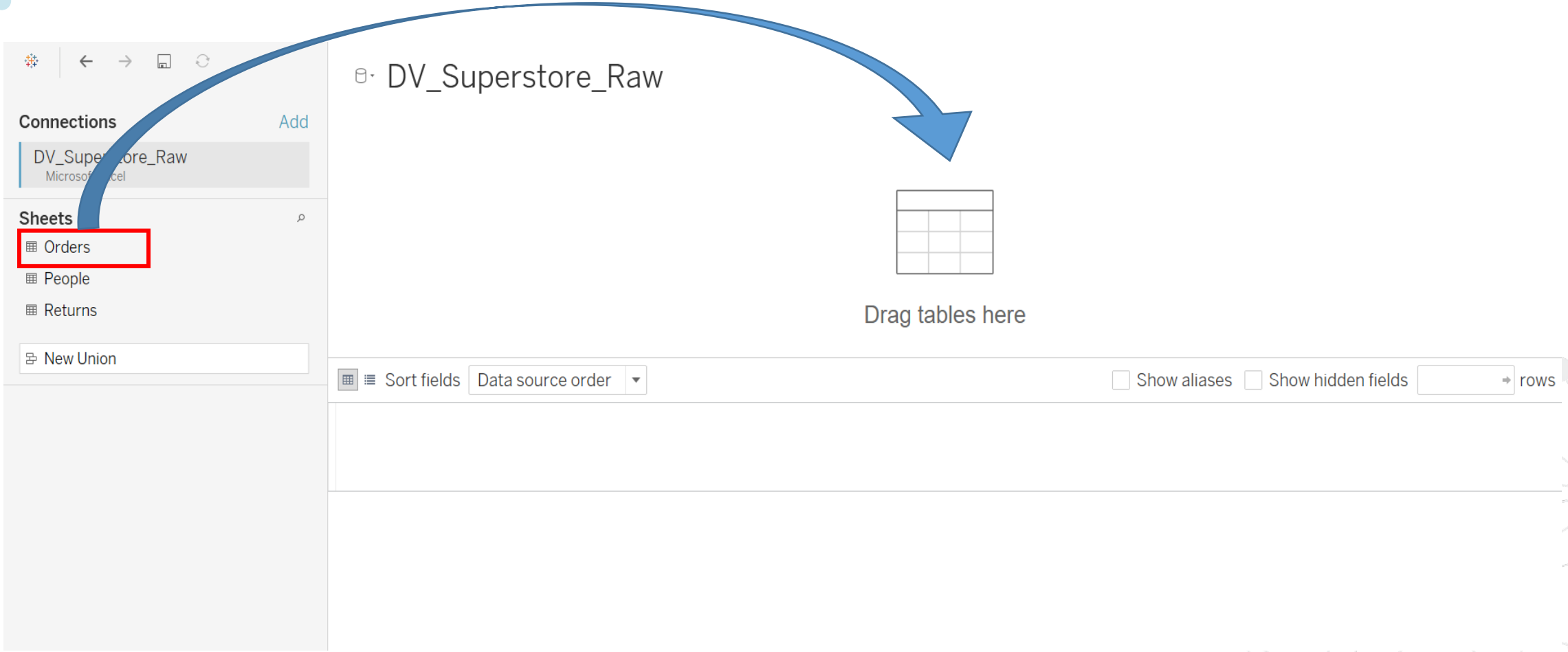
Launch Tableau Desktop

Click on **Microsoft Excel > Open DV_SuperStore_Raw.xlsx**

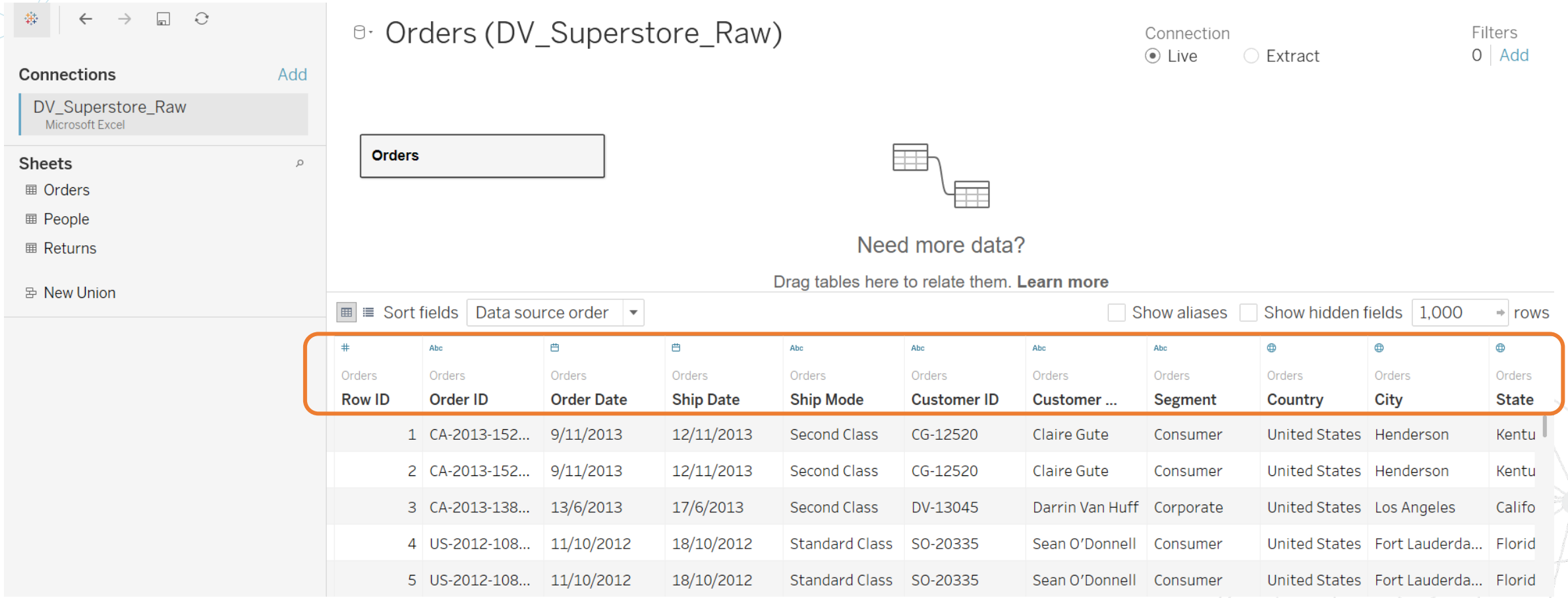


Connect to Data Source

Drag **Orders** table to the Drag Table Here area.



Data Source Page



Our Approach

Orders Row ID	Orders Order ID	Orders Order Date	Orders Ship Date	Orders Ship Mode	Orders Customer ID	Orders Customer ...	Orders Segment	Orders Country	Orders City	Orders State
1	CA-2013-152...	9/11/2013	12/11/2013	Second Class	CG-12520	Claire Gute	Consumer	United States	Henderson	Kentu
2	CA-2013-152...	9/11/2013	12/11/2013	Second Class	CG-12520	Claire Gute	Consumer	United States	Henderson	Kentu
3	CA-2013-138...	13/6/2013	17/6/2013	Second Class	DV-13045	Darrin Van Huff	Corporate	United States	Los Angeles	Califo
4	US-2012-108...	11/10/2012	18/10/2012	Standard Class	SO-20335	Sean O'Donnell	Consumer	United States	Fort Lauderda...	Florid
5	US-2012-108...	11/10/2012	18/10/2012	Standard Class	SO-20335	Sean O'Donnell	Consumer	United States	Fort Lauderda...	Florid

1. Data Profiling
2. Data Cleaning
3. Data Exploration
4. Data Visualisation

Data Profiling - Explore

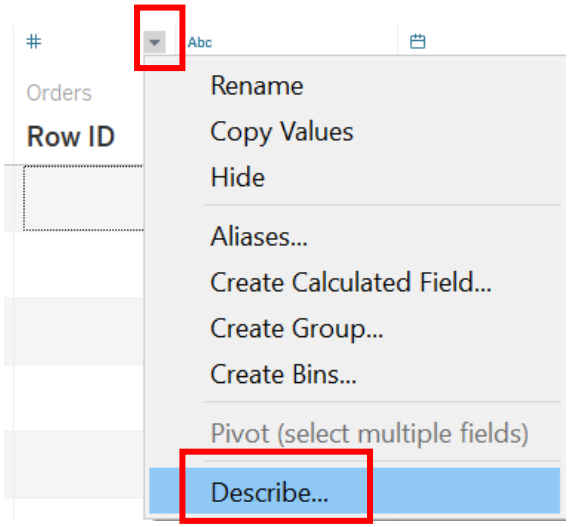
Take a look at the Superstore Dataset in Tableau Desktop.

