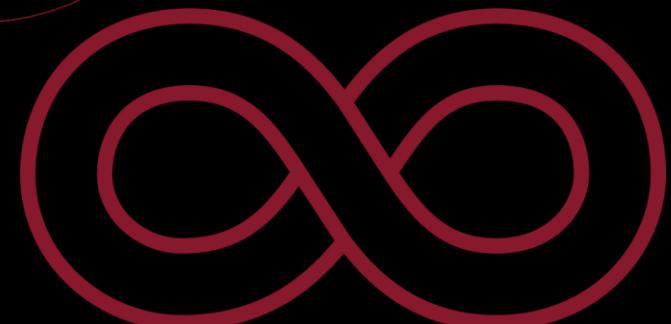


# Excel Module

April 2022

This document remains copyrighted to Natural Intelligence Solutions Technology. This document and content is for the use of the recipients of the email, presentation or meeting only. The document must not be copied, reproduced or shared in any format whether in part or in whole, in any medium without the written consent of Natural Intelligence Solutions Technology. Uncontrolled if printed. All rights reserved.

SCIENCE.  
SOLUTIONS.  
STORIES.



# Contents

## **Overview of training datasets**

Managing multiple worksheets

Data understanding

Conditional formatting and logical operators

Lookup and reference functions

A review of essential functions

Fundamentals of descriptive statistics

Selecting the right charts to visualize your analysis result

Dashboard dos and don'ts



NATURAL INTELLIGENCE SOLUTIONS TECHNOLOGY



# Overview of training datasets

We will be using these 2 datasets for the hands-on exercises.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1	CustomerID	LastName	State	Gender	MaritalStatus	Age	NumberofDependents								
2	15815690	KUMAR	Perak	Female	Yes	40	0								
3	15815660	LAL	Perak	Female	Yes	34	0								
4	15815656	SHARMA	Terengganu	Female	Yes	39	2								
5	15815645	ADI	Kelantan	Male	Yes	37	3								
6	15815628	SHAN	Pahang	Female	Yes	37	1								
7	15815626	ADIPUTERA	Terengganu	Male	Yes	63	5								
8	15815615	ADIPUTRA	Kelantan	Male	Yes	36	3								
9	15815660	ADIPUTERA	Perak	Male	No	74	0								
10	15815552	JAI	Selangor	Female	Yes	42	0								
11	15815534	ADIPUTRA	Perlis	Male	Yes	37	0								
12	15815530	PAL	Perak	Female	Yes	42	5								
13	15815490	AGUNG	Perak	Male	Yes	40	2								
14	15815443	AGGARWAL	Melaka	Female	No	46	0								
15	15815428	AGUS	Pahang	Male	No	34	0								
16	15815420	AHAD	Perak	Male	Yes	47	1								
17	15815364	RAJE	Perlis	Female	Yes	28	4								
18	15815316	ANDIKA	Terengganu	Male	Yes	50	4								
19	15815295	CHANDE	Kelantan	Female	Yes	38	2								
20	15815271	ANJANG	Johor	Male	No	43	1								
21	15815259	CHANDER	Kedah	Female	No	56	0								
22	15815236	ANUAR	Terengganu	Male	No	34	1								
23	15815131	NARA	Johor	Female	Yes	36	1								
24	15815125	ARIFF	Kelantan	Male	Yes	45	5								
25	15815097	RAI	Kuala Lumpur	Female	Yes	34	5								
26	15815095	ASHRAFF	Kelantan	Male	No	54	0								
27	15815070	NATH	Perak	Female	Yes	44	0								
28	15815043	ATAN	Melaka	Male	Yes	49	5								
29	15815040	GOEL	Perak	Female	Yes	42	4								
30	15814998	AWANG	Pahang	Male	Yes	42	4								
31	15814940	BHAT	Perak	Female	Yes	33	4								
32	15814930	RAJI	Perak	Female	Yes	40	2								
33	15814923	BAGUS	Melaka	Male	Yes	38	2								
34	15814846	BAMBANG	Terengganu	Male	Yes	52	2								

- 2 datasets:
- Customer Personal Info**  
~7 columns and 10,000 rows
  - Financial Info**  
~ 30 columns and 10,000 rows

# Description of Customer Personal Info dataset



7 columns (fields) and 10,000 rows

Fields	Description	Type
CustomerID	Primary key to uniquely identify a customer record in dataset	8 digits
LastName	Last name of customer	String
State	Origin of customer	String
Gender	Gender of customer	String
MaritalStatus	Marital status of customer, 2 options: yes – married or no – not married	Binary
Age	Age of customer	2 digits
NumberofDependents	Number of people who are in immediate family	Numeric



# Overview of training datasets

We will be using these 2 datasets for the hands-on exercises.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S
1	MonthlySalary	RiskProfile	CreditScore	CustomerID	NumberBankAccts	FixedDeposit	CreditCard	House	Car	LifeInsurance	MedicalCards	PensionPlan	EducationFund	GovtBonds	Share	HousingLoan	CarLoan	Savings	Loan
2	101349	0.16	619	15634602		1 Yes	Yes	No	Yes	Yes	No	No	Yes	Yes	Yes	No	Yes	65424.71	0
3	112543	0.48	608	15647311		1 No	Yes	No	No	Yes	No	Yes	Yes	No	Yes	Yes	No	23130.27	33658
4	113932	0.13	502	15619304		3 Yes	Yes	No	No	No	No	No	Yes	Yes	Yes	Yes	Yes	19497.81	0
5	93827	0.21	699	15701354		2 No	Yes	No	No	No	No	No	No	Yes	No	No	No	63760.01	1431
6	79084	0.03	850	15737888		1 Yes	Yes	Yes	Yes	Yes	No	No	No	No	Yes	No	No	14422.97	40265
7	149757	0.04	645	15574012		2 Yes	No	Yes	Yes	No	Yes	Yes	Yes	No	No	Yes	Yes	43903.89	26063
8	10063	0.26	822	15592531		2 Yes	Yes	Yes	Yes	Yes	No	No	Yes	No	No	No	No	79283.95	16828
9	119347	0.16	376	15656148		4 Yes	Yes	Yes	No	No	Yes	Yes	No	No	Yes	Yes	Yes	10517.56	11864
10	74941	0.45	501	15792365		2 No	Yes	No	No	Yes	No	No	No	No	Yes	No	No	44416.56	31645
11	71726	0.38	684	15592389		1 Yes	Yes	Yes	No	Yes	Yes	No	No	No	Yes	Yes	No	76232.61	34991
12	80181	0.4	528	15767821		2 No	Yes	Yes	No	Yes	No	Yes	Yes	Yes	Yes	Yes	No	12631.02	0
13	76390	0.34	497	15737173		2 Yes	No	Yes	Yes	No	Yes	Yes	Yes	No	No	No	No	66437.53	18209
14	26261	0.05	476	15632264		2 Yes	No	No	No	Yes	No	Yes	No	No	No	No	No	15942	5332
15	190858	0.49	549	15691483		2 No	Yes	Yes	No	No	Yes	No	No	Yes	No	No	No	74170.58	10541
16	65952	0.28	635	15600882		2 Yes	No	No	No	Yes	Yes	No	No	No	No	No	No	70164.99	34720
17	64327	0.42	616	15643966		2 No	Yes	No	Yes	No	Yes	Yes	Yes	No	Yes	No	51048.98	19521	
18	25098	0.32	653	15737452		1 Yes	Yes	Yes	No	Yes	Yes	No	Yes	No	Yes	Yes	28980.38	0	
19	14406	0.13	549	15788218		2 Yes	No	No	No	Yes	No	No	Yes	No	No	No	No	57603.84	26508
20	158685	0.37	587	15661507		1 No	No	No	Yes	No	No	No	No	Yes	Yes	No	No	12639.9	20227
21	54724	0.38	726	15568982		2 Yes	No	No	No	Yes	No	Yes	No	Yes	No	No	No	18036.71	15789
22	170886	0.37	732	15577657		2 Yes	Yes	Yes	No	No	No	No	Yes	Yes	Yes	No	No	20951.12	44012
23	138555	0.33	636	15597945		2 Yes	Yes	Yes	No	No	Yes	Yes	Yes	No	Yes	No	No	10717.9	0
24	118914	0.23	510	15699309		1 Yes	No	No	Yes	No	No	Yes	Yes	No	No	Yes	Yes	45521.72	1645
25	8488	0.32	669	15725737		2 No	No	Yes	Yes	Yes	No	Yes	Yes	No	Yes	No	No	12143.82	15041
26	187616	0.25	846	15625047		1 Yes	No	No	Yes	Yes	No	Yes	No	No	Yes	Yes	No	76164.47	36780
27	124508	0.4	577	15738191		2 No	Yes	No	Yes	Yes	No	Yes	Yes	Yes	No	Yes	No	40518.12	21112
28	170042	0.21	756	15736816		1 Yes	No	Yes	Yes	Yes	No	No	Yes	Yes	No	Yes	No	33697.16	20356
29	38433	0.15	571	15700772		2 No	No	Yes	Yes	No	Yes	No	Yes	Yes	Yes	Yes	No	16386.4	19683
30	100187	0.08	574	15728693		1 Yes	Yes	Yes	No	Yes	No	Yes	No	Yes	Yes	No	No	22740.32	33808
31	53483	0.19	411	15656300		2 Yes	Yes	No	Yes	Yes	Yes	No	Yes	No	Yes	Yes	No	82037.18	30162
32	140469	0.12	591	15589475		3 Yes	Yes	Yes	No	Yes	Yes	Yes	Yes	No	Yes	Yes	Yes	67255.96	0
33	156732	0.23	533	15706552		1 No	No	No	No	Yes	No	Yes	No	Yes	Yes	No	No	26780.16	9714
34	81899	0.06	553	15750181		2 No	Yes	Yes	No	No	No	No	Yes	No	Yes	Yes	No	25750.44	3038

- 2 datasets:
- Customer Personal Info  
~7 columns and 10,000 rows
  - Financial Info  
~ 30 columns and 10,000 rows

# Description of Financial Info dataset (1)



30 columns (fields) and 10,000 rows

Fields	Description	Type
MonthlySalary	Monthly salary of customer	Numeric
RiskProfile	Evaluation of customer's willingness and ability to take risks	2 decimal points
CreditScore	The creditworthiness of customer, higher score indicates a good credit score	Numeric, range between 300-850
CustomerID	Primary key to uniquely identify a customer record in dataset	8 digits
NumberBankAccts	Number of bank accounts owned by customer	Numeric
FixedDeposit	Customer owns any fixed deposit accounts	Binary
CreditCard	Customer owns any credit cards	Binary
House	Customer owns any houses	Binary
Car	Customer owns any cars	Binary
LifeInsurance	Customer owns any life insurance	Binary
MedicalCards	Customer owns any medical cards	Binary
PensionPlan	Customer owns any pension plans	Binary
EducationFund	Customer owns any education funds	Binary
GovtBonds	Customer owns any government bonds	Binary
Share	Customer owns any stocks or equities	Binary

# Description of Financial Info dataset (2)



Fields	Description	Type
HousingLoan	Customer has any house loans	Binary
CarLoan	Customer has any car loans	Binary
Savings	Amount of savings possessed by customer	Numeric
Loan	Amount of loan owed by customer	Numeric
NetAssets	Total assets minus total liabilities possessed by customer	Numeric
Portfolio Return	Ratio of the net gain or loss realized by investments	4 decimal points
CashFlowRatio	Comparisons of cash flows to other elements of financial statements	4 decimal points
Taxes	Amount of taxes pay to government, such as income tax, real property gains tax etc	Numeric
NumberOfHouses	Number of houses owned by customer	Numeric
NumTransactions	Number of transactions made by customer in a particular month	Numeric
LastTransactionDate	Last transaction date made by customer	Date format
LastTransactionAmt	Last transaction amount made by customer	Numeric
ForeignAssets	Ratio of assets outside Malaysia to total assets possessed by customer	2 decimal points
Churn	Probability of customer that stopped using company's loan product	Binary
LoanApprovalRate	Percentage of customer getting a loan	2 decimal points

Download and save the datasets on your laptop



<https://qrgo.page.link/xcNkm>



# Contents

Overview of training datasets

## **Managing multiple worksheets**

Data understanding

Conditional formatting and logical operators

Lookup and reference functions

A review of essential functions

Fundamentals of descriptive statistics

Selecting the right charts to visualize your analysis result

Dashboard dos and don'ts



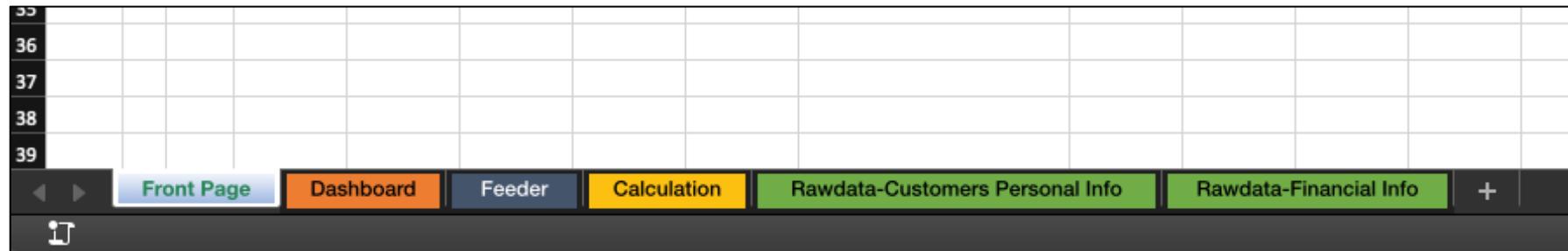
NATURAL INTELLIGENCE SOLUTIONS TECHNOLOGY

# Managing multiple excel worksheets



When starting with an analytics project, you should create multiple worksheets to:

- Organise your project better so that it is easier to find/locate the content
- Provide the users with a structure to identify the source of error
- Provides consistency and best practice across multiple users
- Colour coding the tabs will allow for quick reference to the content of the tab e.g. green = raw data



Worksheet Name	Description
Front Page	State your business objectives in this sheet, what does your dashboard address.
Dashboard	Create your dashboard in this sheet by putting all the charts together.
Feeder	Store your clean dataset or as the linkage between dashboard and raw data.
Calculation	Perform all calculations or create any tables, pivot tables, charts in this sheet.
Raw Data	Store your raw data in this sheet.

# Managing multiple worksheets



- 1. Insert new worksheet**
2. Copy a worksheet
3. Rename a worksheet
4. Change the worksheet tab colour
5. Delete a worksheet

43	181298	0.33	465	15738148		1 No	Yes	Yes	No	No
44	94154	0	556	15687946		1 Yes	Yes	Yes	No	Yes
45	194366	0.01	834	15755196		1 No	No	Yes	No	No
46	158338	0.38	660	15684171		1 Yes	No	No	No	Yes
47	126517	0.28	776	15754849		2 Yes	Yes	No	Yes	Yes
48	119708	0.14	829	15602280		1 Yes	Yes	No	Yes	Yes
49	117623	0.02	637	15771573		1 Yes	No	No	No	Yes
50	90878	0.2	550	15766205		1 No	Yes	No	Yes	Yes
51	194099	0.38	776	15771873		2 Yes	No	No	No	No
52	198059	0.18	698	15616550		2 Yes	Yes	No	Yes	No



43										
44										
45										
46										
47										
48										
49										
50										
51										
52										



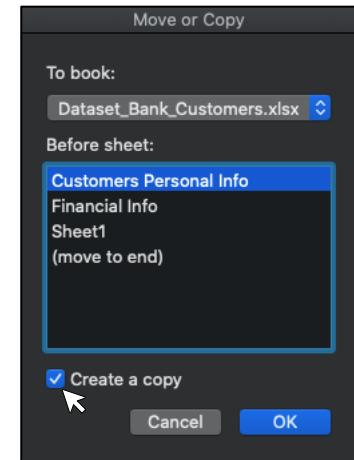
# Managing multiple worksheets



1. Insert new worksheet
2. **Copy a worksheet**
3. Rename a worksheet
4. Change the worksheet tab colour
5. Delete a worksheet

A screenshot of an Excel spreadsheet titled 'Customers Personal Info'. A context menu is open over a cell, with 'Move or Copy...' highlighted in blue. Other options in the menu include 'Insert Sheet', 'Delete', 'Rename', 'View Code', 'Protect Sheet...', 'Tab Colour', 'Hide', 'Unhide...', 'Select All Sheets', and 'Import Image'.

35	15814810	Kambomachi	France	Male	Yes	40	4
36	15814772	Adams	Germany	Male	Yes	49	1
37	15814757	Carter	Spain	Male	Yes	31	3
38	15814750	Ricci	Spain	Male	Yes	34	3
39	15814690	Chukwujekwu	Germany	Female	Yes	64	1
40	15814675	Chien	Germany	Female	Yes	39	0
41	15814664	Scott	Germany	Female	Yes	33	4
42	15814553	Ball	France	Female	Yes	34	1
43	15814536	Conti	France	Male	No	37	0
44	15814519	Kamdibe	France	Female	Yes	37	5
45	15814468	Wei	Germany	Male	Yes	50	0
46	15814465	Ch'in	France	Male	Yes	24	4
47	15814430	Ma	Spain	Male	No	41	1
48	15814405	Chesnokova	France	Female	Yes	46	1
49	15814331	Lung	Germany	Female	No	43	0
50	15814275	Zikoranachidimma	France	Male	No	33	0
51	15814268	Franklin	France	Female	Yes	40	2
52	15814267	Zhdanova	France	Male	No	22	1



A screenshot of the Excel ribbon. It shows two tabs: 'Customers Personal Info (2)' and 'Customers Personal Info'. Below the ribbon, the status bar displays 'Sheet1'.