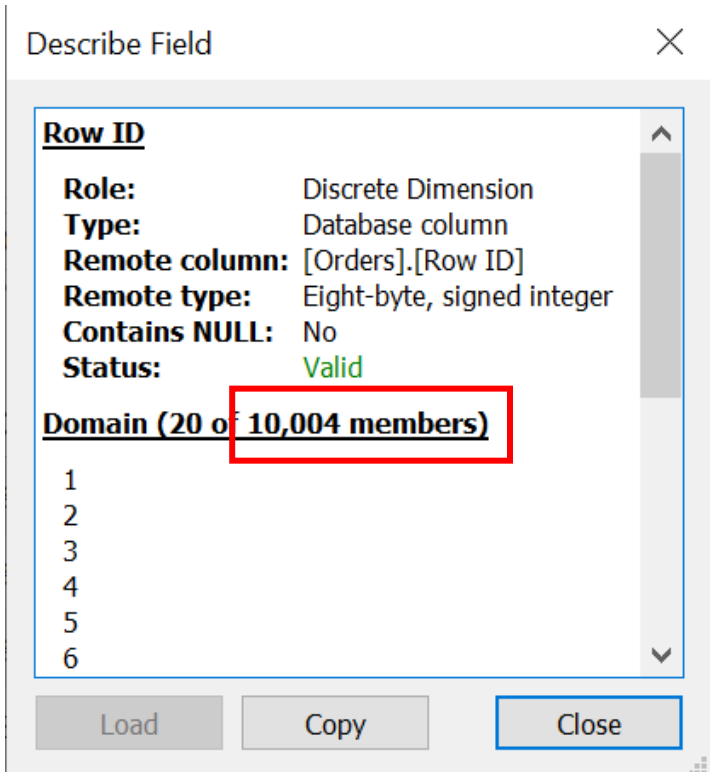
We will explore the following:

- How many **columns and rows** are there in the dataset?
- Do you observe anything strange about the column called **Segment**?
- Do you observe anything strange about the column called **Country**?

Data Profiling

Click ▾ next to column names (header row) for **Row ID** > click **Describe**



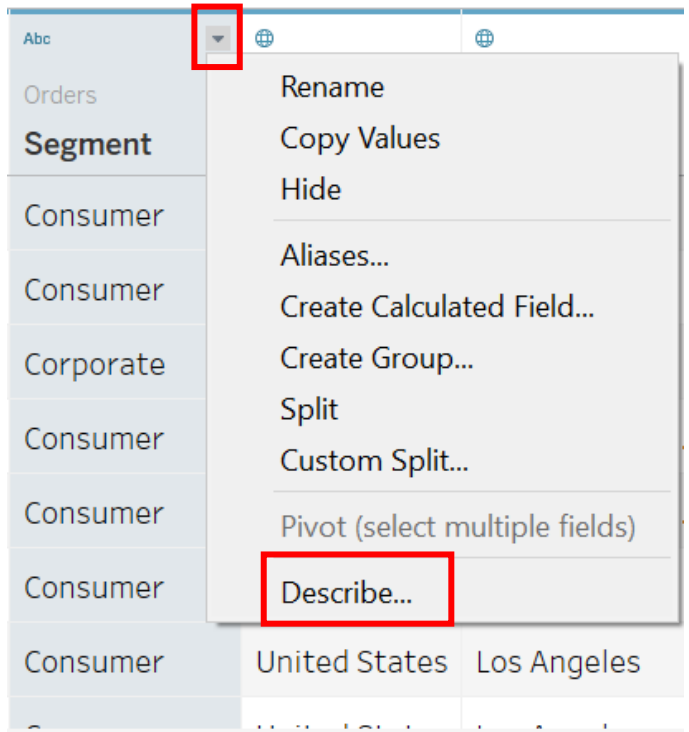


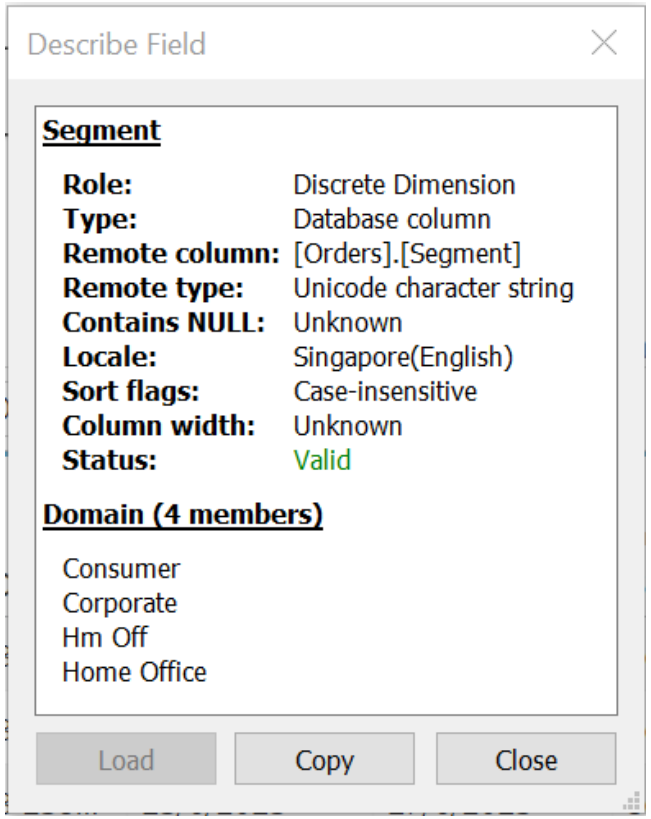
Number of rows	
Number of columns	

Note : This view only shows 20 rows out of a total of 10004 rows.

Data Profile and Audit

Click ▾ next to column names (header row) for **Segment** > click **Describe**



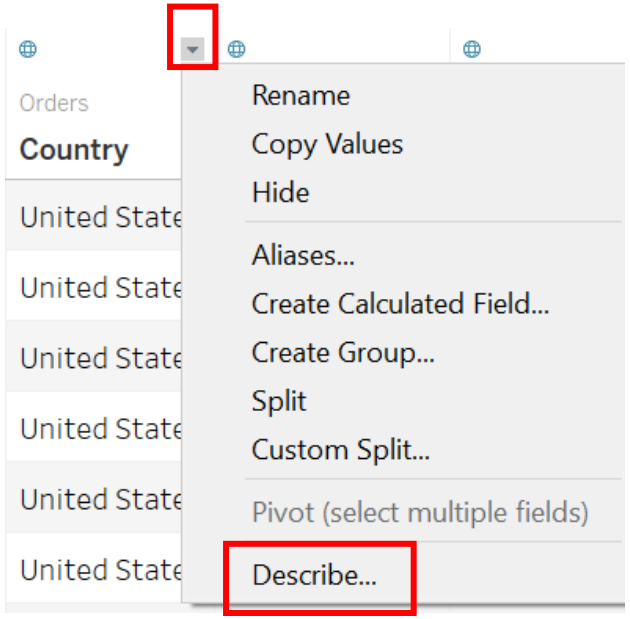


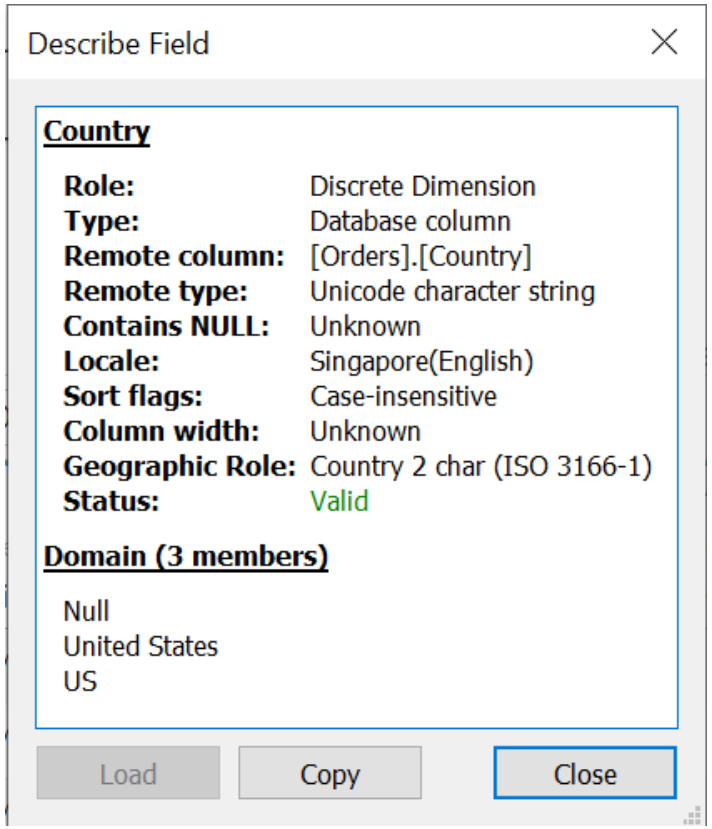
What is the data range/values?

Do you observe anything strange?