35	15814810	Kambomachi	France	Male	Yes	40	4
36	15814772	Adams	Germany	Male	Yes	49	1
37	15814757	Carter	Spain	Male	Yes	31	3
38	15814750	Ricci	Spain	Male	Yes	34	3
39	15814690	Chukwujekwu	Germany	Female	Yes	64	1
40	15814675	Chien	Germany	Female	Yes	39	0
41	15814664	Scott	Germany	Male	Yes	33	4
42	15814553	Ball	France	Female	Yes	34	1
43	15814536	Conti	France	Male	No	37	0
44	15814519	Kamdibe	France	Female	Yes	37	5
45	15814468	Wei	Germany	Male	Yes	50	0
46	15814465	Ch'in	France	Male	Yes	24	4
47	15814430	Ma	Spain	Male	No	41	1
48	15814405	Chesnokova	France	Female	Yes	46	1
49	15814331	Lung	Germany	Female	No	43	0
50	15814275	Zikoranachidimma	France	Male	No	33	0
51	15814268	Franklin	France	Female	Yes	40	2
52	15814267	Zhdanova	France	Male	No	22	1

# Managing multiple worksheets



1. Insert new worksheet
2. Copy a worksheet
3. **Rename a worksheet**
4. Change the worksheet tab colour
5. Delete a worksheet

A screenshot of a spreadsheet application showing a context menu over a worksheet tab. The menu options include: Insert Sheet (F11), Delete, Rename (highlighted in blue), Move or Copy..., View Code, Protect Sheet..., Tab Colour, Hide, Unhide..., Select All Sheets, and Import Image. The worksheet tab 'Customers Personal Info' is selected. The background shows a table with data from rows 36 to 52.

	Customer ID	Name	Country	Gender	Is VIP	Age	Count
36	15814772	Adams	Germany	Male	Yes	49	1
37	15814757	Carter	Spain	Male	Yes	31	3
38	15814750	Ricci	Spain	Male	Yes	34	3
39	15814690	Chukwujekwu	Germany	Female	Yes	64	1
40	15814675	Chien			Yes	39	0
41	15814664	Scott			Yes	33	4
42	15814553	Ball			Yes	34	1
43	15814536	Conti			No	37	0
44	15814519	Kamdibe			Yes	37	5
45	15814468	Wei			Yes	50	0
46	15814465	Ch'in			Yes	24	4
47	15814430	Ma			No	41	1
48	15814405	Chesnokova			Yes	46	1
49	15814331	Lung			No	43	0
50	15814275	Zikoranachidimma			No	33	0
51	15814268	Franklin			Yes	40	2
52	15814267	Zhdanova			No	22	1



A screenshot of the spreadsheet application showing the same table as above, but with the worksheet tab 'Customers Personal Info' renamed to 'Calculation'. The tabs at the bottom are now labeled 'Calculation', 'Customers Personal Info', 'Financial Info', 'Sheet1', and '+'. A red arrow points from the first screenshot to this one, indicating the result of renaming the worksheet.

	Customer ID	Name	Country	Gender	Is VIP	Age	Count
36	15814772	Adams	Germany	Male	Yes	49	1
37	15814757	Carter	Spain	Male	Yes	31	3
38	15814750	Ricci	Spain	Male	Yes	34	3
39	15814690	Chukwujekwu	Germany	Female	Yes	64	1
40	15814675	Chien	Germany	Female	Yes	39	0
41	15814664	Scott	Germany	Male	Yes	33	4
42	15814553	Ball	France	Female	Yes	34	1
43	15814536	Conti	France	Male	No	37	0
44	15814519	Kamdibe	France	Female	Yes	37	5
45	15814468	Wei	Germany	Male	Yes	50	0
46	15814465	Ch'in	France	Male	Yes	24	4
47	15814430	Ma	Spain	Male	No	41	1
48	15814405	Chesnokova	France	Female	Yes	46	1
49	15814331	Lung	Germany	Female	No	43	0
50	15814275	Zikoranachidimma	France	Male	No	33	0
51	15814268	Franklin	France	Female	Yes	40	2
52	15814267	Zhdanova	France	Male	No	22	1

# Managing multiple worksheets



1. Insert new worksheet
2. Copy a worksheet
3. Rename a worksheet
4. **Change the worksheet tab colour**
5. Delete a worksheet



Customer ID	Customer Name	Country	Gender	Is VIP	Age	Region
36	15814772 Adams	Germany	Male	Yes	49	1
37	15814757 Carter	Spain	Male	Yes	31	3
38	15814750 Ricci	Spain	Male	Yes	34	3
39	15814690 Chukwujekwu	Germany	Female	Yes	64	1
40	15814675 Chien				39	0
41	15814664 Scott				33	4
42	15814553 Ball				34	1
43	15814536 Conti				37	0
44	15814519 Kamdibe				5	
45	15814468 Wei				50	0
46	15814465 Ch'in				24	4
47	15814430 Ma	Spain	Male	No	41	1
48	15814405 Chesnokova	France	Female	Yes	46	1
49	15814331 Lung	Germany	Female	No	43	0
50	15814275 Zikoranachidimma	France	Male	No	33	0
51	15814268 Franklin	France	Female	Yes	40	2
52	15814267 Zhdanova	France	Male	No	22	1

# Managing multiple worksheets



1. Insert new worksheet
  2. Copy a worksheet
  3. Rename a worksheet
  4. Change the worksheet tab colour
  5. **Delete a worksheet**



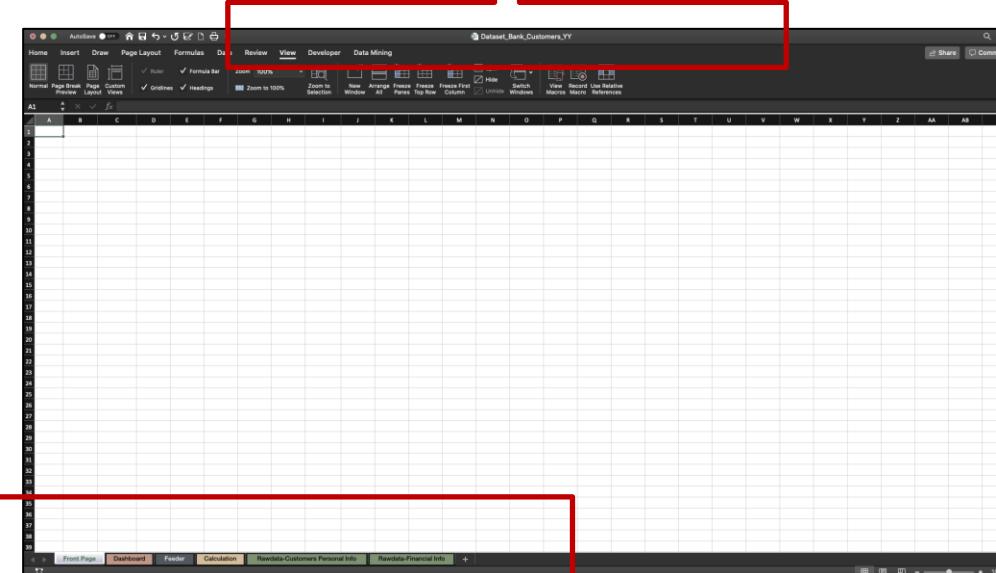
36	142033	0.01	722	15732963	2	Yes	No	Yes	Yes	Yes	Yes	
37	27823	0.35	475	15794171	1	Yes	No	No	No	No	No	
38	114067	0.38	490	15788448	1	No	Yes	Yes	No	Yes		
39	98453	0.09	804	15729599	1	No	Yes	Yes	No	Yes		
40	40813	0.09	850	15717426	1	Yes	No	No	No	Yes	Yes	
41	178074	0.4	582	15585768	2	No	No	No	Yes	Yes	Yes	
42	70154	0.24	472	15619360	1	Yes	No	No	No	No	No	
43	181298	0.33	465	15738148	1	No	Yes	Yes	No	No	No	
44	94154	0	556	15687946	1	Yes	Yes	Yes	No	Yes		
45	194366	0.01	834	15755196	1	No	No	Yes	No	No	No	
46	158338	0.38	660	15684171	1	Yes	No	No	No	Yes		
47	126517	0.28	776	15754849	2	Yes	Yes	No	Yes	Yes	Yes	
48	119708	0.14	829	15602280	1	Yes	Yes	No	Yes	Yes	Yes	
49	117623	0.02	637	15771573	1	Yes	No	No	No	Yes		
50	90878	0.2	550	15766205	1	No	Yes	No	Yes	Yes	Yes	
51	194099	0.38	776	15771873	2	Yes	No	No	No	No	No	
52	198059	0.18	698	15616550	2	Yes	Yes	No	Yes	No	No	



# Setting up your excel environment

Set up your workbook environment as below and save your Excel file name in the following format:

e.g. Dataset\_Bank\_Customers\_[Your Initial].



The screenshot shows the Excel ribbon with the following tabs visible:

- Front Page (highlighted in blue)
- Dashboard (highlighted in orange)
- Feeder (highlighted in dark blue)
- Calculation (highlighted in yellow)
- Rawdata-Customers Personal Info (highlighted in green)
- Rawdata-Financial Info (highlighted in green)
- A plus sign (+) tab

# Contents

Overview of training datasets

Managing multiple worksheets

## **Data understanding**

Conditional formatting and logical operators

Lookup and reference functions

A review of essential functions

Fundamentals of descriptive statistics

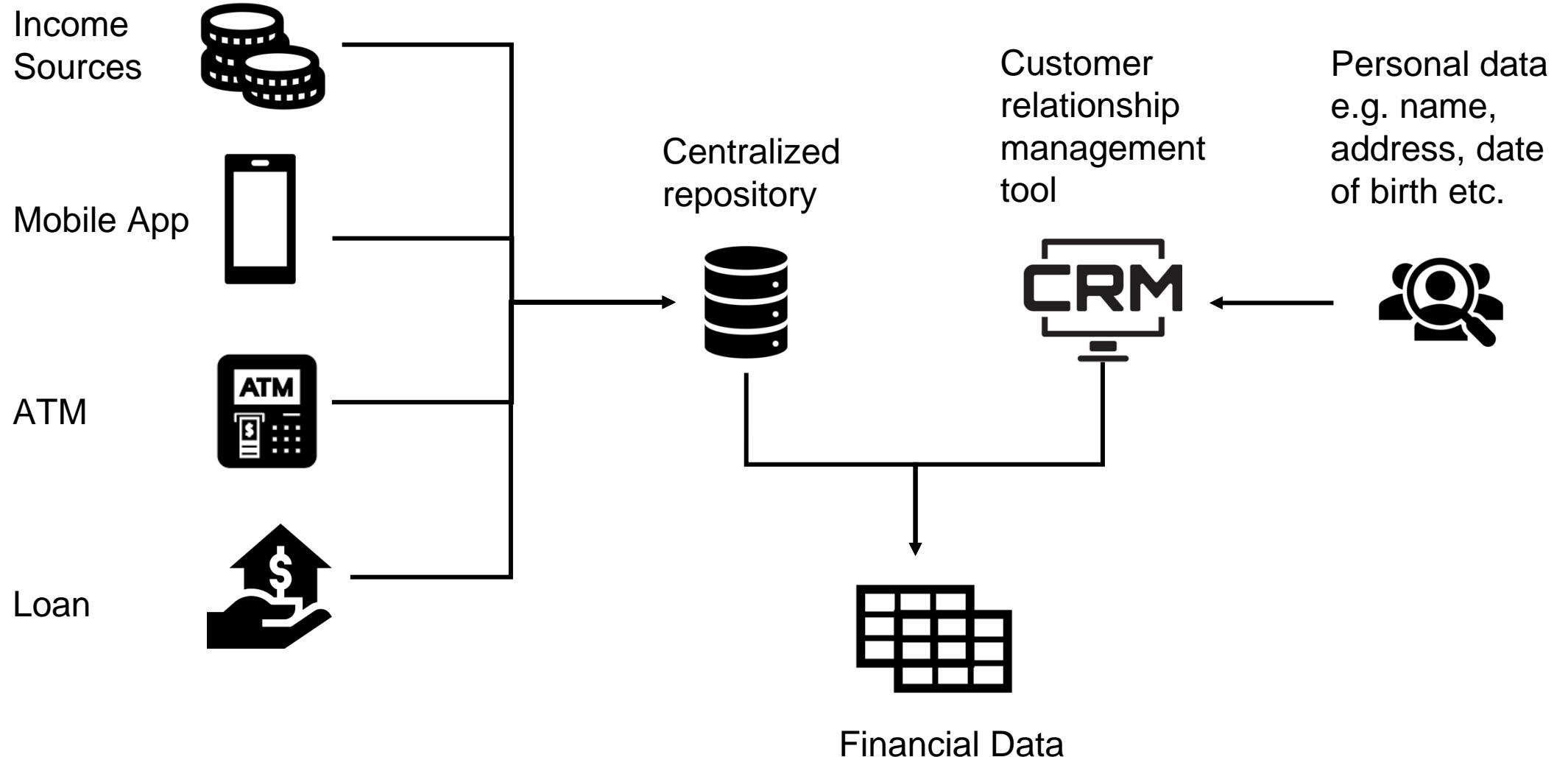
Selecting the right charts to visualize your analysis result

Dashboard dos and don'ts



NATURAL INTELLIGENCE SOLUTIONS TECHNOLOGY

# Example of data and data flow



# Example of financial data



We will be using this financial data as an example to elaborate the topics in this course.

The diagram illustrates the structure of the financial data. It shows a table with 14 columns. The first 13 columns are labeled "Attributes" at the top, and the last column is labeled "Target attribute". A bracket on the left side of the table is labeled "Observations". Brackets at the bottom are labeled "Record data" under the first 13 columns and "Transaction data" under the last column. An arrow points from the "Target attribute" label to the last column of the table.

Attributes													Target attribute
Observations	CustomerID	Surname	State	Gender	Marital Status	Age	Number Bank Accounts	Has Credit Card	...	Last Transaction Date	Last Transaction Amount	Able To Get Loan	
	15634602	Fatimah	Johor	Female	1	42	1	1	...	1/27/2019	2095.3	0	
	15647311	Chin	Selangor	Female	1	41	1	0	...	2/25/2019	9955.2	0	
	15619304	Priya	Kedah	Female	1	42	3	1	...	11/5/17	9123.1	1	
	15701354	Siti	Selangor	Female	1	39	2	0	...	1/7/19	5973.3	0	

Record data

Transaction data

# Terminology



The following terms are used interchangeably:

- **Observation**, object, row, datapoint
- **Attribute**, feature, column, variable
- **Target attribute**, target, label, target variable

A target attribute is the attribute we are interested in modelling, predicting or finding out the answer to.

# Data types



Data type is the most basic and the most common classification of data:

Qualitative (or categorical) data

Nominal Data

Categories without ordering  
e.g. colour, gender

Ordinal Data

Ordered categories (rankings, order or scaling)  
e.g. satisfaction level of the workshop