Data Profile and Audit

Click ▼ next to column names (header row) for **Country** > click **Describe**



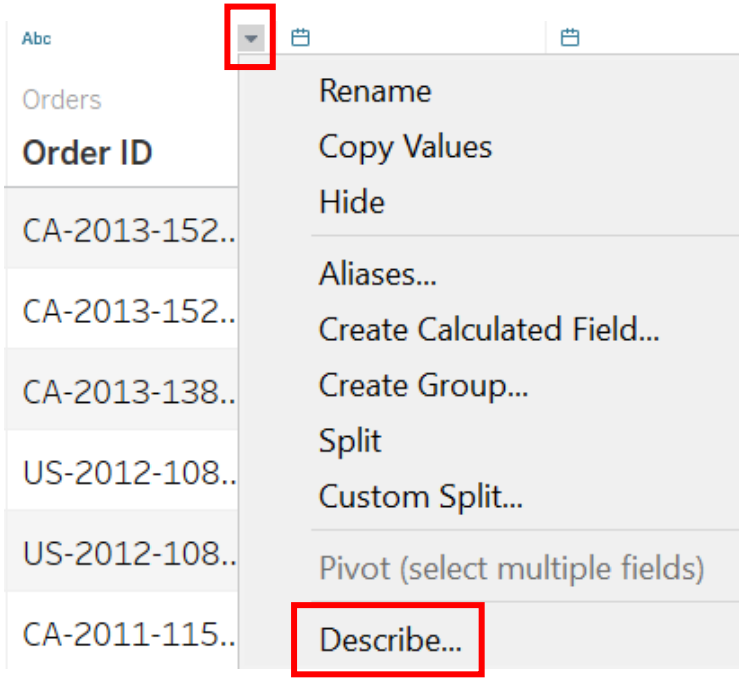


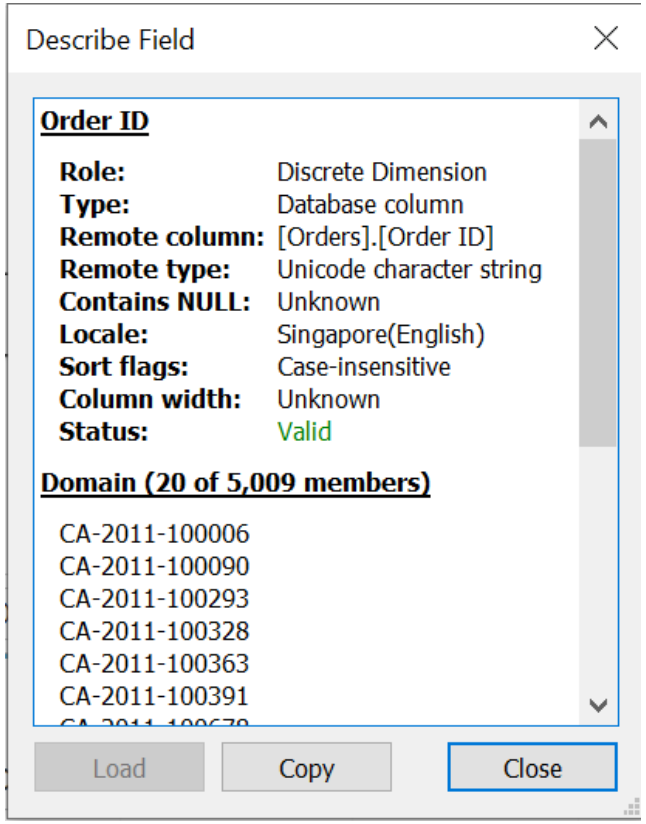
What is the data range/values?

Do you observe anything strange?

Data Profile and Audit

Click ▾ next to column names (header row) for **Order ID** > click **Describe**





What is the data range/values?

Do you observe anything strange?

Data Profiling – Subset

We have completed Data Profiling for these columns. Try out the rest on your own and note them down in your lab sheet.

Column	Data Type	Range/Values	Any Observations?
Segment	nominal/string	Consumer; Corporate; Home Office; **Hm Off (4)**	inconsistency
Country	geographical	United States; **US; Blanks (3)**	inconsistency
State			
Region			
Order ID	nominal/string	e.g. CA-2011-100006 (5009)	Need to split State-Year-Order ID
Product ID			
Sales			
Cost			

What have we just done for Data Profiling?

1. Missing Values – Country
2. Inconsistencies – Segment, Country, State and Region
3. Need splitting – Order ID, Product ID
4. What if I want to know the profit / duration to ship out the product?

What else can we do for Data Profiling?

According to the video earlier, there are 3 levels in Data Profiling :

1. **Column level** - Min, Max, Mean, Mode, Percentile, Standard Deviation, Frequency → Learned in DAVA (Statistics node)
   Tableau Desktop is not the best for this. Can try Tableau Prep, Python, KNIME.

2. **Metadata level** - Data Type, Length, Duplicates, Occurrence of Null Values, Typical String Patterns → Done in **hands-on exercise**

3. **Table level** - primary key, foreign key → Learned under Data Integration. No need to do data integration for Superstore dataset.

You should definitely look into all three levels when dealing other datasets.

Temasek POLYTECHNIC | School of Informatics & IT

How To Do Data Cleaning

Music from https://pixabay.com/music/introoutro-beautiful-piano-logo-161975/

How To Do Data Cleaning?

- Different types of data will require different types of cleaning.
- This section is one systematic approach you can try.
    1. Remove Duplicates
    2. Resolve Inconsistencies
    3. Handle Missing Values
    4. Handle Anomalies

Data Cleaning : Remove Duplicates

How Do They Come About

- Duplicates most likely arise during **data collection**, such as when you combine datasets from multiple places.

What Do They Look Like

- Same record appear in multiple rows in the same data set.
- E.g. same customer record appearing multiple times.

How To Fix Them

- Remove the duplicated rows

Data Cleaning : Resolve Inconsistencies

How Do They Come About

- May arise during data entry, data transfer, or "poor housekeeping“

What Do They Look Like

- Data contains inconsistencies such as :
- e.g. Mixed codes : Rating has mixture of “1,2,3” and “A,B,C”
- e.g. Mis-labelled classes : Manager vs Mgr, NA vs Nil
- e.g. Capitalization : finance vs Finance
- e.g. Typo : doctor vs docter

How To Fix Them

- Standardize by finding and replacing

Data Cleaning : Handle Missing Values

How Do They Come About

- Could be many causes such as :
  - Nonresponse (refusal to answer a question) due to sensitive info e.g. income
  - Irrelevant attribute for the corresponding observation
  - New fields introduced during data collection (thus missing values for prior records)
  - System error (i.e. faulty sensors) or human error

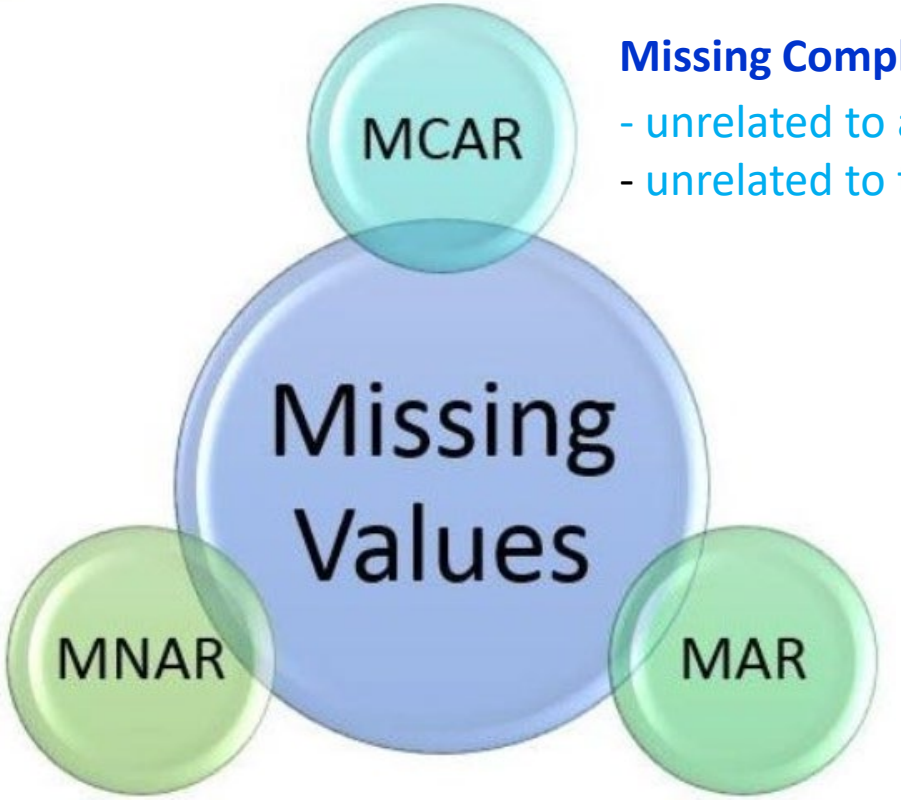
What Do They Look Like

- Often encoded as blanks, NaN or other placeholders

How To Fix Them

- Coming up in the next few slides

Data Cleaning : Handle Missing Values



Missing Completely at Random

- unrelated to any other variables and
- unrelated to the variable with missing values itself

Missing Not at Random

- related to the variable itself

Missing at Random

- related to other variables

https://medium.com/analytics-vidhya/different-types-of-missing-data-59c87c046bf7

Data Cleaning : Handle Missing Values

Delete Rows

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

If only a few rows missing and due to MCAR

Delete Column

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

If majority missing (e.g. >60%)

Data Cleaning : Handle Missing Values

Imputation

- Numeric Data
  - Replace by mean or median of entire column or neighbouring data
  - Recode as a new group called “Missing”
- Categorical Data
  - Replace by mode of entire column or neighbouring data
  - Recode as a new group called “Missing”

ID	Gender	Rating
1	Male	6
2	Male	2
3	Female	1
4	Male	4.75
5	Female	5
6	Female	9
7	Male	3
8	Female	4
9	Female	4.75
10	Male	8

Decide on imputation method based on MCAR, MAR, MNAR

Data Cleaning : Handling Anomalies

How They Come About

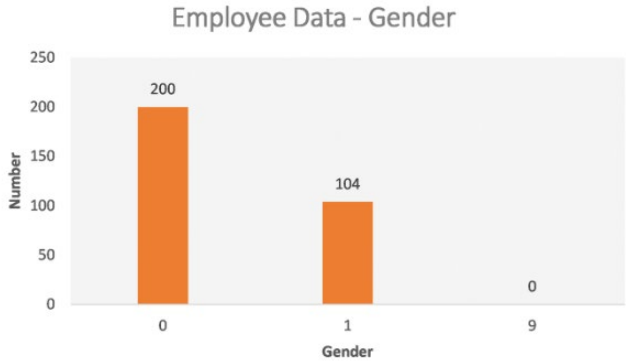
- Anomalies are things that deviate from what is standard, normal, or expected.
- They may arise from outliers or from data errors.
- Data errors are erroneous values
  - may be due to faulty instruments or data entry errors
- Outliers are non-erroneous values
  - May be caused by unusual values that are far away from the norm

Data Cleaning : Handling Anomalies

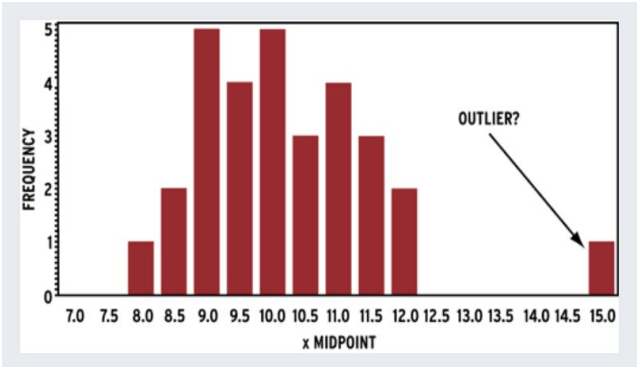
How To Detect Them

- For Categorical Data - Frequency Table or Bar Chart

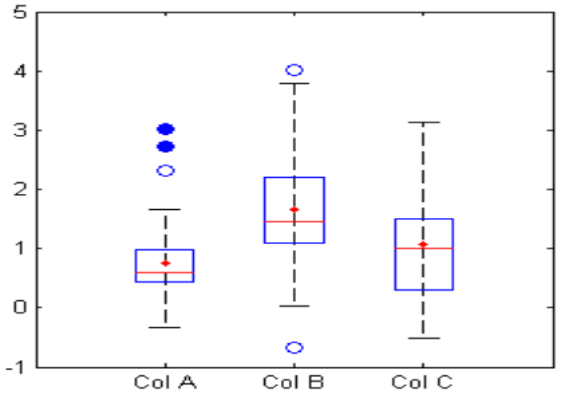
REGION	SALES
North	$29034
South	$56728
Soth	$12000
East	$27000
West	$89092



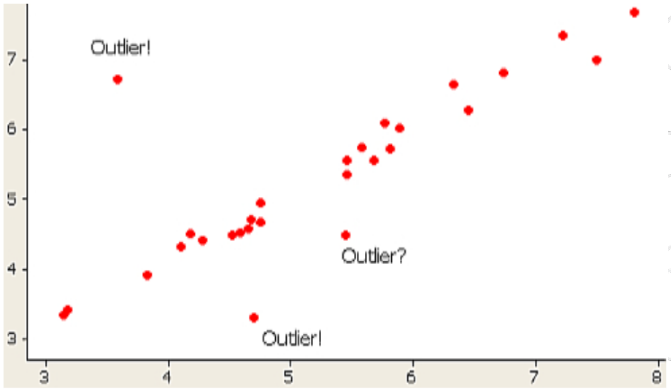
- For Numeric Data - Histogram, Box Plot, Scatter Plot



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



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



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

Data Cleaning : Handling Anomalies

Once an anomaly is detected, we will need to check for underlying causes and determine if it is classified as an outlier or a data error.

How To Fix Them - Outliers

- It is sometimes useful to exclude (filter) them as they can veer the mean value.
- However, do it with caution.

How To Fix Them – Data Errors

- Erroneous data should be excluded (filter them).

HandsOn

Data Cleaning

*We are using the same worksheet as previous activity.

Data Cleaning

In the previous hands-on exercise on Data Profiling, we explored the Superstore dataset and discovered the following:

1. Missing Values – Country
2. Inconsistencies – Segment, Country, State and Region
3. Need splitting – Order ID, Product ID
4. Need new Columns - What if I want to know the profit / duration to ship out the product?

Here, we will pick up from where we left off and execute the act of data cleaning.

Data Cleaning

Aim : You will learn to handle the following in Tableau Desktop:

(1) missing values - remove rows using filter
(2) inconsistency - standardize the values
(3) splitting data fields - split by delimiter
(4) creating new columns - to extract new insights.

Data Cleaning : Remove Rows

Problem 1 : Missing Values for **Country**	Resolution : Remove the rows

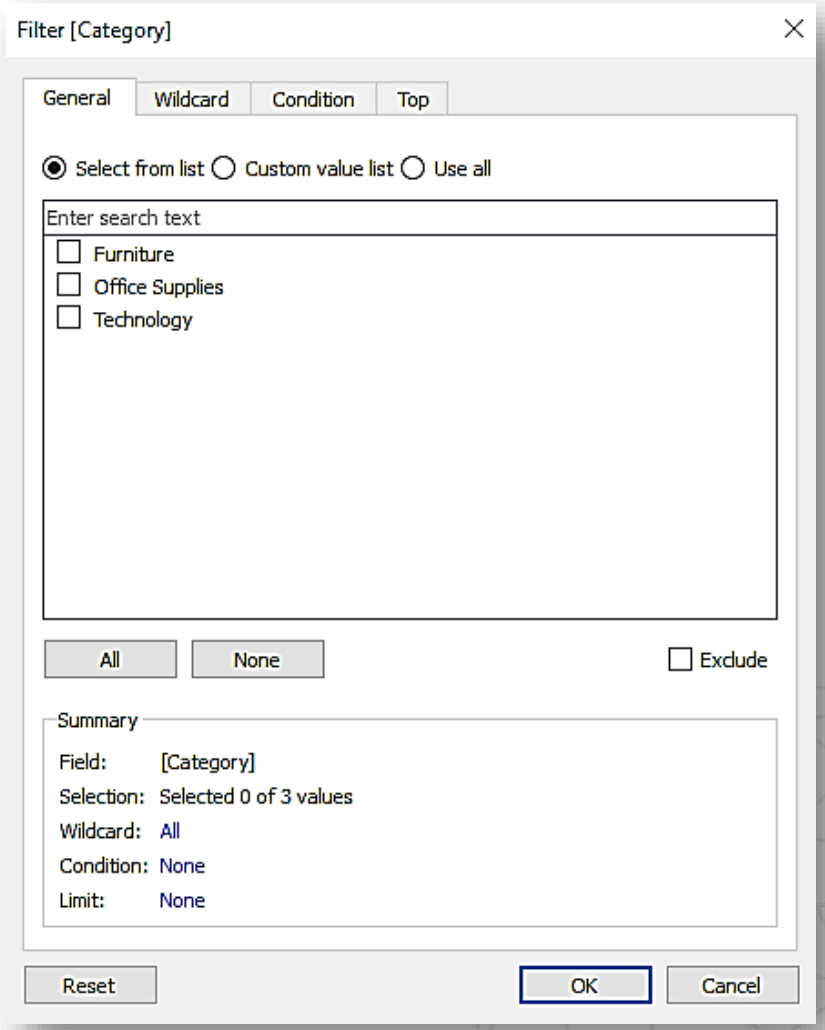
Question : How can we remove empty value in the Country column?

How can we get the answer: Create filter

Filtering

In Tableau, you can filter by Dimension and choose to limit the data values as follows:

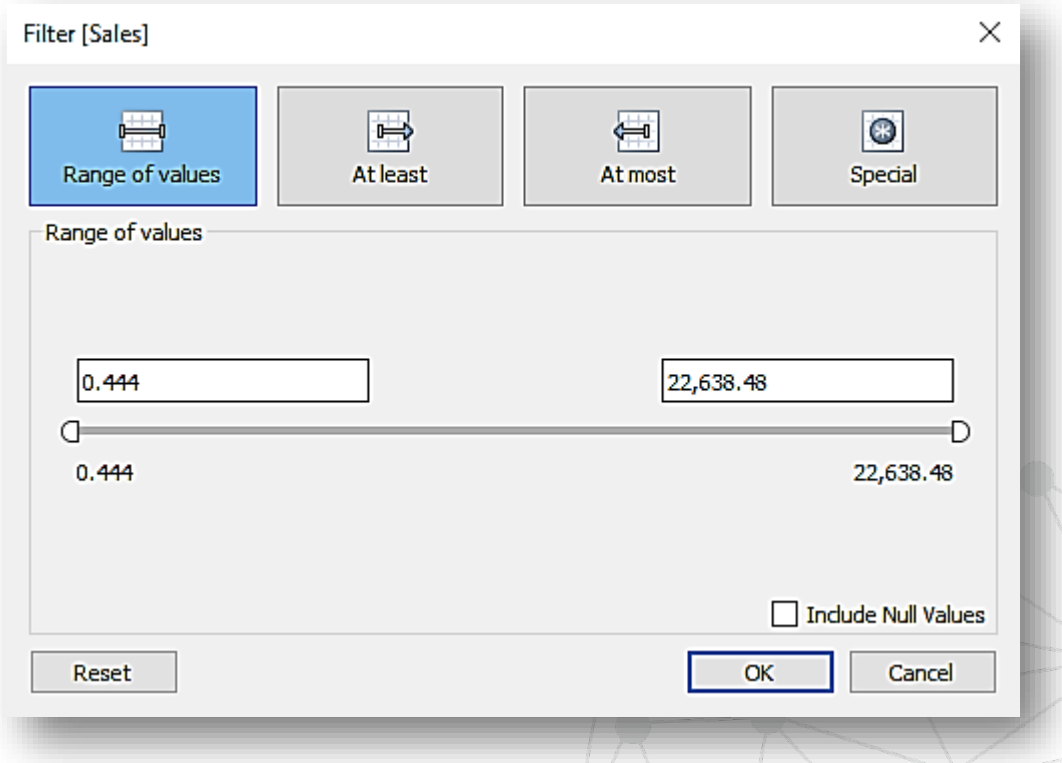
General	Select data values to include or exclude
Wildcard	Values that matches/ Starts with/End With
Condition	Filter by field/range/formula
Top	Filter by Top ‘N’