Quantitative (or numerical) data

- Discrete
- Continuous

Interval Data

Differences between measurements but no true zero  
e.g. temperature, year

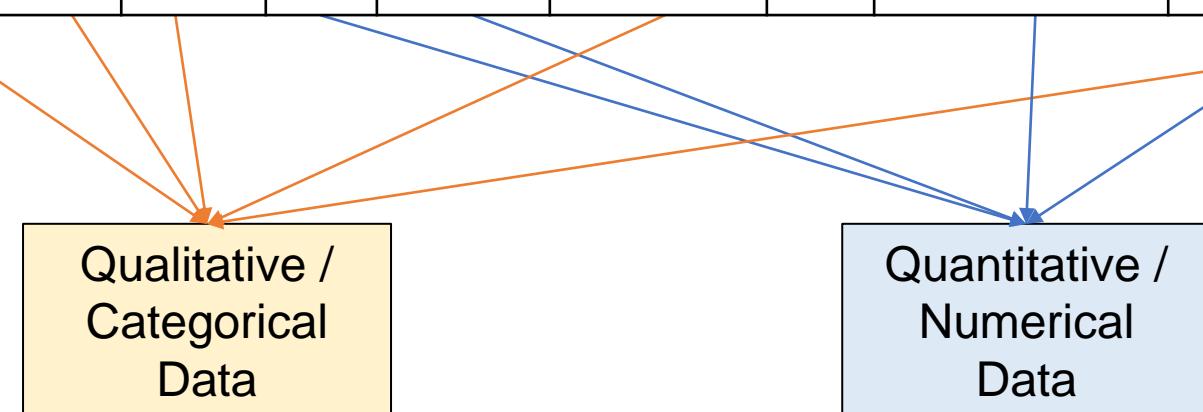
Ratio Data

Differences between measurements, true zero exists  
e.g. height, weight, money

# Returning to financial data example



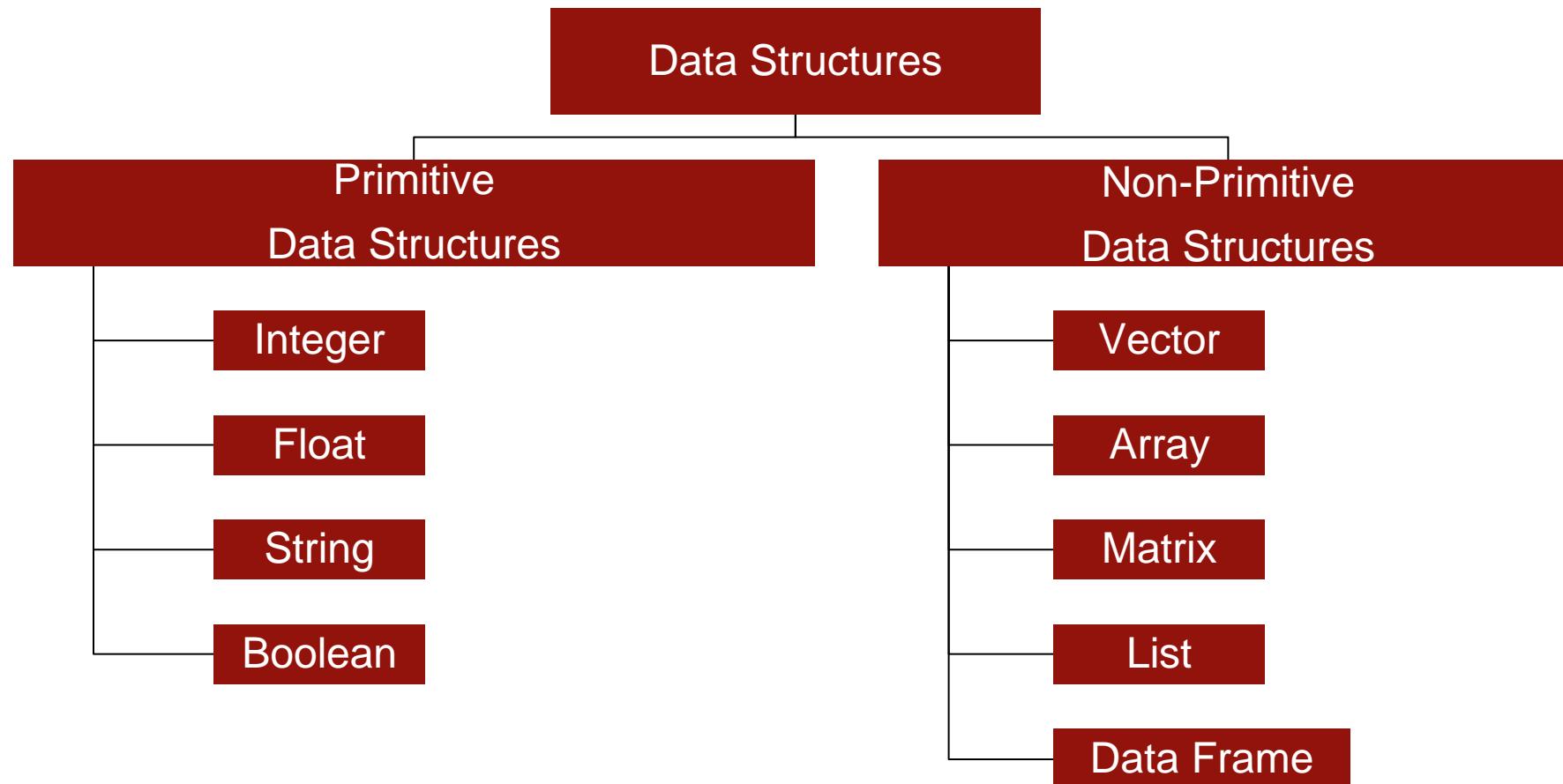
CustomerID	Surname	State	Gender	Marital Status	Age	Number Bank Accounts	Has Credit Card	...	Last Transaction Date	Last Transaction Amount	Able To Get Loan
15634602	Fatimah	Johor	Female	1	42	1	1	...	1/27/2019	2095.3	0
15647311	Chin	Selangor	Male	1	41	1	0	...	2/25/2019	9955.2	0
15619304	Priya	Kedah	Female	1	42	3	1	...	11/5/17	9123.1	1
15701354	Siti	Selangor	Female	1	39	2	0	...	1/7/19	5973.3	0



# Data structures



A data structure is a specialized format for organizing, processing, retrieving and storing data. We need to organize data in a structured manner so that we can retrieve the information easier and faster.



# Primitive data structure



Primitive data structures are basic structures and are directly operated upon by machine instructions.

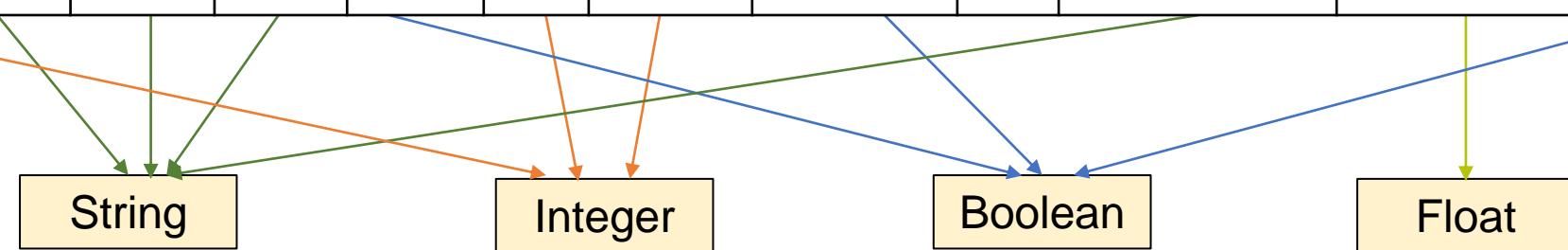
Data structures	Used for
Integer	It is a data type which allows all values without fraction part. We can use it for whole numbers.
Float	It is a data type which use for storing fractional numbers.
String	It is a data type which use for alphanumeric characters.
Boolean	It is a data type which representing logical values.

# Returning to financial data example



What is the primitive data structure for each data field?

CustomerID	Surname	State	Gender	Marital Status	Age	Number Bank Accounts	Has Credit Card	...	Last Transaction Date	Last Transaction Amount	Able To Get Loan
15634602	Fatimah	Johor	Female	1	42	1	1	...	1/27/2019	2095.3	0
15647311	Chin	Selangor	Male	1	41	1	0	...	2/25/2019	9955.2	0
15619304	Priya	Kedah	Female	1	42	3	1	...	11/5/17	9123.1	1
15701354	Siti	Selangor	Female	1	39	2	0	...	1/7/19	5973.3	0



# Non-primitive data structures



The non-primitive data structures emphasize on structuring of a group of homogeneous or heterogeneous data items.

Data Structures	Used for
Array	A collection of cells with a fixed size where all cells hold the same type (integers or characters or reals or whatever).
Vector	A one dimensional array (fixed size, all cell types the same).
Matrix	A two-dimensional array (fixed size, all cell types the same).
n-dimensional array	A three or more dimensional array (fixed size, all cell types the same).
List	A list can hold items of different types.
Data Frame	A data frame is called a table in most languages. Each column holds the same type, and the columns can have header names.

# Non-primitive data structure example



What is the non-primitive data structure for each data field?

EmpID	Employee_Name	MarriedID	State	Postcode	DateofHire	Termination Status	Termination Reason	...	Engagement Survey	PayRate	Performance Score
1202031821	Choong Ying Li	1	Selangor	42472	7/5/11	Yes	unhappy	...	1.03	28.0	Needs Improvement
1107027551	Norul Atiah binti Osman	0	Selangor	41778	1/10/11	Yes	unhappy	...	1.3	22.75	Fully Meets
1012023226	Sharniya Nair a/p Devadas	1	Selangor	41752	16/5/11	Yes	unhappy	...	1.16	25.2	Fully Meets
1411071324	Mohamad Rashdan bin Mohammad Sharuddin	1	Selangor	42128	26/4/10	Yes	career change	...	4.35	29.0	PIP
1306057810	Siti Fathiah binti Saliman	0	Selangor	42128	7/7/14	No	N/A - still employed	...	1.51	25.0	Exceeds

Data Frame

List  
(different type)

Vector /  
1-D Array  
(same type)

2 x 2 Matrix /  
2-D Array

1.03	28.0
1.3	22.75

3-D Array

# Contents

- Overview of training datasets
- Managing multiple worksheets
- Data understanding
- Conditional formatting and logical operators**
- Lookup and reference functions
- A review of essential functions
- Fundamentals of descriptive statistics
- Selecting the right charts to visualize your analysis result
- Dashboard dos and don'ts



NATURAL INTELLIGENCE SOLUTIONS TECHNOLOGY



# Dealing with the duplicate records

Duplicate data is a serious issue for any company using multiple platforms to manage their data. It occurs when an exact copy of a record is created as a different entry in the same database.

1. Record data – validate and remove the exact duplications
2. Transactional data – validate and aggregate the values if necessary

CustomerID	LastName	State	Gender	MaritalSt	Age	NumberofDependents	AgeGreaterAndEqualTo60	NrOfDependentLessThan2	StateSelangor
15798964	NKEMAKONAM	Sarawak	Male	No	33	1	FALSE	TRUE	TRUE
15798964	NKEMAKONAM	Sarawak	Male	No	33	1	FALSE	TRUE	TRUE
15794171	LOMBARDO	Sabah	Female	No	45	1	FALSE	TRUE	TRUE
15794171	LOMBARDO	Sabah	Female	No	45	1	FALSE	TRUE	TRUE
15788448	WATSON	Pahang	Male	Yes	31	3	FALSE	FALSE	TRUE
15788448	WATSON	Pahang	Male	Yes	31	3	FALSE	FALSE	TRUE
15788218	HENDERSON	Pahang	Female	No	24	1	FALSE	TRUE	TRUE
15788218	HENDERSON	Pahang	Female	No	24	1	FALSE	TRUE	TRUE
15769959	AJULUCHUKWU	Kedah	Female	Yes	53	1	FALSE	TRUE	TRUE
15769959	AJULUCHUKWU	Kedah	Female	Yes	53	1	FALSE	TRUE	TRUE
15768163	GRIFFIN	Melaka	Female	Yes	46	4	FALSE	FALSE	TRUE
15768163	GRIFFIN	Melaka	Female	Yes	46	4	FALSE	FALSE	TRUE
15750181	SANDERSON	Johor	Male	Yes	41	0	FALSE	TRUE	TRUE
15750181	SANDERSON	Johor	Male	Yes	41	0	FALSE	TRUE	TRUE
15738191	MACLEAN	Johor	Male	Yes	25	4	FALSE	FALSE	TRUE
15738191	MACLEAN	Johor	Male	Yes	25	4	FALSE	FALSE	TRUE
15738148	CLARKE	Pahang	Female	Yes	51	4	FALSE	FALSE	TRUE
15738148	CLARKE	Pahang	Female	Yes	51	4	FALSE	FALSE	TRUE
15737452	ROMEO	Selangor	Male	No	58	0	FALSE	TRUE	FALSE
15737452	ROMEO	Selangor	Male	No	58	0	FALSE	TRUE	FALSE
15736816	YOUNG	Terengganu	Male	Yes	36	4	FALSE	FALSE	TRUE
15736816	YOUNG	Terengganu	Male	Yes	36	4	FALSE	FALSE	TRUE
15732963	CLEMENTS	Melaka	Female	Yes	29	2	FALSE	FALSE	TRUE
15732963	CLEMENTS	Melaka	Female	Yes	29	2	FALSE	FALSE	TRUE
15729599	LORENZO	Kedah	Male	Yes	33	0	FALSE	TRUE	TRUE
15729599	LORENZO	Kedah	Male	Yes	33	0	FALSE	TRUE	TRUE
15728693	MCWILLIAMS	Melaka	Female	Yes	43	2	FALSE	FALSE	TRUE
15728693	MCWILLIAMS	Melaka	Female	Yes	43	2	FALSE	FALSE	TRUE
15725737	MOSMAI	Negeri Semb	Male	No	46	0	FALSE	TRUE	TRUE
15725737	MOSMAI	Negeri Semb	Male	No	46	0	FALSE	TRUE	TRUE
15719294	WOOD	Perlis	Female	Yes	29	2	FALSE	FALSE	TRUE
15719294	WOOD	Perlis	Female	Yes	29	2	FALSE	FALSE	TRUE
15719276	TAO	Terengganu	Male	Yes	35	0	FALSE	TRUE	TRUE

# Some examples of conditional formatting



A conditional format changes the appearance of cells on the basis of conditions that you specify. If the conditions are true, the cell range is formatted; if the conditions are false, the cell range is not formatted.

- 1. Highlight duplicates in column**
2. Highlight cells with formulas
3. Highlight top or bottom values
4. Highlight data with a colour scale
5. Highlight upcoming expiry dates/ weekend dates



# Conditional formatting

Use conditional formatting to help you visually explore and analyse data, detect critical issues, and identify patterns and trends.

The screenshot shows the Microsoft Excel ribbon with the 'Home' tab selected (step 1). In the 'Conditional Formatting' dropdown menu (step 2), the 'New Rule...' option is highlighted (step 3). The Excel interface includes a worksheet titled 'Dataset\_Bank\_Customers\_YY' with columns A through J and rows 1 through 34. The bottom navigation bar shows tabs for 'Front Page', 'Dashboard', 'Feeder', 'Calculation', 'Customers Personal Info', and 'Financial Info'. The status bar at the bottom right indicates a 100% zoom level.

# Try it now – highlight duplicate values



Go to Calculation worksheet, use conditional formatting to highlight duplicate records in CustomerID column.

The screenshot shows a Microsoft Excel spreadsheet titled "Dataset\_Bank\_Customers\_YY". The spreadsheet contains data for 10048 customers across various columns including CustomerID, LastName, State, Gender, MaritalStatus, Age, NumberofDependents, and AgeGreaterAndEqualTo. The "Calculation" tab is selected at the bottom. The "Conditional Formatting" dropdown menu is open, and the "Highlight Cells Rules" option is selected. A submenu is displayed with "Duplicate Values..." highlighted in blue. The status bar at the bottom shows "Average: 15690874.59 Count: 10048 Sum: 1.57646E+11".

# Filter duplicate values



You can filter the duplicate records by selecting either Cell Colour or Font Colour to view the highlighted duplicate records.

The screenshot shows a Microsoft Power BI interface with the 'Dataset\_Bank\_Customers\_YY' dataset open. The table has columns: CustomerID, Surname, Forename, Gender, Age, NumberofDependents, AgeGreaterAndEqualTo60, NrOfDependentLessThan2, and StateSelangor. A filter is applied to the 'CustomerID' column, specifically targeting 'Cell Colour'. The dropdown menu shows 'None' is selected, and 'Cell Colour' is chosen. Under 'Cell Colour', 'Rose' is selected, which highlights all rows where the CustomerID value is 15798964. Other CustomerID values like 15565701, 15565706, 15565714, 15565779, 15565796, 15565806, and 15738191 are also present in the list but not highlighted. The Power BI ribbon at the top includes tabs for Home, Insert, Review, View, Developer, Data Mining, Share, Comments, Sort & Filter, Find & Select, Ideas, and Sensitivity. The bottom navigation bar includes Front Page, Dashboard, Feeder, Calculation, Customers Personal Info, Financial Info, and other tabs.



# Try it now – remove duplicate records

The data we are validating now is Record data, so we can remove the duplicate records.  
Highlight the entire dataset and click Remove Duplicates under Data tab in Ribbon.

A screenshot of Microsoft Excel showing a dataset titled "Dataset\_Bank\_Customers\_YY". The Data ribbon tab is selected. The "Remove Duplicates" button in the Data Tools group is highlighted with a cursor. The dataset contains 34 rows of customer information, including columns for CustomerID, LastName, State, Gender, MaritalStatus, Age, NumberofDependents, AgeGreaterAndEqualTo60, NrOfDependentLessThan2, and StateSelangor. The rows show various entries for customers from different states like Sarawak, Sabah, Pahang, Johor, Kedah, Melaka, and Terengganu.

	CustomerID	LastName	State	Gender	MaritalStatus	Age	NumberofDependents	AgeGreaterAndEqualTo60	NrOfDependentLessThan2	StateSelangor
1	15798964	NKEMAKONAM	Sarawak	Male	No	33	1	FALSE	TRUE	TRUE
2	15798964	NKEMAKONAM	Sarawak	Male	No	33	1	FALSE	TRUE	TRUE
3	15794171	LOMBARDO	Sabah	Female	No	45	1	FALSE	TRUE	TRUE
4	15794171	LOMBARDO	Sabah	Female	No	45	1	FALSE	TRUE	TRUE
5	15788448	WATSON	Pahang	Male	Yes	31	3	FALSE	FALSE	TRUE
6	15788448	WATSON	Pahang	Male	Yes	31	3	FALSE	FALSE	TRUE
7	15788218	HENDERSON	Pahang	Female	No	24	1	FALSE	TRUE	TRUE
8	15788218	HENDERSON	Pahang	Female	No	24	1	FALSE	TRUE	TRUE
9	15769959	AJULUCHUKWU	Kedah	Female	Yes	53	1	FALSE	TRUE	TRUE
10	15769959	AJULUCHUKWU	Kedah	Female	Yes	53	1	FALSE	TRUE	TRUE
11	15768163	GRIFFIN	Melaka	Female	Yes	46	4	FALSE	FALSE	TRUE
12	15768163	GRIFFIN	Melaka	Female	Yes	46	4	FALSE	FALSE	TRUE
13	15750181	SANDERSON	Johor	Male	Yes	41	0	FALSE	TRUE	TRUE
14	15750181	SANDERSON	Johor	Male	Yes	41	0	FALSE	TRUE	TRUE
15	15738191	MACLEAN	Johor	Male	Yes	25	4	FALSE	FALSE	TRUE
16	15738191	MACLEAN	Johor	Male	Yes	25	4	FALSE	FALSE	TRUE
17	15738148	CLARKE	Pahang	Female	Yes	51	4	FALSE	FALSE	TRUE
18	15738148	CLARKE	Pahang	Female	Yes	51	4	FALSE	FALSE	TRUE
19	15737452	ROMEO	Selangor	Male	No	58	0	FALSE	TRUE	FALSE
20	15737452	ROMEO	Selangor	Male	No	58	0	FALSE	TRUE	FALSE
21	15736816	YOUNG	Terengganu	Male	Yes	36	4	FALSE	FALSE	TRUE
22	15736816	YOUNG	Terengganu	Male	Yes	36	4	FALSE	FALSE	TRUE
23	15732963	CLEMENTS	Melaka	Female	Yes	29	2	FALSE	FALSE	TRUE
24	15732963	CLEMENTS	Melaka	Female	Yes	29	2	FALSE	FALSE	TRUE
25	15729599	LORENZO	Kedah	Male	Yes	33	0	FALSE	TRUE	TRUE
26	15729599	LORENZO	Kedah	Male	Yes	33	0	FALSE	TRUE	TRUE
27	15728693	MCWILLIAMS	Melaka	Female	Yes	43	2	FALSE	FALSE	TRUE
28	15728693	MCWILLIAMS	Melaka	Female	Yes	43	2	FALSE	FALSE	TRUE
29	15725737	MOSMAN	Negeri Semb	Male	No	46	0	FALSE	TRUE	TRUE
30	15725737	MOSMAN	Negeri Semb	Male	No	46	0	FALSE	TRUE	TRUE
31	15719294	WOOD	Perlis	Female	Yes	29	2	FALSE	FALSE	TRUE
32	15719294	WOOD	Perlis	Female	Yes	29	2	FALSE	FALSE	TRUE
33	15719276	TAO	Terengganu	Male	Yes	35	0	FALSE	TRUE	TRUE
34										

# Check it now



Go to your dataset in Financial Info worksheet. Can you find any duplicate records?