Filtering

In Tableau, you can also filter by Measure and choose to limit the data values as follows:

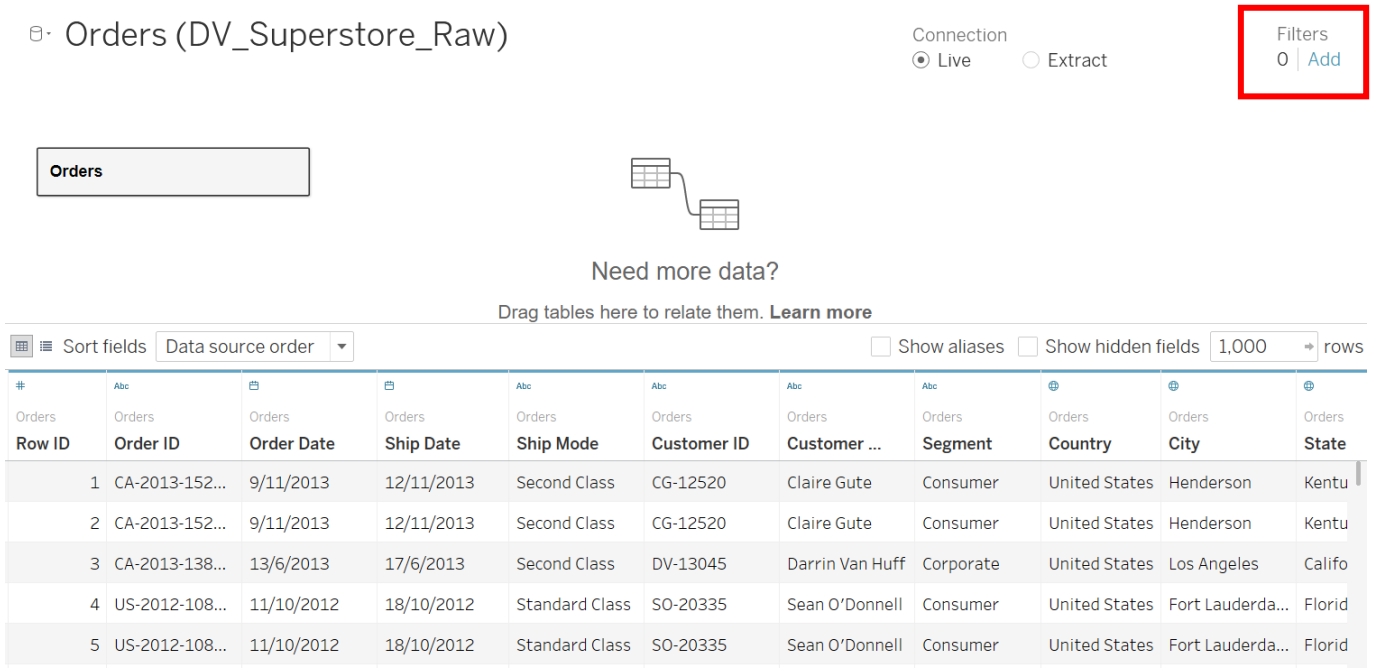
Range of values	Include values within the range
At least	Include values above a specified value
At most	Include values below a specified value
Special	Filter null or non-null value



How to Create Filter

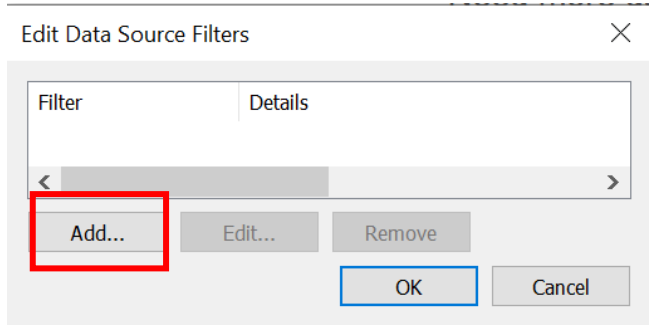
Question : How can we remove empty value in the Country column?

1. Click on 'Add' under Filters (top right)



Orders Row ID	Orders Order ID	Orders Order Date	Orders Ship Date	Orders Ship Mode	Orders Customer ID	Orders Customer ...	Orders Segment	Orders Country	Orders City	Orders State
1	CA-2013-152...	9/11/2013	12/11/2013	Second Class	CG-12520	Claire Gute	Consumer	United States	Henderson	Kentu
2	CA-2013-152...	9/11/2013	12/11/2013	Second Class	CG-12520	Claire Gute	Consumer	United States	Henderson	Kentu
3	CA-2013-138...	13/6/2013	17/6/2013	Second Class	DV-13045	Darrin Van Huff	Corporate	United States	Los Angeles	Califo
4	US-2012-108...	11/10/2012	18/10/2012	Standard Class	SO-20335	Sean O'Donnell	Consumer	United States	Fort Lauderda...	Florid
5	US-2012-108...	11/10/2012	18/10/2012	Standard Class	SO-20335	Sean O'Donnell	Consumer	United States	Fort Lauderda...	Florid

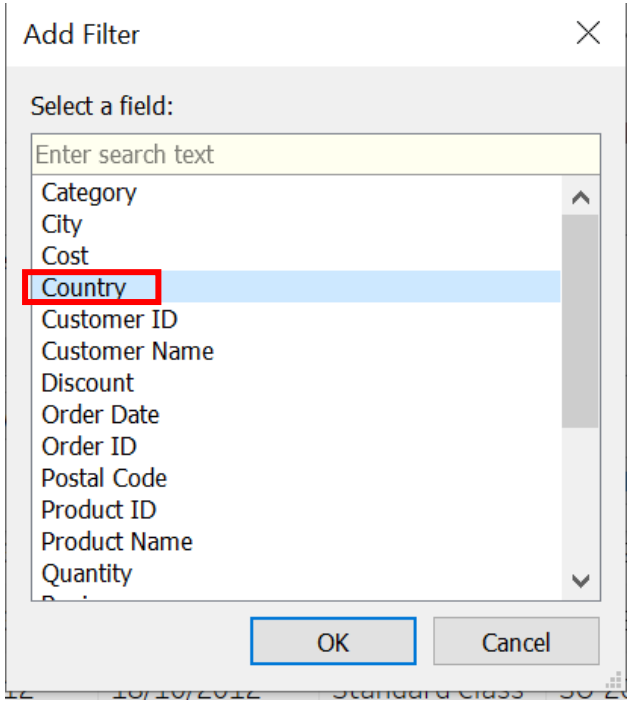
2. Click on Add again



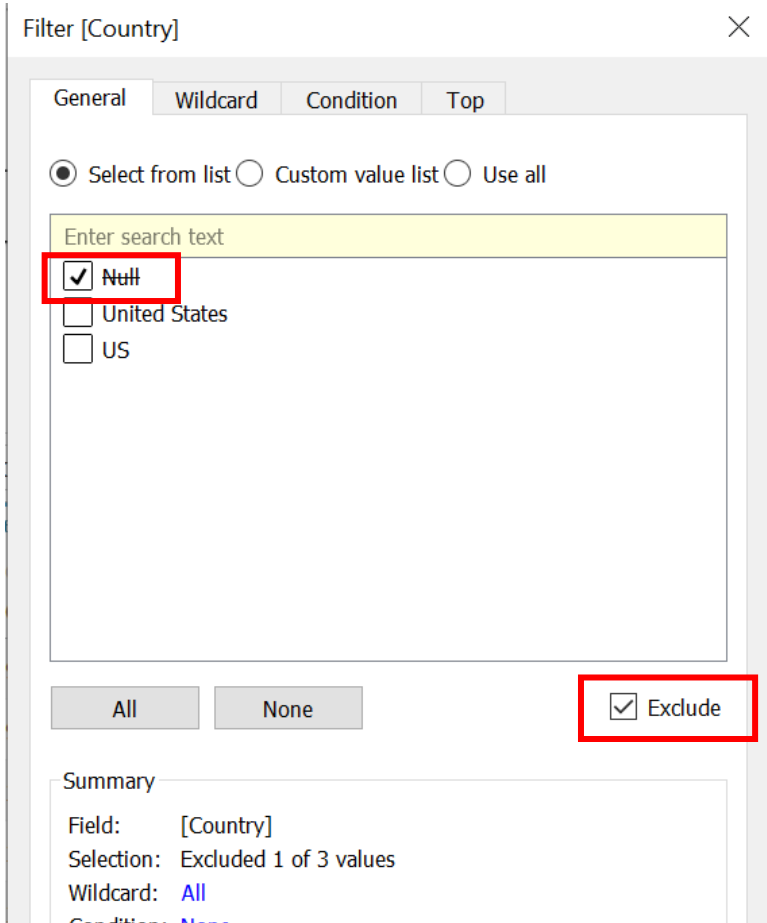
How to Create Filter

Question : How can we remove empty value in the Country column?