Repeat these steps:

1. Apply conditional formatting to highlight the duplicate values.
2. Filter the records by cell or font colour.
3. If there is duplicate records, you need to select the entire dataset and apply remove duplicates under Data tab in Ribbon.
4. If there is no duplication, your data is now ready for next step – manipulation and analysis.

# Logical operators



**Logical operators** can be used in formulas to perform comparisons, and to build formula criteria. A common use of logical operators is found in Excel's IF function. We will use the logical operators again and again in this 4-days training during data understanding and data preparation phases.

Operator	Description	Example	Results
=	Equal to	=A1=B1	If the value in A1 is equal to the value in B1, the comparison result is TRUE, otherwise FALSE.
<>	Not equal to	=A1<>B1	If the value in A1 is not equal to the value in B1, the comparison result is TRUE, otherwise FALSE.
>	Greater than	=A1>B1	If the value in A1 is greater than the value in B1, the comparison result is TRUE, otherwise FALSE.
>=	Greater than and equal to	=A1>=B1	If the value in A1 is greater than and equal to the value in B1, the comparison result is TRUE, otherwise FALSE.
<	Less than	=A1<B1	If the value in A1 is less than the value in B1, the comparison result is TRUE, otherwise FALSE.
<=	Less than and equal to	=A1<=B1	If the value in A1 is less than and equal to the value in B1, the comparison result is TRUE, otherwise FALSE.

# Try it now



Go to **Calculation** worksheet, sort the CustomerID by ascending order.

1. Check if age is greater and equal to 60. (create a new column, column name: Age>=60)
  - a. Click on Calculation sheet
  - b. Add a new column: “Age>=60”
  - c. In cell H2, type “=F2>=60”, and hit enter button
  - d. Double click at the bottom right of cell H2, formula will be filled in this new column
  - e. The answer will be either TRUE or FALSE
2. Check if Number of Dependents is less than 2. (create a new column, column name: Dependents<2)
3. Check if State is not equal to Selangor. (create a new column, column name: StateSelangor)

# Contents

- Overview of training datasets
- Managing multiple worksheets
- Data understanding
- Conditional formatting and logical operators
- Lookup and reference functions**
- A review of essential functions
- Fundamentals of descriptive statistics
- Selecting the right charts to visualize your analysis result
- Dashboard dos and don'ts



NATURAL INTELLIGENCE SOLUTIONS TECHNOLOGY

# Lookup and reference functions



Lookup & Reference functions help you to work with data array, and are particularly useful when you need to cross reference between different data sets. They perform tasks such as providing information about a range, returning the location of a given address or value, or looking up specific values.

For example, we have 2 datasets: Customers Personal Info & Financial Info, and we want to know the transaction amount made by the customers from Selangor. But the information of the customer's origin cannot be found in transactional data, so we need to consolidate these 2 information together: transactions amount from financial info dataset into customer personal info dataset. We can use lookup & reference functions to insert the required values based on the unique key – CustomerID.

Function	Description	Syntax
VLOOKUP	Look up a value in a table by matching on the <b>first column</b>	=VLOOKUP(lookup_value, table_array, col_index_num, [range_lookup])
HLOOKUP	Look up a value in a table by matching on the <b>first row</b>	=HLOOKUP(lookup_value, table_array, row_index_num, [range_lookup])
INDEX	Get a <b>value</b> in a list or table based on location	=INDEX(array, row_num, [col_num])
MATCH	Get the <b>position</b> of an item in an array	=MATCH(lookup_value, lookup_array, [match_type])



# LOOKUP function

HLOOKUP is similar to VLOOKUP, but it searches a row instead of a column, and the result is offset by a row index number. The V in VLOOKUP stands for vertical search (in a single column), while the H in HLOOKUP stands for horizontal search (within a single row).

1. The dataset / table we have.

	A	B	C	D	E	F	G
1	CustomerID	LastName	State	Gender	MaritalStatus	Age	NumberofDependents
2	15565701	FERRI	Johor	Female	Yes	39	0
3	15565706	RAHMAN	Terengganu	Male	Yes	35	0
4	15565714	ADI	Perlis	Male	Yes	47	1
5	15565779	KENT	Kedah	Female	Yes	30	4
6	15565796	HAY	Terengganu	Male	Yes	48	0
7	15565806	AMMAR	Terengganu	Male	No	38	1
8	15565878	ALEX	Pahang	Male	Yes	29	3
9	15565879	RILEY	Kedah	Female	No	28	1
10	15565891	ADAM	Johor	Male	No	39	0
11	15565996	ALAN	Terengganu	Male	Yes		

VLOOKUP function

	A	B	C	D	E	F	G	H	I	J	K
1	CustomerID	15565701	15565706	15565714	15565779	15565796	15565806	15565878	15565879	15565891	15565996
2	LastName	FERRI	RAHMAN	ADI	KENT	HAY	AMMAR	ALEX	RILEY	ADAM	ALAN
3	State	Johor	Terengganu	Perlis	Kedah	Terengganu	Terengganu	Pahang	Kedah	Johor	Terengganu
4	Gender	Female	Male	Male	Female	Male	Male	Male	Female	Male	Male
5	MaritalStatus	Yes	Yes	Yes	Yes	Yes	No	Yes	No	No	Yes
6	Age	39	35	47	30	48	38	29	28	39	44
7	NumberofDependents	0	0	1	4	0	1	3	1	0	3

HLOOKUP function

2. The output table we want to generate.

CustomerID	LastName	Age
15565996		
15565879		
15565714		
15565796		
15565701		
15565806		
15565706		
15565878		
15565779		
15565891		



# VLOOKUP function

=VLOOKUP(lookup\_value, table\_array, col\_index\_num, [range\_lookup])

The dataset / table we have.

	A	B	C	D	E	F	G
1	CustomerID	LastName	State	Gender	MaritalStatus	Age	NumberofDependents
2	15565701	FERRI	Johor	Female	Yes	39	0
3	15565706	RAHMAN	Terengganu	Male	Yes	35	0
4	15565714	ADI	Perlis	Male	Yes	47	1
5	15565779	KENT	Kedah	Female	Yes	30	4
6	15565796	HAY	Terengganu	Male	Yes	48	0
7	15565806	AMMAR	Terengganu	Male	No	38	1
8	15565878	ALEX	Pahang	Male	Yes	29	3
9	15565879	RILEY	Kedah	Female	No	28	1
10	15565891	ADAM	Johor	Male	No	39	0
11	15565996	ALAN	Terengganu	Male	Yes	44	3

b. table array

1	CustomerID	LastName	State	Gender	MaritalStatus	Age	NumberofDependents
2	15565701	FERRI	Johor	Female	Yes	39	0
3	15565706	RAHMAN	Terengganu	Male	Yes	35	0
4	15565714	ADI	Perlis	Male	Yes	47	1
5	15565779	KENT	Kedah	Female	Yes	30	4
6	15565796	HAY	Terengganu	Male	Yes	48	0
7	15565806	AMMAR	Terengganu	Male	No	38	1
8	15565878	ALEX	Pahang	Male	Yes	29	3
9	15565879	RILEY	Kedah	Female	No	28	1
10	15565891	ADAM	Johor	Male	No	39	0
11	15565996	ALAN	Terengganu	Male	Yes	44	3

c. col index num

starting with 1 for the left-most column of table-array (CustomerID) to the return value (LastName)

The table we want to generate.

	CustomerID	LastName	Age
17	CustomerID	LastName	Age
18	15565996	=VLOOKUP(A18,\$A\$1:\$B\$11,2,0)	
19	15565879		
20	15565714		
21	15565796		
22	15565701		
23	15565806		
24	15565706		
25	15565878		
26	15565779		
27	15565891		
28			

a. lookup value

d. [range lookup]

TRUE / 1: Approximate match, that is, if an exact match is not found, use the closest match below the lookup\_value.

FALSE / 0: Exact match, that is, if an exact match not found, then it will return an error.



# HLOOKUP function

=HLOOKUP(lookup\_value, table\_array, row\_index\_num, [range\_lookup])

The dataset / table we have.

	A	B	C	D	E	F	G	H	I	J	K
1	CustomerID	15565701	15565706	15565714	15565779	15565796	15565806	15565878	15565879	15565891	15565996
2	LastName	FERRI	RAHMAN	ADI	KENT	HAY	AMMAR	ALEX	RILEY	ADAM	ALAN
3	State	Johor	Terengganu	Perlis	Kedah	Terengganu	Terengganu	Pahang	Kedah	Johor	Terengganu
4	Gender	Female	Male	Male	Female	Male	Male	Male	Female	Male	Male
5	MaritalStatus	Yes	Yes	Yes	Yes	Yes	No	Yes	No	No	Yes
6	Age	39	35	47	30	48	38	29	28	39	44
7	NumberofDependents	0	0	1	4	0	1	3	1	0	3

b. table array

c. row index num

starting with 1 for the first row of table-array (CustomerID) to the return value (LastName)

The table we want to generate.

	CustomerID	LastName	Age
17			
18		15565996	=HLOOKUP(A18,\$A\$1:\$K\$2,2,0)
19		15565879	
20		15565714	
21		15565796	
22		15565701	
23		15565806	
24		15565706	
25		15565878	
26		15565779	
27		15565891	

a. lookup value

d. [range lookup]

TRUE / 1: Approximate match, that is, if an exact match is not found, use the closest match below the lookup\_value.

FALSE / 0: Exact match, that is, if an exact match not found, then it will return an error.

# Try it now



Go to **Calculation** worksheet, lookup data from **Financial Info** worksheet.

1. NumberBankAccts (add a new column, column name: NumberBankAccts)
  - a. Add a new column at column K: "NumberBankAccts"
  - b. At cell K2, type =VLOOKUP(A2,'Rawdata-Financial Info'!D:E,2,0) and hit enter button
  - c. Double click at the bottom right of cell K2, formula will be filled in column K
  - d. The value at column K is now lookup from Financial Info dataset column E based on the unique key CustomerID
2. Savings (add a new column, column name: Savings)
3. LastTransactionDate (add a new column, column name: LastTransactionDate)
4. LastTransactionAmt (add a new column, column name: LastTransactionAmt)

# Try it now



Create a new worksheet named **Exercise**, write the formulas to lookup the values:

1.

CustomerID	15565714	15576269	15692718	15789874	15602084	15783305	15757933
LastName							
State							
Age							
House							
HousingLoan							

2. CustomerID 15757933

LastName	Age	House	HousingLoan

3. Change the CustomerID in Question 2 and compare your answers with your neighbours.
- 15576269
  - 15789874

# How about this



Go back to your **Calculation** worksheet, lookup data from **Financial Info** worksheet.

1. MonthlySalary (add a new column, column name: MonthlySalary) ??
2. CreditScore (add a new column, column name: CreditScore) ??

\* There is a limitation to perform lookup function for Q5 and Q6, because the CustomerID unique key is at the right of lookup value in Financial Info dataset. An alternative way to perform lookup function is by using INDEX MATCH function. We will revisit this exercise again later on.

# VLOOKUP vs INDEX MATCH



Differences between VLOOKUP and INDEX MATCH

VLOOKUP	INDEX MATCH
Require a static column reference	Require a dynamic column reference
It breaks the formula each time when you add or delete a new column	You can add or delete columns without distorting the array
Total length of your lookup criteria should not exceed 255 characters, otherwise you will end up having the #VALUE! error	Can lookup values more than 255 characters in length
Only work if the lookup value is in the first column, cannot look to its left	Performs the lookup both horizontally and vertically, doesn't require the lookup value to be in the first column, it can be anywhere

# VLOOKUP vs INDEX MATCH



The INDEX function is often used with the MATCH function. We can say it is an alternative way to do VLOOKUP.

Dataset / table

	A	B	C	D	E	F
1	MonthlySalary	CreditScore	CustomerID	LastName	State	Gender
2	90212	698	15565701	FERRI	Johor	Female
3	83256	612	15565706	RAHMAN	Terengganu	Male
4	96518	601	15565714	ADI	Perlis	Male
5	188258	627	15565779	KENT	Kedah	Female
6	74511	745	15565796	HAY	Terengganu	Male
7	30584	532	15565806	AMMAR	Terengganu	Male
8	197963	631	15565878	ALEX	Pahang	Male
9	56186	845	15565879	RILEY	Kedah	Female
10	56214	709	15565891	ADAM	Johor	Male
11	154640	653	15565996	ALAN	Terengganu	Male

We can get the same result by performing the 2 functions below.

17	CustomerID	LastName	Age		
18	15565996	ALAN	=VLOOKUP(A18,C1:H11,6,0)		
19	15565879	RILEY			
20	15565714	ADI			
21	15565796	HAY			

VLOOKUP

17	CustomerID	LastName	Age		
18	15565996	ALAN	=INDEX(\$H\$2:\$H\$11,MATCH(A18,\$C\$2:\$C\$11,0))		
19	15565879	RILEY			
20	15565714	ADI			
21	15565796	HAY			

INDEX & MATCH

# INDEX & Match functions



Before we can perform INDEX MATCH function, let us first learn the INDEX and MATCH function separation.

INDEX can be used to get **value** at known position.

=INDEX(array, row\_num, [col\_num])

MATCH can be used to find **position** in a list.

=MATCH(lookup\_value, lookup\_array, [match\_type])

Dataset / table:

	A	B	C	D	E	F	G
1	CustomerID	LastName	State	Gender	MaritalStatus	Age	NumberofDependents
2	15565701	FERRI	Johor	Female	Yes	39	0
3	15565706	RAHMAN	Terengganu	Male	Yes	35	0
4	15565714	ADI	Perlis	Male	Yes	47	1
5	15565779	KENT	Kedah	Female	Yes	30	4
6	15565796	HAY	Terengganu	Male	Yes	48	0
7	15565806	AMMAR	Terengganu	Male	No	38	1
8	15565878	ALEX	Pahang	Male	Yes	29	3
9	15565879	RILEY	Kedah	Female	No	28	1
10	15565891	ADAM	Johor	Male	No	39	0
11	15565996	ALAN	Terengganu	Male	Yes	44	3

Formulas:

16	Position	CustomerID	LastName
17	4	=INDEX(A2:A11,A17)	=INDEX(B2:B11,A17)
18	7	=INDEX(A2:A11,A18)	=INDEX(B2:B11,A18)

Formulas:

16	CustomerID	Position
17	15565779	=MATCH(A17,A2:A11,0)
18	15565878	=MATCH(A18,A2:A11,0)

Output:

16	Position	CustomerID	LastName
17	4	15565779	KENT
18	7	15565878	ALEX

Output:

16	CustomerID	Position
17	15565779	4
18	15565878	7

# Try it now



Go to Exercise worksheet, perform INDEX or MATCH functions from Calculation worksheet.

## 1. Find the value

- 3<sup>rd</sup> position at LastName column
  - insert a new worksheet and rename it as “Exercise”, create the table shown in the picture
  - in cell C2, type =INDEX(Calculation!B:B,A2) and hit enter button
- 700<sup>th</sup> position at Age column
- 1000<sup>th</sup> position at CustomerID column
- 5000<sup>th</sup> position at LastName

	A	B	C
1	Position	Column Name	Value
2	3	LastName	
3	700	Age	
4	1000	CustomerID	
5	5000	LastName	
6			
7	Header	Value	Position
8	CustomerID	15571571	
9	CustomerID	15574213	
10	LastName	HANNAH	
11	CustomerID	15568104	

## 2. Find the position

- 15571571 from CustomerID column
  - in cell C7, type =MATCH(B7,Calculation!\$A\$2:\$A\$10048,0), and hit enter button
- 15574213 from CustomerID column
- HANNAH from LastName column (\*\* there is a double spacing before HANNAH)
- 15568104 from Customer ID column

# Dynamic formulas – INDEX MATCH



The INDEX function is often used with the MATCH function. We can say it is an alternative way to do VLOOKUP.

=INDEX(array, row\_num, [col\_num])

↑ replace

=MATCH(lookup\_value, lookup\_array, [match\_type])

Dataset / table

	A	B	C	D	E	F	G	H
1	MonthlySala	CustomerID	LastName	State	Gender	MaritalStatus	Age	NumberofDependents
2	90212	15565701	FERRI	Johor	Female	Yes	39	0
3	83256	15565706	RAHMAN	Terengganu	Male	Yes	35	0
4	96518	15565714	ADI	Perlis	Male	Yes	47	1
5	188258	15565779	KENT	Kedah	Female	Yes	30	4
6	74511	15565796	HAY	Terengganu	Male	Yes	48	0
7	30584	15565806	AMMAR	Terengganu	Male	No	38	1
8	197963	15565878	ALEX	Pahang	Male	Yes	29	3
9	56186	15565879	RILEY	Kedah	Female	No	28	1
10	56214	15565891	ADAM	Johor	Male	No	39	0
11	154640	15565996	ALAN	Terengganu	Male	Yes	44	3

The value we want to lookup.

	CustomerID	MonthlySalary
17		
18	15565996	=INDEX(A2:A11,MATCH(A18,B2:B11,0))
19	15565779	

# Try it now



Go to **Calculation** worksheet, lookup data from **Financial Info** worksheet using INDEX & MATCH.

1. NumberBankAccts (add a new column, column name: NumberBankAccts\_1)
2. Savings (add a new column, column name: Savings\_1)
3. LastTransactionDate (add a new column, column name: LastTransactionDate\_1)
4. LastTransactionAmt (add a new column, column name: LastTransactionAmt\_1)
5. MonthlySalary (add a new column, column name: MonthlySalary)
6. CreditScore (add a new column, column name: CreditScore)

Hint:  
Use col\_num in this syntax.

=INDEX(array, row\_num, [col\_num])

↑  
replace

=MATCH(lookup\_value, lookup\_array, [match\_type])

# Try it now



Go to **Exercise** worksheet, lookup data from **Calculation** worksheet.