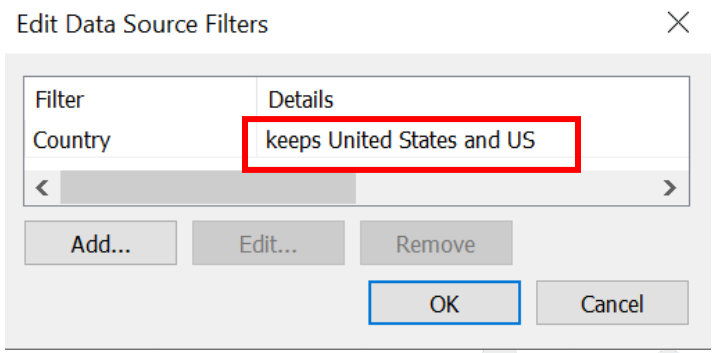
3. Select Country field



4. Check NULL & Exclude checkbox



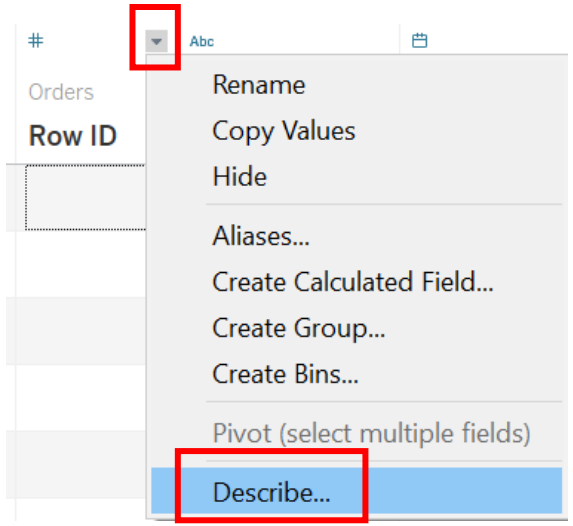
5. Make sure :

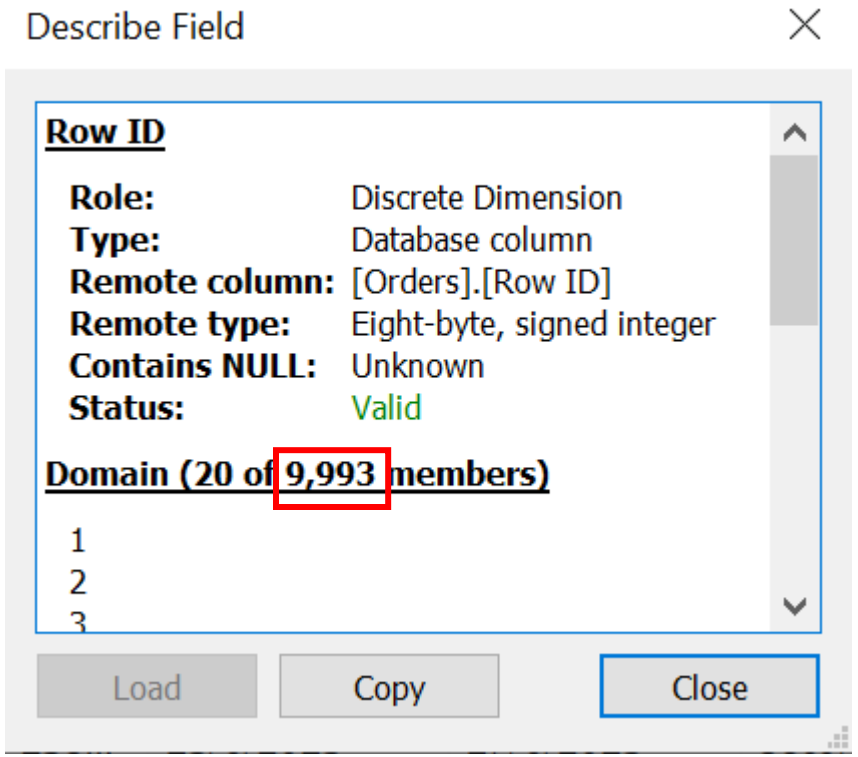


How to Create Filter

Question : How can we remove empty value in the Country column?

How many rows are left?





Number of rows	9993
Number of column	21

Do you get this?

Data Cleaning : Find and Replace

Problem 2a : Inconsistencies in **Country**	Resolution : Standardize everything

We will replace ‘US’ with ‘United States’.

Question : How can we standardize the values in the Country column?

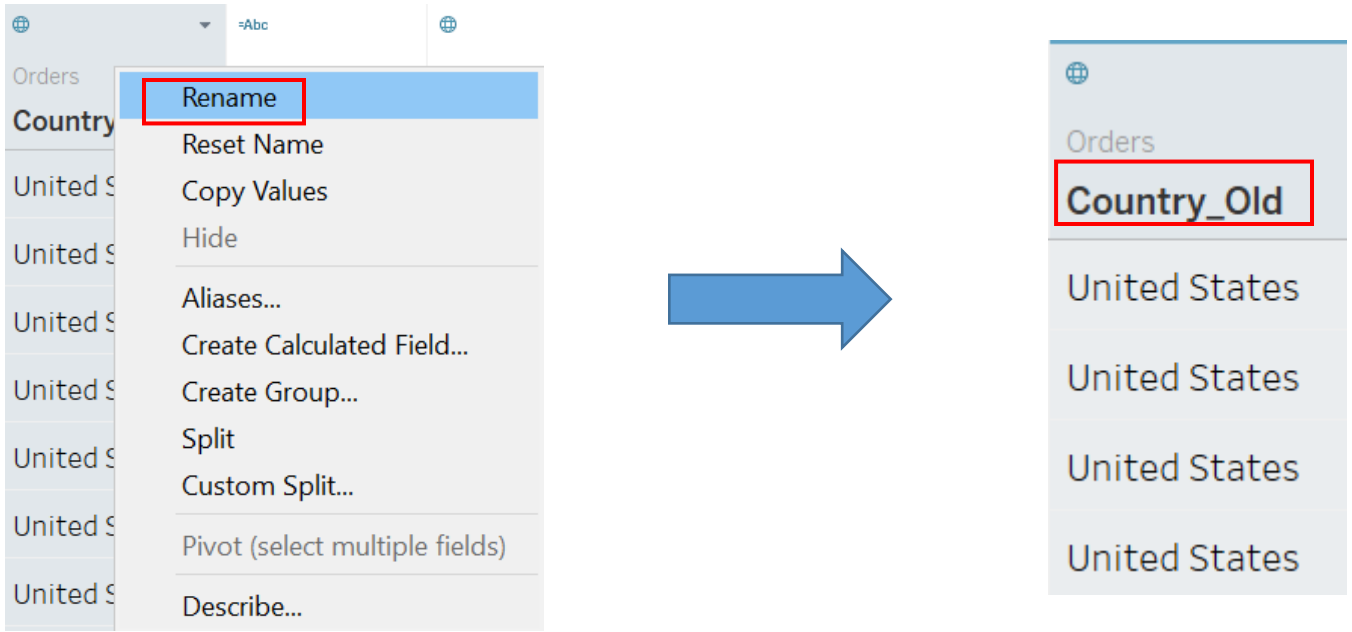
How can we get the answer: Create a new field

Data Cleaning : Column Rename

Problem 2a : Inconsistencies in **Country**	Resolution : Standardize everything

First, let's rename the original column as "Country_Old".

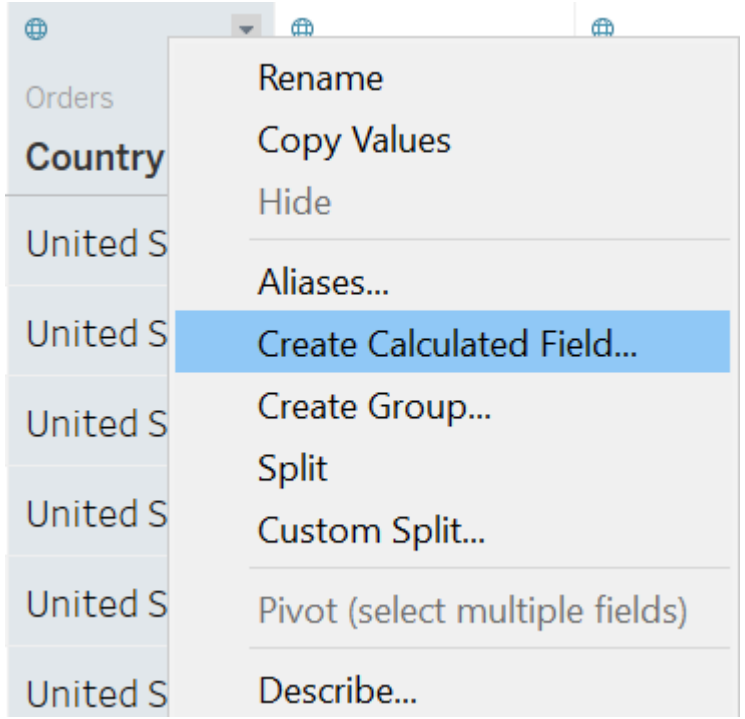
1. Right-click column header for "Country" > Select Rename > Type "Country_Old".



Create a Calculated Field - Replace

Question : How can we standardize the values in the Country column?

2. Right Click on **Country_Old** field and select **Create Calculated field**



3. Type in 'Country' as name of the field > Add code

REPLACE([Country_Old], 'US', 'United States')

Create a Calculated Field - Replace

Question : How can we standardize the values in the Country column?

4. Check that there is only **1 distinct value** in Country field

Orders Country_Old	Calculation Country
United States	United States
United States	United States
US	United States
US	United States
United States	United States

Data Cleaning : Find and Impute

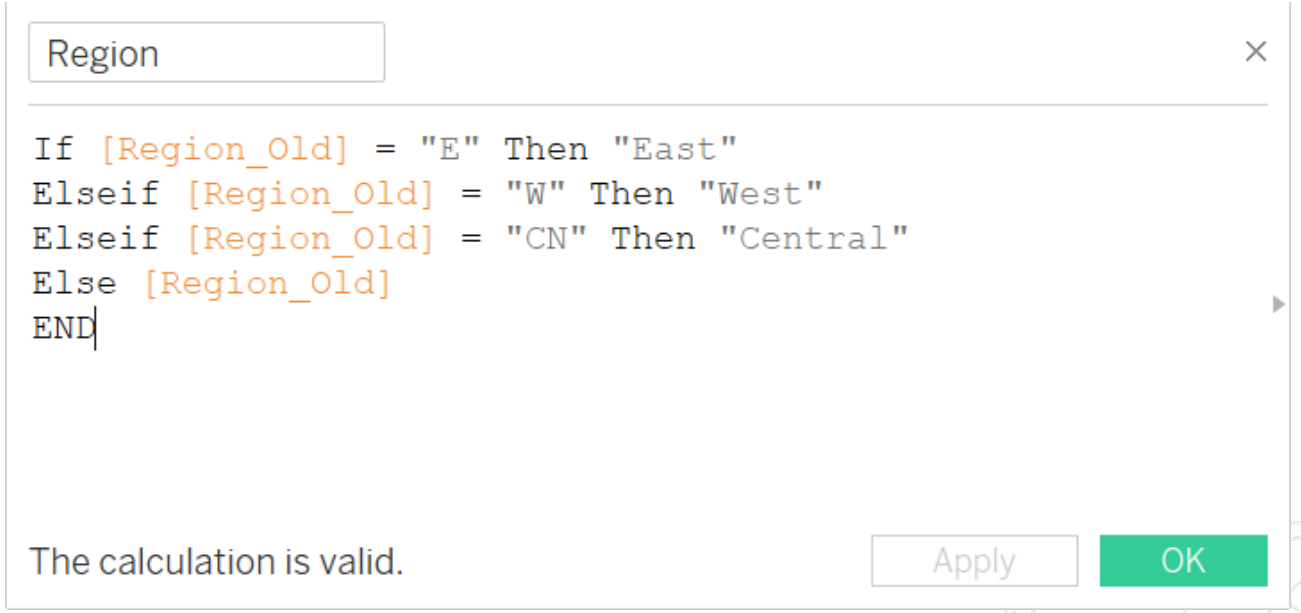
Problem 2b : Inconsistencies in **Region**	Resolution : Standardize to respective region

We will standardized values in Region column (E, W, CN)

1. Let's rename the original column as "Region_Old".

2. Right Click on **Region_Old** field and select **Create Calculated field**

3. Type in 'Region' as name of the field > Add code



If [Region_Old] = "E" Then "East"
Elseif [Region_Old] = "W" Then "West"
Elseif [Region_Old] = "CN" Then "Central"
Else [Region_Old]
End

Create a Calculated Field – If-then-else

Problem 2b : Inconsistencies in **Region**	Resolution : Standardize to respective region

4. Check that there are only **4 distinct values** in Region field

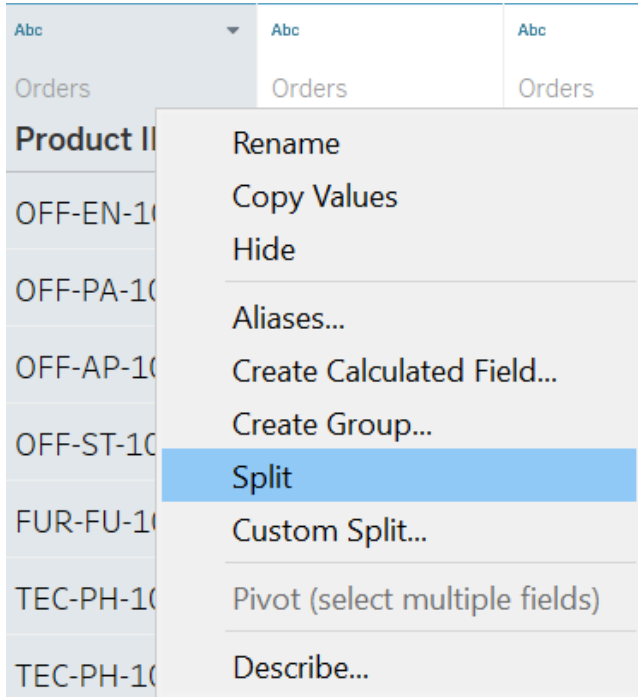
Abc	=Abc
Orders	Calculation
Region_Old	**Region**
East	East
East	East
East	East
E	East
East	East
East	East

Data Cleaning: Split Data

Problem 3 : Prod ID made of cat, sub-cat, 8-digit Prod ID	Resolution : Split to get ID

We will perform a split to separate the 8 digit Product ID from the Product ID column.

1. Right-click on column header of **Product ID** > Select **Split**



2. Scroll to the far right to see the split

=Abc Calculation Product ID - Split 1	=Abc Calculation Product ID - Split 2	=# Calculation Product ID - Split 3
FUR	BO	10001798
FUR	CH	10000454
OFF	LA	10000240
FUR	TA	10000577
OFF	ST	10000760

Data Cleaning: Column Rename

Problem 3 : Prod ID made of cat, sub-cat, 8-digit Prod ID	Resolution : Split to get ID

3. Rename the columns to get this :

Abc Orders Product ID_Old	# Calculation Product ID
FUR-BO-10001798	10001798
FUR-CH-10000454	10000454
OFF-LA-10000240	10000240
FUR-TA-10000577	10000577

Split Function

You may have noticed that there are two options in Tableau : **Split** and **Custom Spit**.
What the difference?

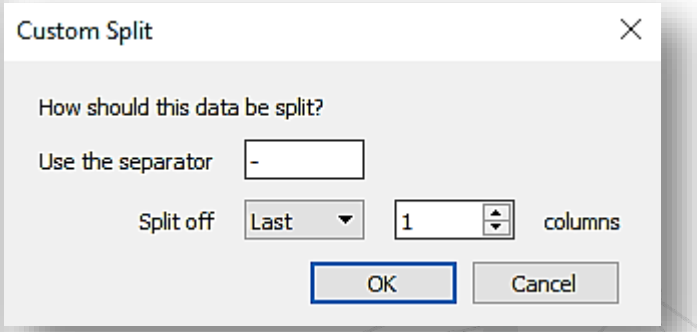
Split

Abc Orders **Customer ID**	=Abc Calculation **Code**	=# Calculation **No**
CG-12520	CG	12520
CG-12520	CG	12520
DV-13045	DV	13045
SO-20335	SO	20335
SO-20335	SO	20335
BH-11710	BH	11710
BH-11710	BH	11710
BH-11710	BH	11710

Split by delimiter (-)

Custom Split

Abc Orders **Order ID**	=Abc Calculation **Code**
CA-2013-152156	152156
CA-2013-152156	152156
CA-2013-138688	138688
US-2012-108966	108966
US-2012-108966	108966
CA-2011-115812	115812
CA-2011-115812	115812
CA-2011-115812	115812