How can you use INDEX MATCH function to find the value for both row and column information?

a. Create the table shown in the picture in range A11:D14

b. In cell B12, type

```
=INDEX(Calculation!$A:$P, MATCH(Exercise!A12, Calculation!$A:$A, 0),MATCH(Exercise!B11,  
Calculation!$1:$1, 0))
```

c. The result in B12 is “TING”, apply the same formula to fill in the values in the table.

CustomerID	LastName	State	LastTransactionAmt
15571571	TING		
15574213			
15576196			

Hint:

Use col\_num in this syntax.

```
=INDEX(array, [row_num], [col_num])  
=MATCH(lookup_value, lookup_array, [match_type])  
=MATCH(lookup_value, lookup_array, [match_type])
```

replace  
replace

# Contents

- Overview of training datasets
- Managing multiple worksheets
- Data understanding
- Conditional formatting and logical operators
- Lookup and reference functions
- A review of essential functions**
- Fundamentals of descriptive statistics
- Selecting the right charts to visualize your analysis result
- Dashboard dos and don'ts



NATURAL INTELLIGENCE SOLUTIONS TECHNOLOGY

# A review of essential functions



**Logical functions** can be used to:

1. Check if a condition is true or false
2. Combine multiple conditions together

For example, we want to validate the customer data with several criteria such as (1) married and (2) female customers, then we can use logical functions to tell us any records that have fulfilled the criteria.

Function	Description	Syntax
AND	Test multiple conditions with AND.	=AND(logical1, [logical2], ...)
OR	Test multiple conditions with OR.	=OR(logical1, [logical2], ...)
IF	Test for a specific condition.	=IF(logical_test, value_if_true, value_if_false)
IFERROR	Returns a specified alternate value if a formula results in an #DIV/0! error.	=IFERROR(value, value_if_error)
IFNA	Returns a specified alternate value if a formula results in an #N/A error.	=IFNA(value, value_if_na)
IFS	Test multiple conditions, return first true.	=IFS(logical_test1, value1, [logical_test2, value2], ...)

# Try it now



In **Calculation** Worksheet, apply logical functions to check for following criteria

- A. Married female with aged 40 (add a new column, column name: A)
  - a. create a new column at column Q
  - b. in cell Q2, type =AND(D2="Female",E2="Yes",F2=40), and hit enter button. In this formula we are using logical operators to perform the comparison in AND function.
- B. Male from Selangor and has more than 3 dependents (add a new column, column name: B)
- C. Age below 40 or has more than 3 dependents (add a new column, column name: C)
- D. Male come from either Selangor or Kuala Lumpur (add a new column, column name: D)
- E. If CreditScore greater than 600, then mark as Y, otherwise as N (add a new column, column name: E)
- F. If CreditScore greater than 600 and MonthlySalary greater than 50,000, then mark as PASS, if CreditScore less than 400 and MonthlySalary more than and equal to 100,000, then mark as FAIL, otherwise mark as CHECK (add a new column, column name: F)
- G. If output in A is NOT True, then lookup Churn value from Financial Info dataset, otherwise leave it blank (add a new column, column name: G)

# A review of essential functions



**Text function** can be used to convert a number to text in a number format.

When we consolidate the data from different sources, and the date for each rows is input with different format, now we can use text function to standardise the date format into a specific and standard format.

Syntax	Input	Examples	Output
TEXT(value, format_text)	8 March 2015	=TEXT(B2, "mm/dd/yyyy")	03/08/2015
		=TEXT(B2, "dd/mm/yyyy")	08/032015
		=TEXT(B2, "dd-mm-yyyy")	08-03-2015
		=TEXT(B2, "dddd, mmmm d, yyyy")	Sunday, March 8, 2015
	5.5	=TEXT(B6, "0.00")	5.50
		=TEXT(B6, "#%")	550%
		=TEXT(B6, "\$#,##0.00")	\$5.50
		=TEXT(B6, "+ \$#,##0.00;- \$#,##0.00;\$0.00")	+ \$5.50
		=TEXT(B6, "- \$#,##0.00;- \$#,##0.00;\$0.00")	- \$5.50
		=TEXT(B6, "# ?/?")	5 1/2
		=TEXT(B6, "0.00E+00")	5.50E+00
		=TEXT(B6, "~# !")	~6 !

# A review of essential functions



## Date and time functions

This function is useful when we want to extract a specific part of a date. For example, instead of looking at transactions amount at every single date, your boss would like to look at a high level summary, which is at year level or month level. Then you can use date and time function to extract the specific date part and summarise the report.

Function	Description	Syntax
DATE	Create a valid date from year, month, and day	=DATE(year,month,day)
YEAR	Get the year from a date	=YEAR(serial_number)
MONTH	Get the month as a number (1-12) from a date	=MONTH(serial_number)
DAY	Get the day as a number (1-31) from a date	=DAY(Serial_number)
DAYS	Get days between dates	=DAYS(end_date, start_date)
TODAY	Get the current date	=TODAY()
NOW	Get the current date and time	=NOW()

# Try it now



In **Calculation Worksheet**,

1. Standardize the date format in column LastTransactionDate to “dd-mm-yy” format (add a new column, column name: dd-mm-yy)  
(Hint: use =TEXT function)
2. Create a new column to store YEAR only from LastTransactionDate (add a new column, column name: Year)
3. Create a new column to store MONTH only from LastTransactionDate (add a new column, column name: Month)
4. Create a new column to store DAY only from LastTransactionDate (add a new column, column name: Day)
5. Convert the value in column LastTransactionAmt to “\$#,##0.00” (add a new column, column name: AmountInDollar)
6. Get the days difference from today date and LastTransactionDate (add a new column, column name: Aging)

# A review of essential functions



Functions that can be used to manipulate text data. All these functions can be used to transform and standardise the data. For example, the text length of “Female” in gender column appear to be too long in data table, instead of showing the whole text, you can just show “F” to represent “Female” by using LEFT function to extract the first alphabet from the text.

Function	Description	Syntax
CONCAT	Join text values without delimiter	=CONCAT(text1, [text2], ...)
LEFT	Extract text from the left of a string	=LEFT(text, [num_chars])
MID	Extract text from inside a string	=MID(text, start_num, num_chars)
RIGHT	Extract text from the right of a string	=RIGHT(text, [num_chars])
TRIM	Remove extra spaces from text	=TRIM(text)
LEN	Get the length of text	=LEN(text)
LOWER	Convert text to lower case	=LOWER(text)
UPPER	Convert text to upper case	=UPPER(text)

# Try it now



In **Calculation Worksheet**,

1. Trim the extra space in LastName column (add a new column, column name: Name\_TrimSpace)
2. Get the length of each text in LastName column after trim the extra space (add a new column, column name: Name\_Length)
3. Get the first alphabet for all LastName (add a new column, column name: Name\_FirstAlphabet)
4. Get the second alphabet onwards for all LastName and lowercase them (add a new column, column name: Name\_Lower)
5. Concatenate output from question 3 & 4 (add a new column, column name: Name\_Concat)
6. Perform one formula to combine all the criteria from question 1 ~ 5 (add a new column, column name: Name\_Final)  
(\* the answer will be exactly same as the answer in question 5.)
7. Capitalise the text in column State (add a new column, column name: State\_Cap)
8. Get the first alphabet for Gender (add a new column, column name: Gender\_Short)
9. Create a new column to read the LastTransactionDate with mm-yy format (add a new column, column name: mm-yy)

# Contents

- Overview of training datasets
- Managing multiple worksheets
- Data understanding
- Conditional formatting and logical operators
- Lookup and reference functions
- A review of essential functions
- Fundamentals of descriptive statistics**
- Selecting the right charts to visualize your analysis result
- Dashboard dos and don'ts



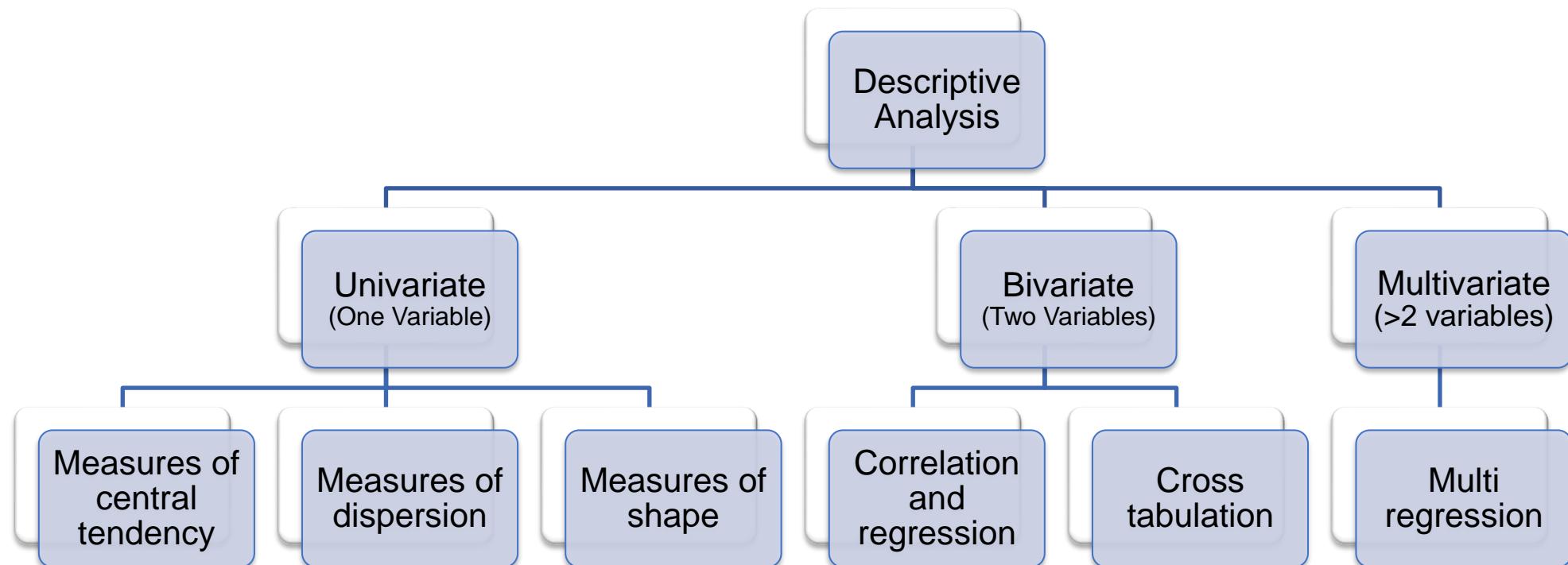
NATURAL INTELLIGENCE SOLUTIONS TECHNOLOGY



# Statistical analysis

Statistical analysis is a component of data analytics. Statistics is basically a science that involves data collection, data interpretation and finally, data validation. Statistical data analysis is a procedure of performing various statistical operations.

Statistical analysis is useful to explore and summarise the information from dataset.



# Univariate analysis



Univariate analysis is the simplest form of analysing data.

- Examination of the distribution of cases on **only one variable** at a time
- Because Univariate analysis does not involve relationships between two or more variables, its purpose is more toward descriptive rather than explanatory.
- Describe patterns found in univariate data include:
  1. Central tendency: mean, mode and median
  2. Dispersion: range, variance, maximum, minimum, quartiles (including the interquartile range), and standard deviation
  3. Shape: skewness, kurtosis



# Distribution of data



The distribution of a statistical data set is a listing or function showing all the possible values of the data and how often they occur.

- When a distribution of categorical data is organized, you see the number or percentage of individuals in each group.
- When a distribution of numerical data is organized, they're often ordered from smallest to largest, broken into reasonably sized groups (if appropriate), and then put into graphs and charts to examine the shape, center, and amount of variability in the data.
- One of the most well-known distributions is called the normal distribution, also known as the bell-shaped curve.

# Normal distribution

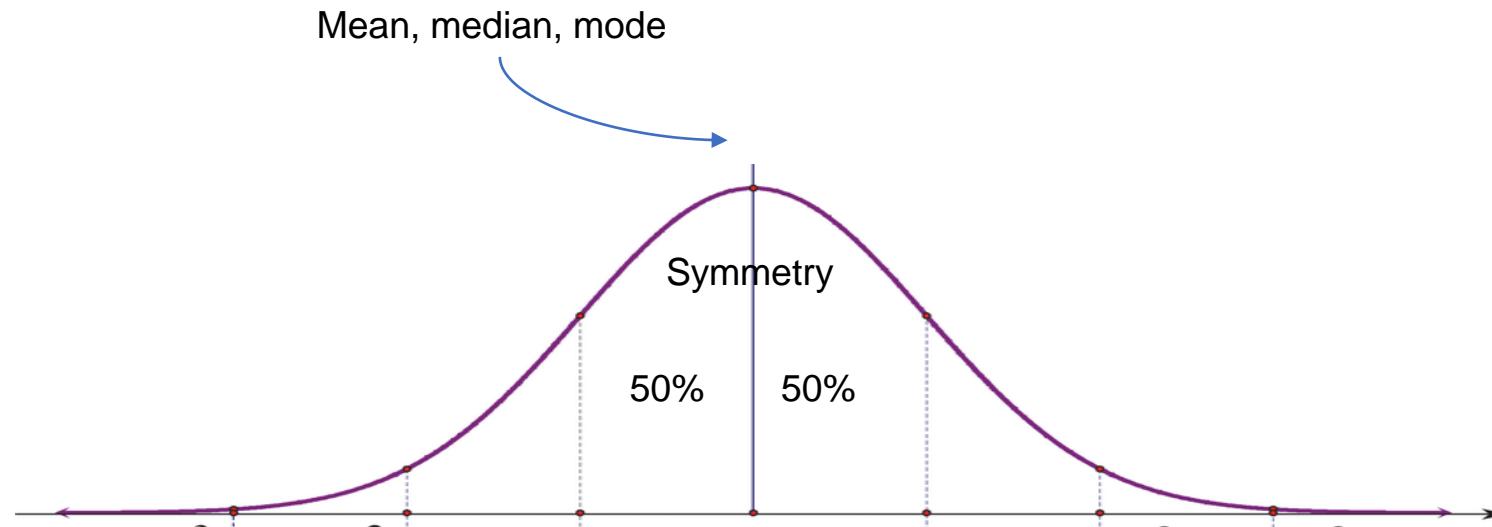


Many things closely follow a Normal Distribution:

- height of people
- size of things produced by machines
- errors in measurements
- blood pressure
- marks on a test

We say the data is “normally distributed”:

- mean = median = mode
- symmetry about the center
- 50% of values less than the mean and 50% greater than the mean

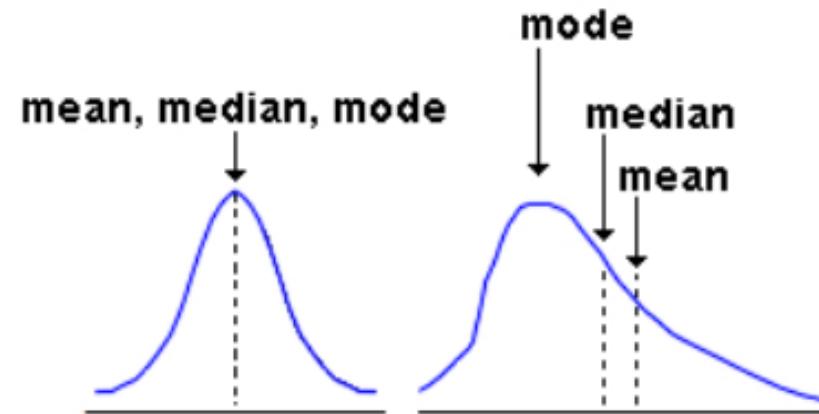


# Frequency distributions: measures of central tendency



Measures of central tendency reflect the central tendencies of a distribution:

- **Mode** reflects the attribute with the greatest frequency
- **Median** reflects the attribute that cuts the distribution in half
- **Mean** reflects the average; sum of attributes divided by number of cases

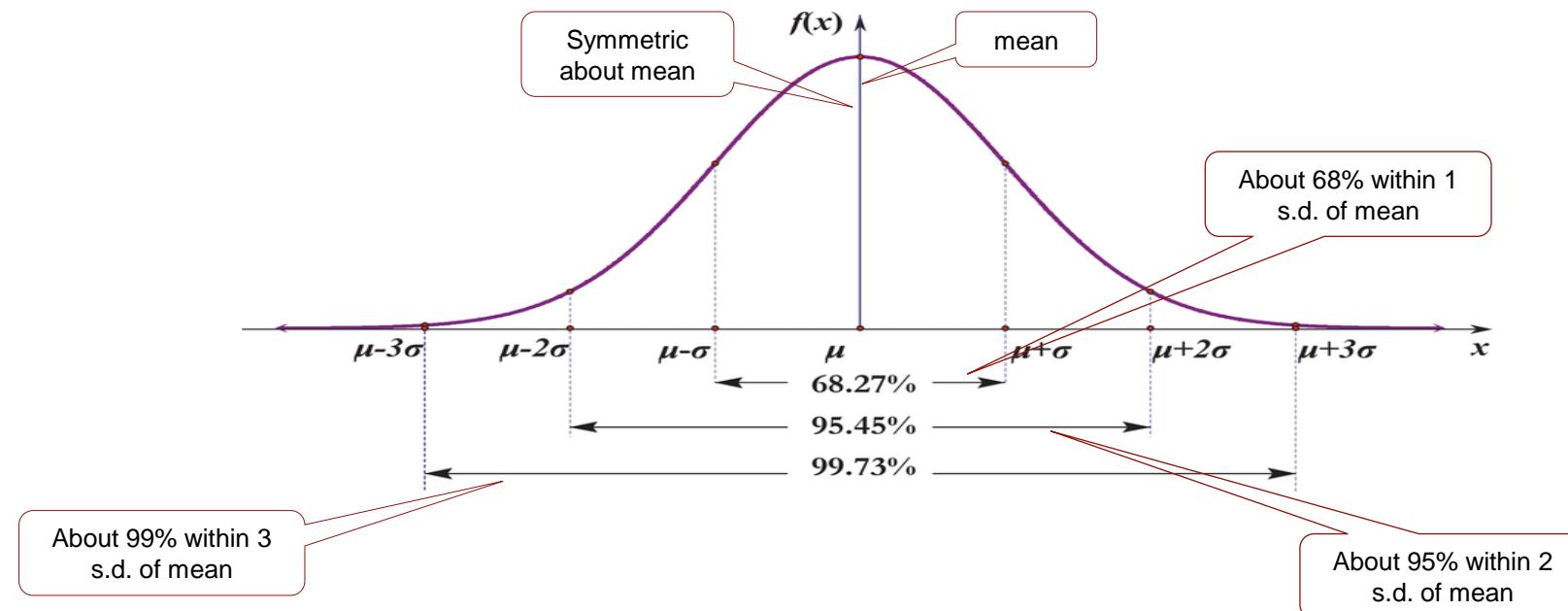


# Frequency distributions: measures of dispersion



Measures of dispersion reflect the spread or distribution of the distribution:

- **Range** is the difference between largest & smallest scores; high – low
- **Variance** is the average of the squared differences between each observation and the mean
- **Standard deviation** is the square root of variance



# Frequency distributions: measures of shape

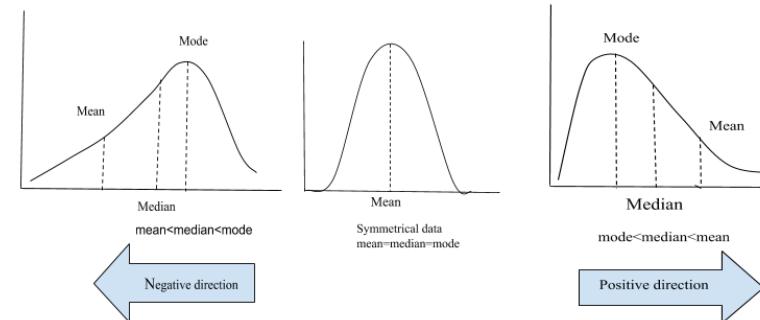


Measures of shape describe the distribution (or pattern) of the data within a dataset :

## 1. Skewness

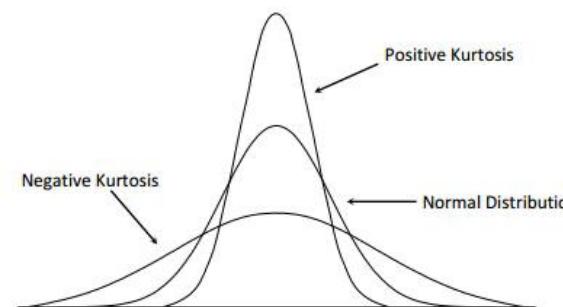
The mean, median and mode are all measures of the center of a set of data. The skewness of the data can be determined by how these quantities are related to one another

By studying the shape of the data we can discover the relation between the mean, median and mode



## 2. Kurtosis

The measurement of a function's "tailedness", often visually described by the sharpness of the peak values. Kurtosis is often explained in terms of the central peak. Higher values of it, indicate a higher, sharper peak; lower values indicate a lower, less distinct peak



# A review of essential functions



## Maths & Statistical Functions (1)

For example, you can use this function to aggregate the transaction amount to prepare a monthly report.

Function	Description	Syntax
SUM	Add numbers together.	=SUM(number1, [number2], [number3].....)
SUMIF	Sum numbers in a range that meet supplied criteria.	=SUMIF(range, criteria, [sum_range])
SUMIFS	Sum cells that match multiple criteria.	=SUMIFS( sum_range, criteria_range1, criteria1, [criteria_range2, criteria2], ...)

# A review of essential functions



## Maths & Statistical Functions (2)

For example, by calculating average monthly transaction amount made by each age group customer, this will allow us to understand the average spending behaviour of customer by age group.

Function	Description	Syntax
AVERAGE	Get the average of a group of numbers.	=AVERAGE(number1, [number2], ...)
AVERAGEA	Get the average of a group of numbers and text.	=AVERAGEA(value1, [value2], ...)
AVERAGEIF	Get the average of numbers that meet criteria.	=AVERAGEIF(range, criteria, [average_range])
AVERAGEIFS	Average cells that match multiple criteria.	=AVERAGEIFS(average_range, criteria_range1, criteria1, [criteria_range2, criteria2], ...)

Averages are useful because they:

- summarise a large amount of data into a single value
- indicate that there is some variability around this single value within the original data.

# A review of essential functions



## Maths & Statistical Functions (3)

For example, we can use this function to count the number of customers either at grand total level or drill down to a specific criteria for reporting.

Function	Description	Syntax
COUNT	Count numbers.	=COUNT(value1, value2....)
COUNTA	Count the number of non-blank cells.	=COUNTA(value1, [value2], ...)
COUNTBLANK	Count cells that are blank.	=COUNTBLANK(range)
COUNTIF	Count cells that match criteria.	=COUNTIF(range, criteria)
COUNTIFS	Count cells that match multiple criteria.	=COUNTIFS(criteria_range1, criteria1, [criteria_range2, criteria2]...)

# A review of essential functions



## Maths & Statistical Functions (4)

For example we can use this functions to find the oldest customer or highest transaction amount.

Function	Description	Syntax
MAX	Get the largest value, ignores the logical values.	=MAX(number1, [number2], ...)
MAXA	Get the largest value, takes the logical values into account in the search process, TRUE = 1, and FALSE = 0.	=MAXA(value1, [value2], ...)
MAXIFS	Get maximum value with criteria.	=MAXIFS(max_range, criteria_range1, criteria1, [criteria_range2, criteria2], ...)
MIN	Get the smallest value.	=MIN(number1, [number2], ...)
MINA	Get the smallest value.	=MINA(value1, [value2], ...)
MINIFS	Get minimum value with criteria.	=MINIFS(max_range, criteria_range1, criteria1, [criteria_range2, criteria2], ...)

# A review of essential functions



## Maths & Statistical Functions (5)

For example, to understand the distribution of the age of customers, by summarising the median, mode, standard deviation of the age, we can get the information of the age of majority customers etc.

Function	Description	Syntax
MEDIAN	Calculate the middle value of a given set of numbers.	=MEDIAN(number1, [number2], ...)
MODE	Calculate the most frequently occurring value from a list of supplied numbers. If there are two or more most frequently occurring values in the supplied data, the function returns the lowest of the values.	=MODE(number1, [number2], ...)
RANK	Rank a number against a range of numbers.	=RANK(number, ref, [order])
STDEV	Get the standard deviation in a sample.	=STDEV(number1, [number2],...)
VAR	Get the sample variance of a set of values.	=VAR(number1, [number2],...)

# A review of essential functions



## Maths & Statistical Functions (6)

Function	Description	Syntax
QUARTILE	Calculate the specified quartile of a set of supplied numbers.	=QUARTILE(array, quart)
SKEW	Calculate the skewness of the distribution of a supplied set of values.	=SKEW(number1, [number2], ... )
KURT	Calculate the kurtosis of a supplied set of values.	=KURT(number1, [number2], ... )



# Example of univariate analysis

Taking the results of a pre-training assessment as an example.  
We can describe the results of the participants / data as below:

	A	B
1	No	Score %
2	1	34.38
3	2	50.00
4	3	22.92
5	4	43.75
6	5	35.42
7	6	33.33
8	7	27.08
9	8	31.25
10	9	36.46
11	10	35.42
12	11	45.83
13	12	28.13
14	13	44.79
15	14	38.54
16	15	60.42
17	16	57.29
18	17	46.88
19	18	48.96
20	19	39.58
21	20	56.25
22	21	43.75
23	22	29.17
24	23	46.88
25	24	15.63
26	25	35.42
27	26	38.54
28	27	54.17
29	28	52.08
30	29	62.50
31	30	70.83
32	31	66.67
33	32	50.00
34	33	58.33
35	34	67.71
36	35	70.83
37	36	60.42
38	37	77.08
39	38	69.79

## Measure of central tendency

mean	47.012
mode	35.417
median	46.354

## Measure of dispersion

min	15.625
max	77.083
range	61.458
variance	226.056
standard deviation	15.035
1st quartile	35.417
2nd quartile	46.354
3rd quartile	58.073

## Measure of shape

skewness	0.114
kurtosis	(0.696)

# Try it now



Perform the descriptive analysis for the customers personal info and financial info in **Calculation** worksheet. Explore the data by describing the variables using

- a. Measures of central tendency
  - b. Measures of dispersion
  - c. Measures of shape
- 
- 1. Age
  - 2. Savings
  - 3. Monthly Salary

## Checklist:

- 1. mean
- 2. mode
- 3. median
- 4. range
- 5. variance
- 6. maximum value
- 7. minimum value
- 8. 1<sup>st</sup> quartile
- 9. 2<sup>nd</sup> quartile
- 10. 3<sup>rd</sup> quartile
- 11. 4<sup>th</sup> quartile
- 12. standard deviation
- 13. skewness
- 14. kurtosis

# Contents

- Overview of training datasets
- Managing multiple worksheets
- Data understanding
- Conditional formatting and logical operators
- Lookup and reference functions
- A review of essential functions
- Fundamentals of descriptive statistics
- Selecting the right charts to visualize your analysis result**
- Dashboard dos and don'ts



NATURAL INTELLIGENCE SOLUTIONS TECHNOLOGY

# Data visualisation

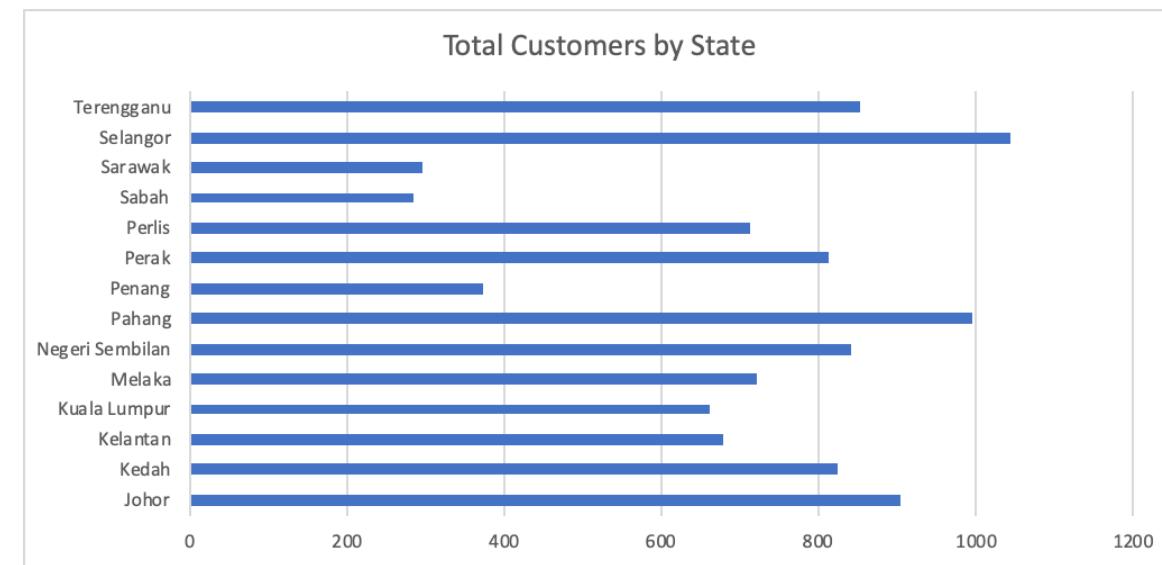


90% of the information transmitted to the brain is visual. Data visualization methods refer to the creation of graphical representations of information.

To prepare the report for your new boss, if the whole report is in table formats, it is hard for you to present the information of customer demographic in a clear and concise way and your boss may need to spend more time and effort to understand those information. It is crucial to present your data and information in a straightforward and clear way, that is to visualise it.

Which one of the below is easier to understand and interpret?

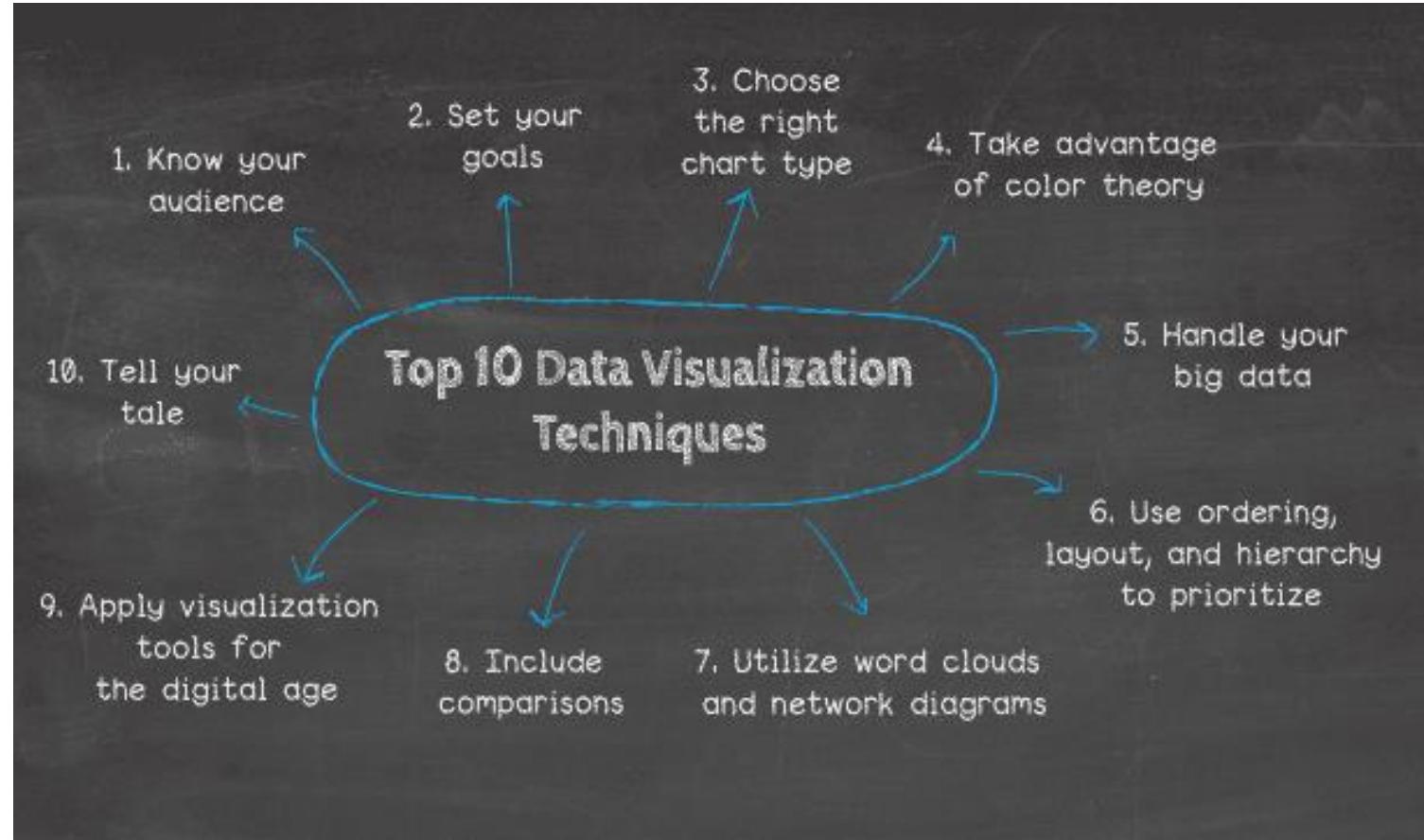
Total customers by state	
State	Total Customers
Johor	904
Kedah	824
Kelantan	679
Kuala Lumpur	662
Melaka	722
Negeri Sembilan	841
Pahang	994
Penang	374
Perak	813
Perlis	712
Sabah	284
Sarawak	297
Selangor	1042
Terengganu	852



# Top 10 data visualisation techniques



Visualization plays an important part of data analytics because it turns a complicated datasets into visuals and make them easier to understand.



<https://www.datapine.com/blog/data-visualization-techniques-concepts-and-methods/>

# Top 10 data visualisation techniques



Ten essential data visualization techniques you should know.

Data visualisation techniques	Description
1. Know the audience	Ensuring the content is both inspiring and tailored to the audience is one of the most essential data visualization techniques.
2. Set the goals	To set a clear-cut set of aims, objectives, and goals prior to building the management reports, graphs, and additional visuals.
3. Choose the right chart type	Must select the right charts for your specific project, audience, and purpose.
4. Take advantage of colour theory	Selecting the right colour scheme for the presentational assets will help enhance efforts significantly. You should always try to keep your colour scheme consistent throughout your data visualizations, using clear contrasts to distinguish between elements (e.g. positive trends in green and negative trends in red). As a guide, people, on the whole, use red, green, blue, and yellow as they can be recognized and deciphered with ease.
5. Handle your big data	Identify data availability, decide which is the most valuable, and label each branch of information clearly to make it easy to separate, analyse, and decipher.

# Top 10 data visualisation techniques



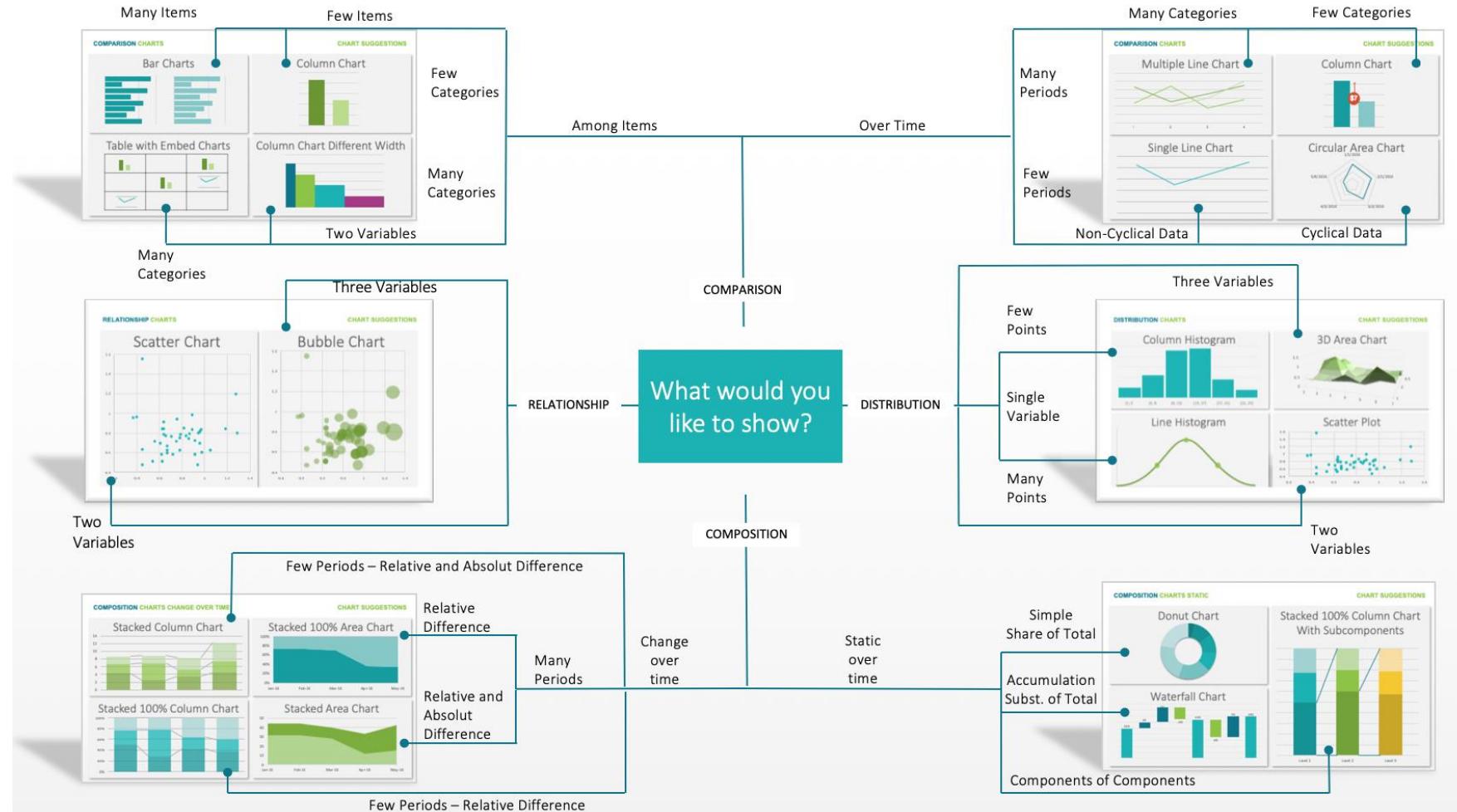
Ten essential data visualization techniques you should know.

Data visualisation techniques	Description
6. Use ordering, layout, and hierarchy to prioritize	Utilise ordering, layout, and hierarchy to make the presentation efforts speedier, simpler, and more successful. Categorized your data and broken it down to the branches of information that you think to be most valuable to your organization.
7. Utilize word clouds and network diagrams	To handle semi-structured or decidedly unstructured sets of data efficiently, the services of network diagrams or cloud words are recommended. A word cloud is an image developed with words used for particular text or subject, in which the size of each word indicates its frequency or importance within the context of the information.
8. Include comparisons	Highlighting strengths, weaknesses, trends and peaks that everyone can ponder and act upon.
9. Apply visualisation tools for the digital age	A task-specific, interactive online dashboard or tool offers a digestible, intuitive, comprehensive, and interactive mean of collecting, collating, arranging, and presenting data with ease – ensuring that your techniques have the most possible impact while taking up a minimal amount of your time.
10. Tell your tale	When data is presenting in a visual format with the aim of communicating an important message or goal, storytelling will engage the audience and make it easy for people to understand with minimal effort.



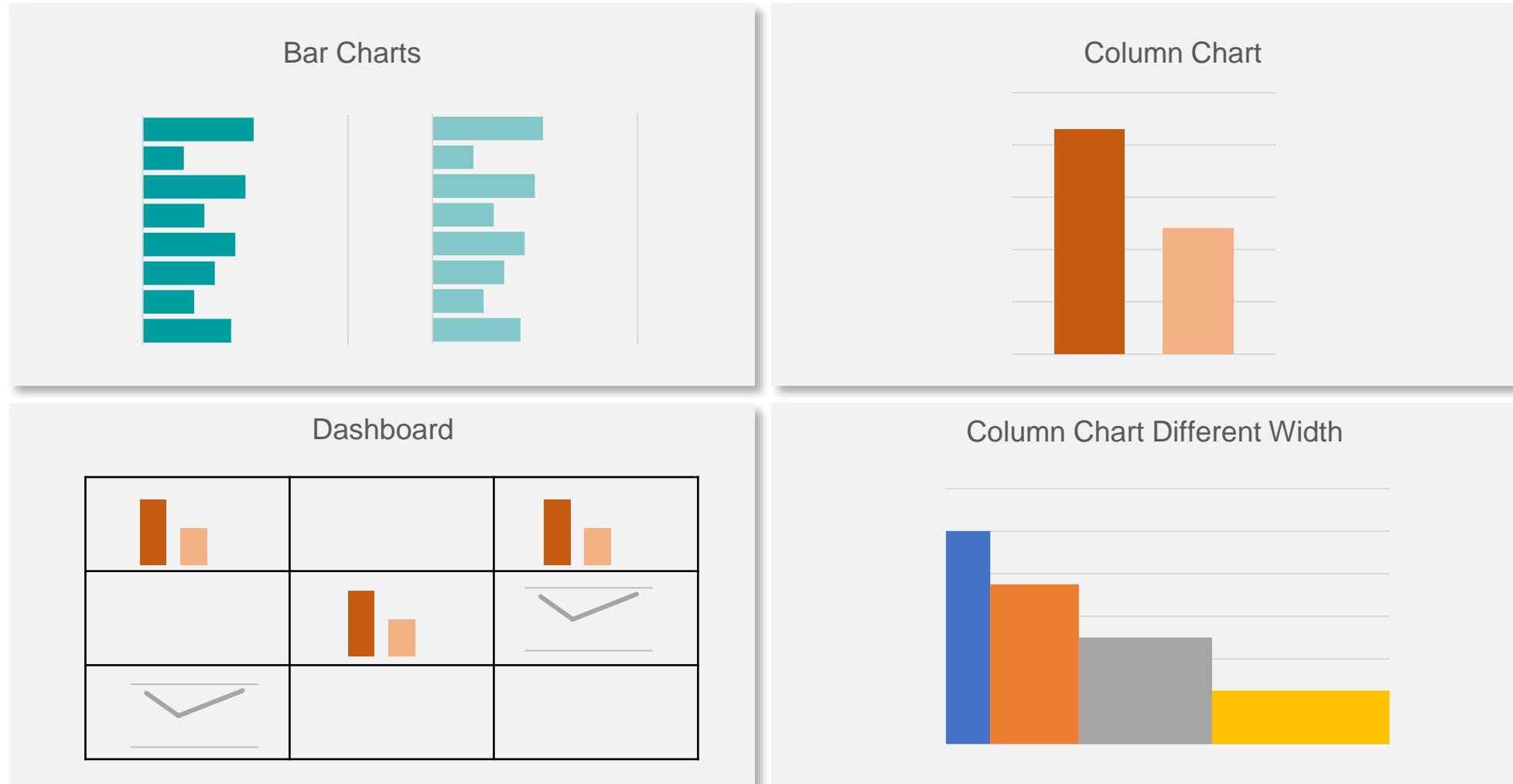
# How to select the right charts for your data

If you want to visualize your data, make sure you use the right charts. While your data might work with multiple chart types, it's up to you to select the one that ensures your message is clear and accurate. Remember, data is only valuable if you know how to visualize it and give context.



- Comparison charts among items

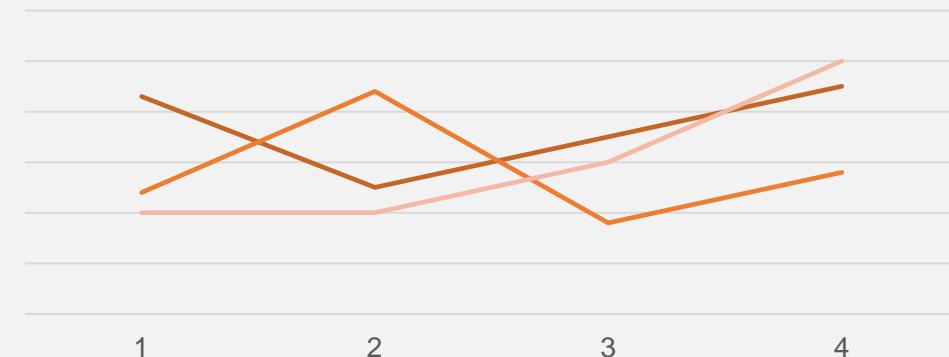
When we want to compare one or more datasets/ variables. For example, compare the number of customers by gender.



- Comparison charts change over time

When we want to compare items or show differences over time. For example, to compare the amount of transactions made by female or male customers in past 5 years.

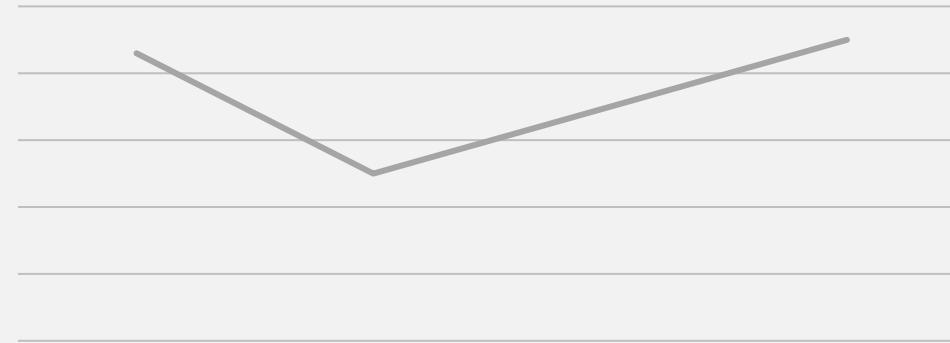
Multiple Line Chart



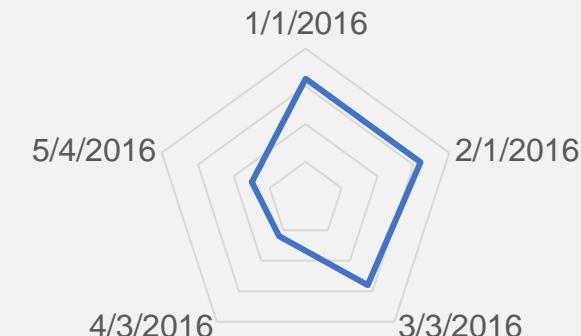
Column Chart



Single Line Chart

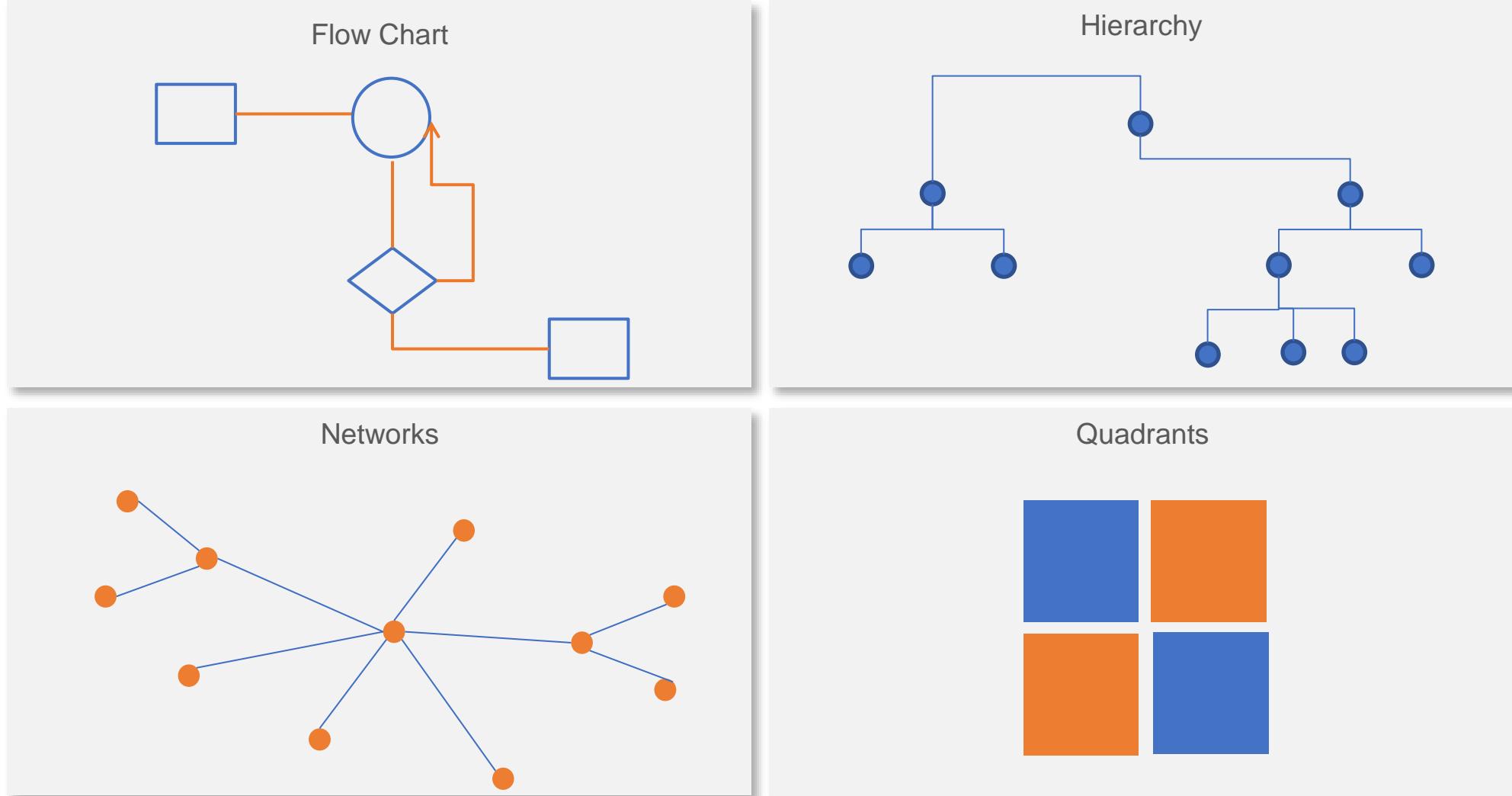


Circular Area Chart



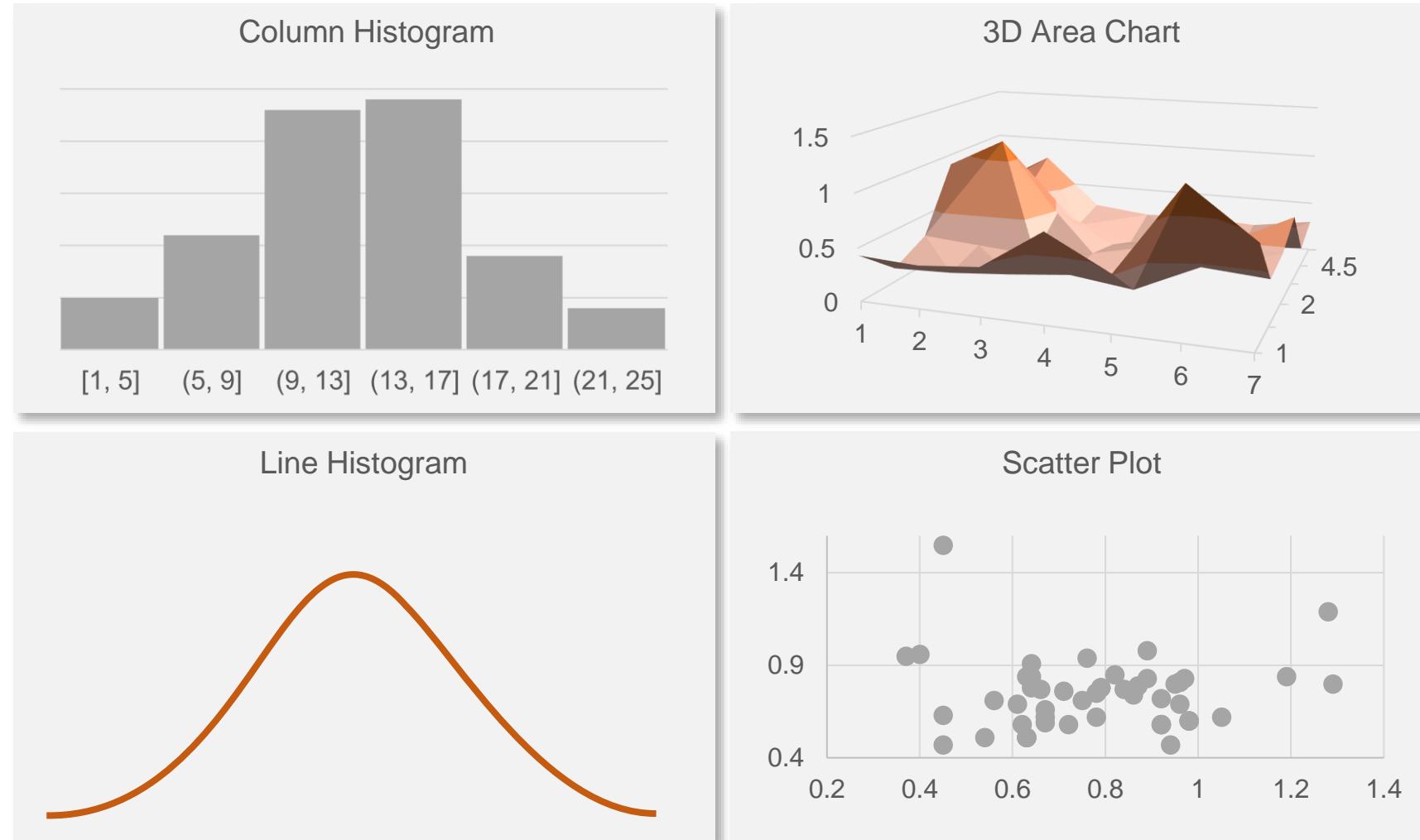
## • Relationship additions

When we want to establish (or show) relationship between 2 or more variables. For example, what is the relationship for female customers who made more number of transactions impact the approval of her loan application.