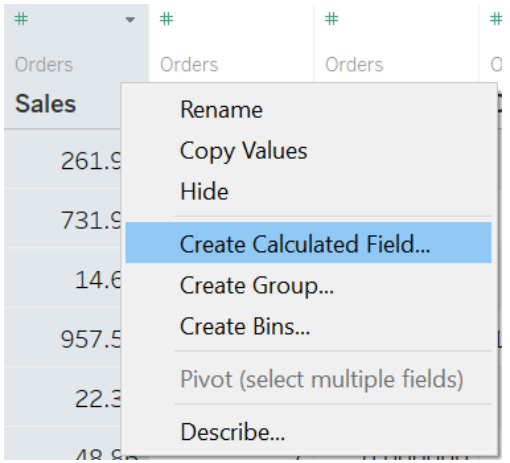


Split by delimiter (-) Last Column

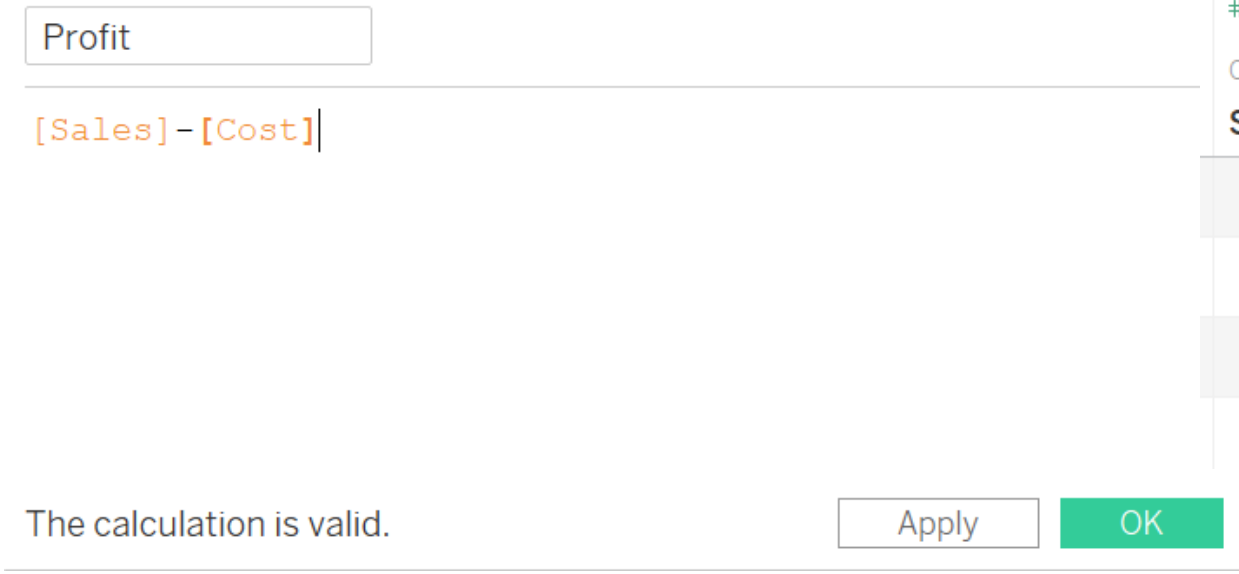
Data Cleaning : Add New Columns

Problem 4 : What if I want to know the Profit?	Resolution : Add Col / Use Formula

1. Right-Click on **Sales** field > Select **Create Calculated field**



2. Type in '**Profit**' as name of the field > Add code



Orders Sales	Calculation Profit	Orders Cost
261.96	41.91	220.05
731.94	219.58	512.36
14.62	6.87	7.75
957.58	-383.03	1,340.61

[Sales] - [Cost]

Data Cleaning : Save Changes

Once data cleaning is completed, save changes and exit Power Query Editor.

Click on **File** tab → **Save As**

What have we done for Data Cleaning?

- Missing Value diagnostic mechanism like **MCAR, MAR and MNAR** helps to understanding the underlying reasons for the missing data. We can then prescribe the appropriate treatment for rectifying the missing value.
- Possible ways to **handle missing values** are deletion of rows, deletion of columns or by imputation.
- For **numeric** data, we can impute with mean, median or mean of neighbouring data.
- For **categorical** data, we can impute with the mode.

Temasek POLYTECHNIC | School of Informatics & IT

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

Music from https://pixabay.com/music/introoutro-beautiful-piano-logo-161975/

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.

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.

Why Data Standards and Metadata are Important

Data standards and metadata enable data to be FAIR…

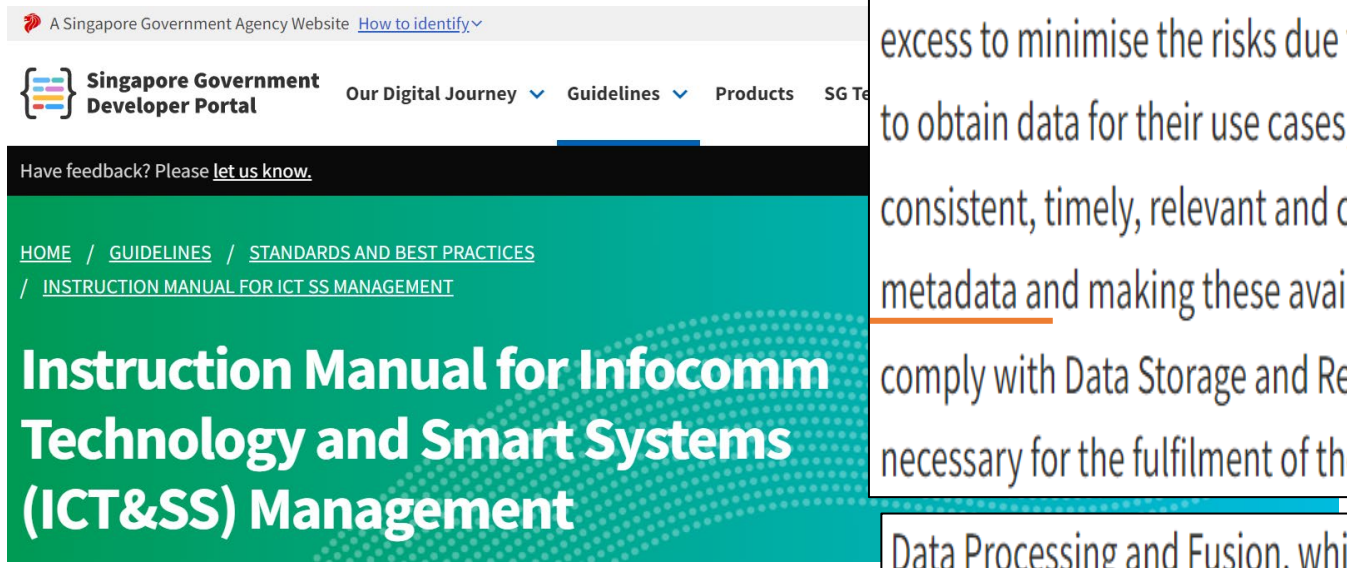
- **F**indable: easily searchable
- **A**ccessible: easy to use
- **I**nteroperable: easily combined
- **R**e-useable: easy to share

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

Temasek POLYTECHNIC | School of Informatics & IT

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

Music from https://pixabay.com/music/introoutro-beautiful-piano-logo-161975/

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?

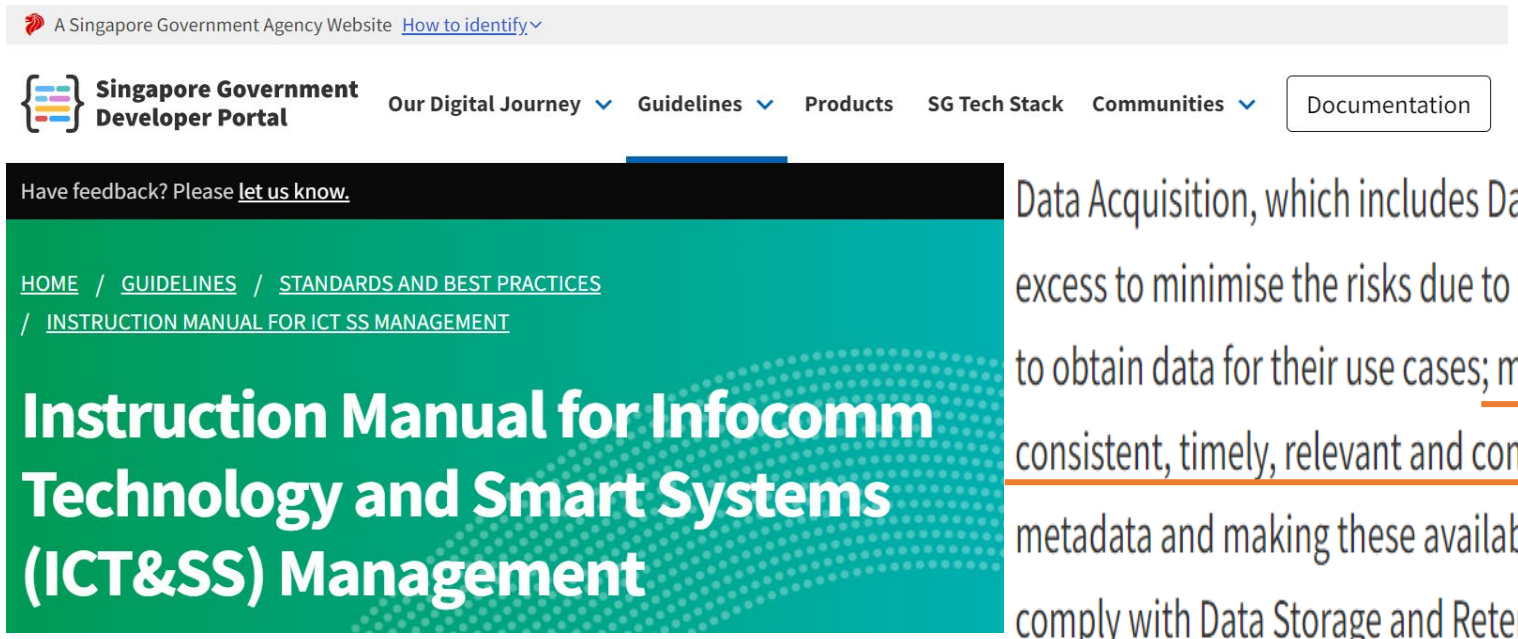
Relevance - Does it meet business needs?

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

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

Singapore Context

Accuracy Consistency
Timeliness Relevance
Completeness



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

Practical 2

Before Class	**Concepts**	Data Attributes Data Quality Dimension
During Class	**Hands-on**	Recap Unpivoting Data Profiling Data Cleaning Data Standards and Meta Data Data Quality Dimension
After Class	**Hands-on**	*Revise today's class with LMS: Apr/Oct – Week X (…)* *Go through Additional Resources slides* - *Explore MCAR, MAR, MNAR* - *Watch video on "What is Data Profiling"* - *Explore further in Data Profiling and Data Cleaning* - *Read up further on Data Standards*