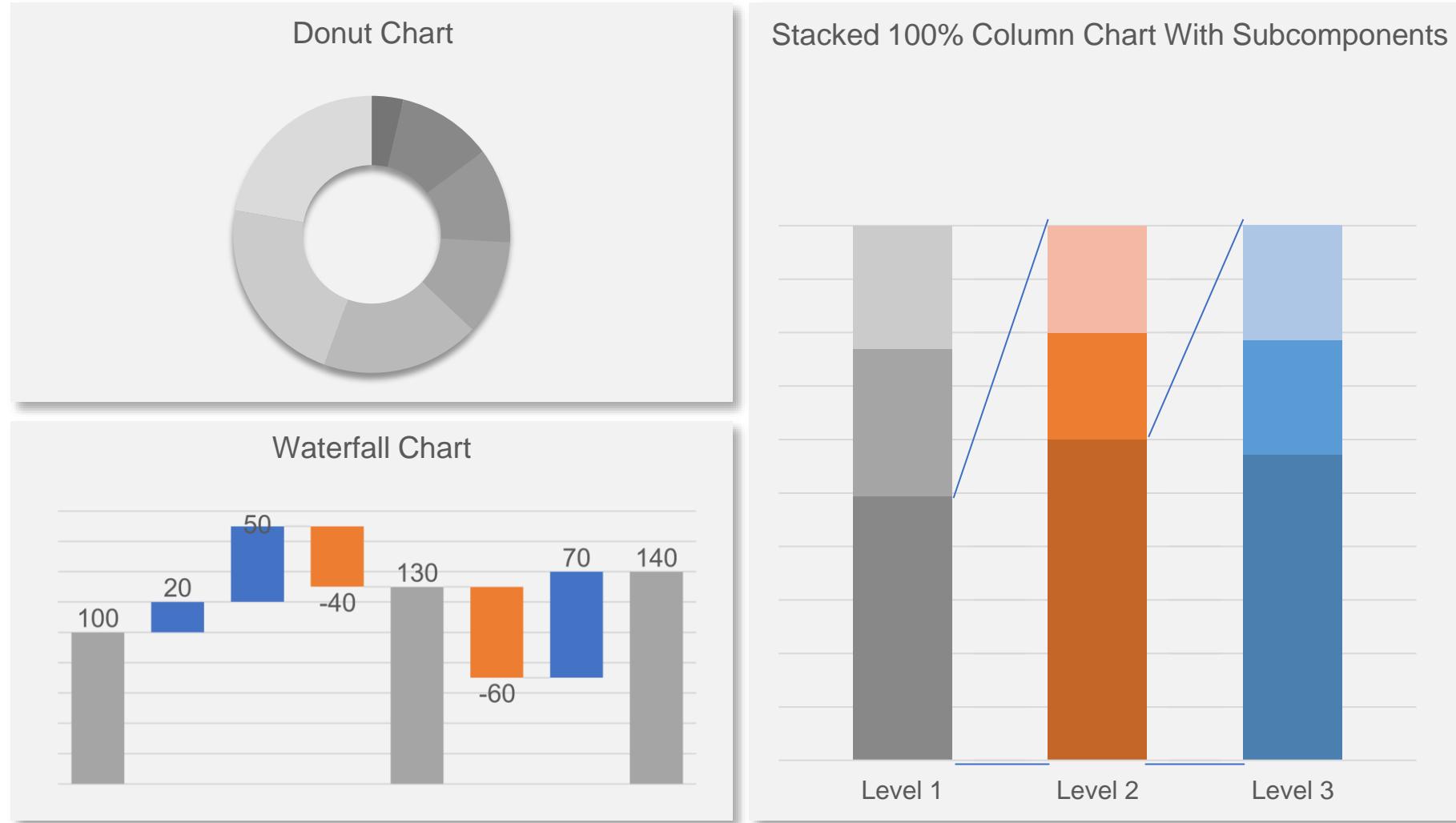
- Distribution charts

When we want to show the distribution of a set of values to understand the outliers, normal ranges etc. For example, what is the distribution/ number of customers according to the age group.



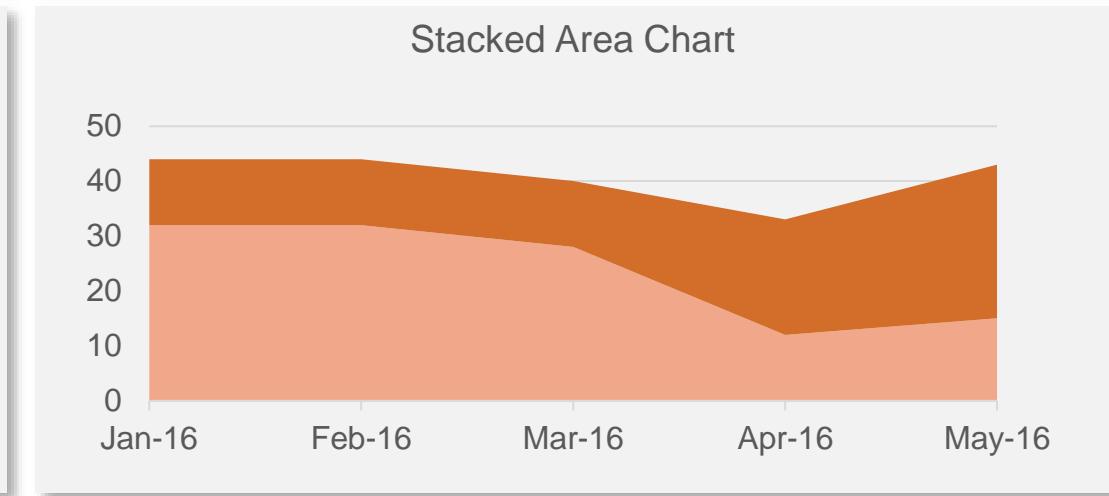
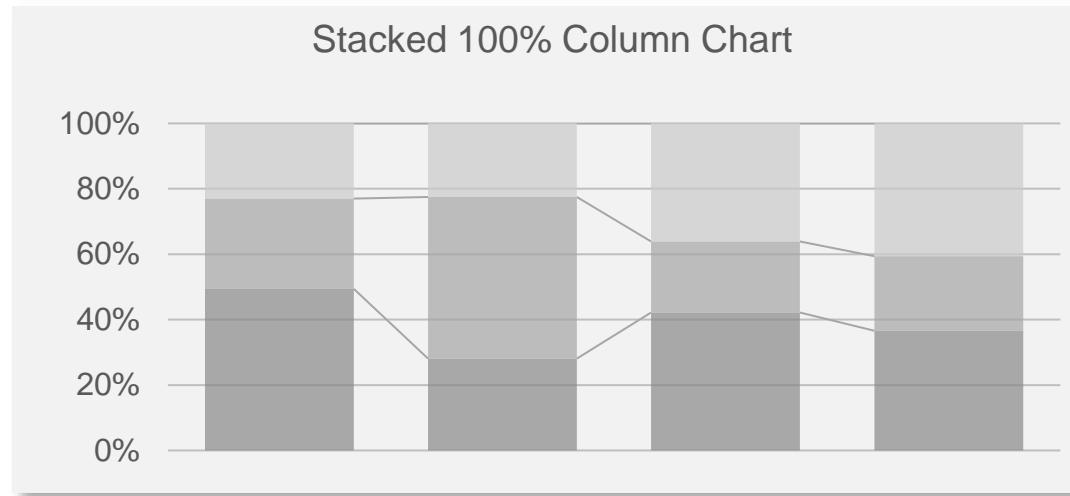
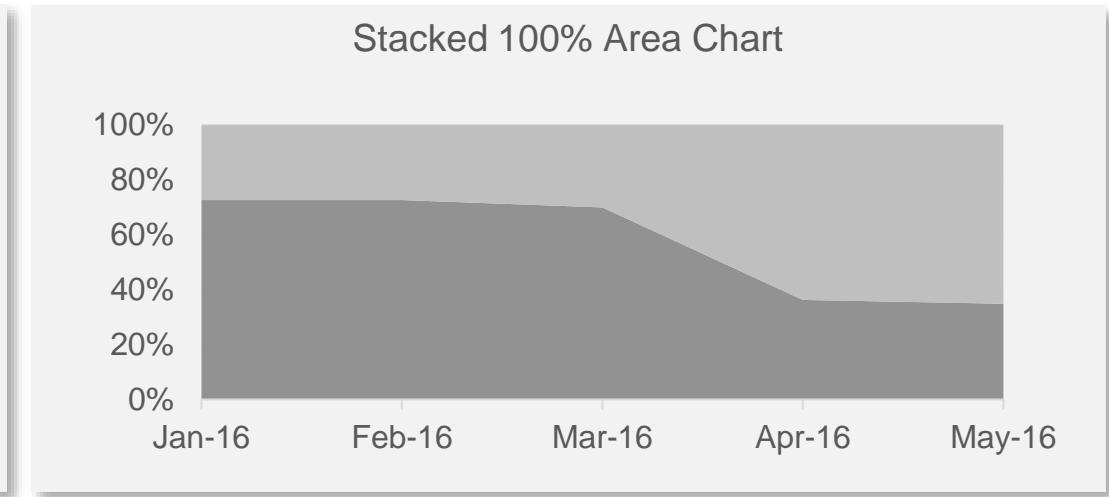
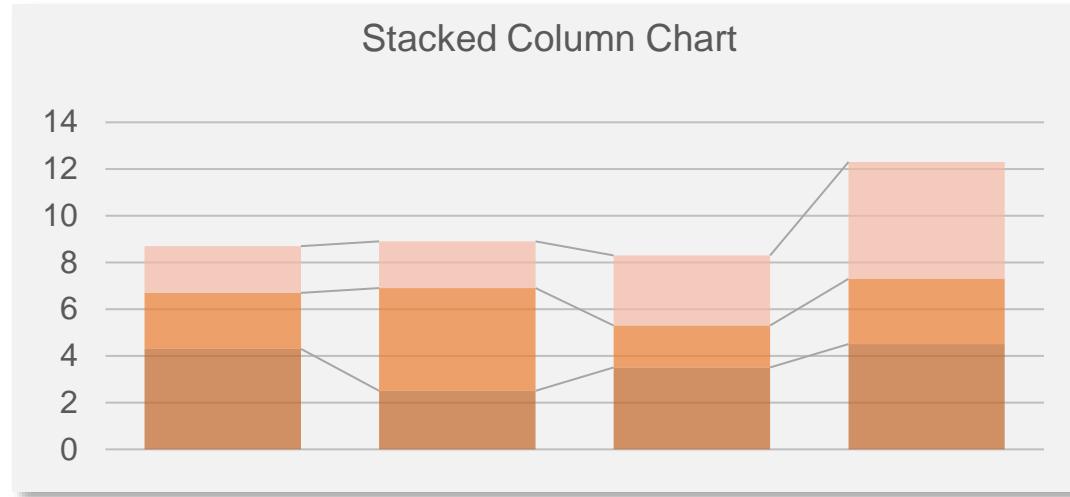
- Composition charts static

When we want to show how various parts comprise the whole. For example, the contribution of the number of customers in each state to total customers.



- Composition charts change over time

When want to understand the trend over time of some variables. For example, the contribution of the number of customers in each state to total customers in past 5 years.



# Data manipulation



Now, you have performed all the necessary data preparation steps on your raw data:

1. Consolidation,
2. Validation,
3. Standardisation,
4. Calculation,
5. Aggregation etc

However you will realised that you have created a lot of new columns during data preparation, and some of the existing columns from raw data or validation columns are no longer required for dashboarding and reporting. So, before you start to visualise the data (CRISP-DM modelling phase), you need to manipulate your data to make it easier to read and create tables and charts for reporting.

The data manipulation steps include:

1. Insert new columns
2. Delete columns
3. Move columns
4. Cut & paste columns
5. Copy & paste columns



# Data manipulation steps

Inserting / Adding Columns – Simply select the column to the right of the column you wish to add, right click, and select “Insert”.

A screenshot of Microsoft Excel showing a context menu open over a table. The menu options are: Cut, Copy, Paste, Paste Special, Insert, Delete, Clear, Selection, Format Cells..., Column Width..., Hide, Unhide, Import Image. The 'Insert' option is highlighted with a red box. The table has columns A through H and rows 1 to 25. The first row contains headers: Emp ID, Name Prefix, Last Name. The second row contains data: 677509, Drs., Walker. The third row contains data: 940761, Ms., Robinson. The fourth row contains data: 93819, Mr., Russell. The fifth row contains data: 99687, Mr., Bailey. The sixth row contains data: 39712, Ms., Baker. The seventh row contains data: 80086, Mr., Murphy. The eighth row contains data: 162402, Hon., Young. The ninth row contains data: 231469, Mr., Peterson. The tenth row contains data: 386158, Mr., Flores. The eleventh row contains data: 301576, Hon., King. The twelfth row contains data: 441771, Mr., Watson. The thirteenth row contains data: 528509, Hon., Scott. The fourteenth row contains data: 912990, Mr., Diaz. The fifteenth row contains data: 214352, Ms., Stewart. The sixteenth row contains data: 890290, Hon., Lee. The seventeenth row contains data: 622406, Mr., Scott. The eighteenth row contains data: 979607, Ms., Lewis. The nineteenth row contains data: 969580, Prof., Edwards. The twentieth row contains data: 426038, Hon., Turner. The twenty-first row contains data: Joan, C, Stewart.

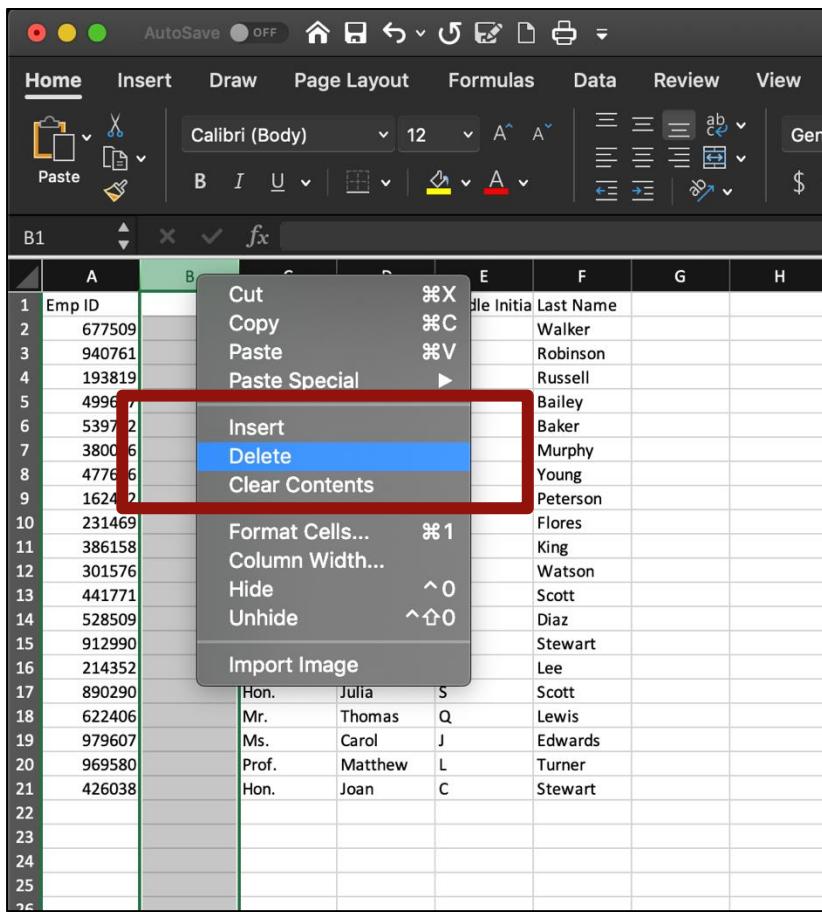


A screenshot of Microsoft Excel showing the same table after a new column 'C' has been inserted between columns B and D. The table now has columns A, B, C, D, E, F, G, H and rows 1 to 25. The first row contains headers: Emp ID, Name Prefix, First Name, Middle Initial, Last Name. The second row contains data: 677509, Drs., Lois, H, Walker. The third row contains data: 940761, Ms., Brenda, S, Robinson. The fourth row contains data: 193819, Mr., Benjamin, R, Russell. The fifth row contains data: 499687, Mr., Patrick, F, Bailey. The sixth row contains data: 539712, Ms., Nancy, T, Baker. The seventh row contains data: 380086, Mrs., Carol, V, Murphy. The eighth row contains data: 477616, Hon., Frances, B, Young. The ninth row contains data: 162402, Hon., Diana, T, Peterson. The tenth row contains data: 231469, Mr., Ralph, L, Flores. The eleventh row contains data: 386158, Mrs., Melissa, Q, King. The twelfth row contains data: 301576, Hon., Wayne, G, Watson. The thirteenth row contains data: 441771, Mrs., Cheryl, O, Scott. The fourteenth row contains data: 528509, Hon., Paula, G, Diaz. The fifteenth row contains data: 912990, Mr., Joshua, T, Stewart. The sixteenth row contains data: 214352, Ms., Theresa, G, Lee. The seventeenth row contains data: 890290, Hon., Julia, S, Scott. The eighteenth row contains data: 622406, Mr., Thomas, Q, Lewis. The nineteenth row contains data: 979607, Ms., Carol, J, Edwards. The twentieth row contains data: 969580, Prof., Matthew, L, Turner. The twenty-first row contains data: 426038, Hon., Joan, C, Stewart.



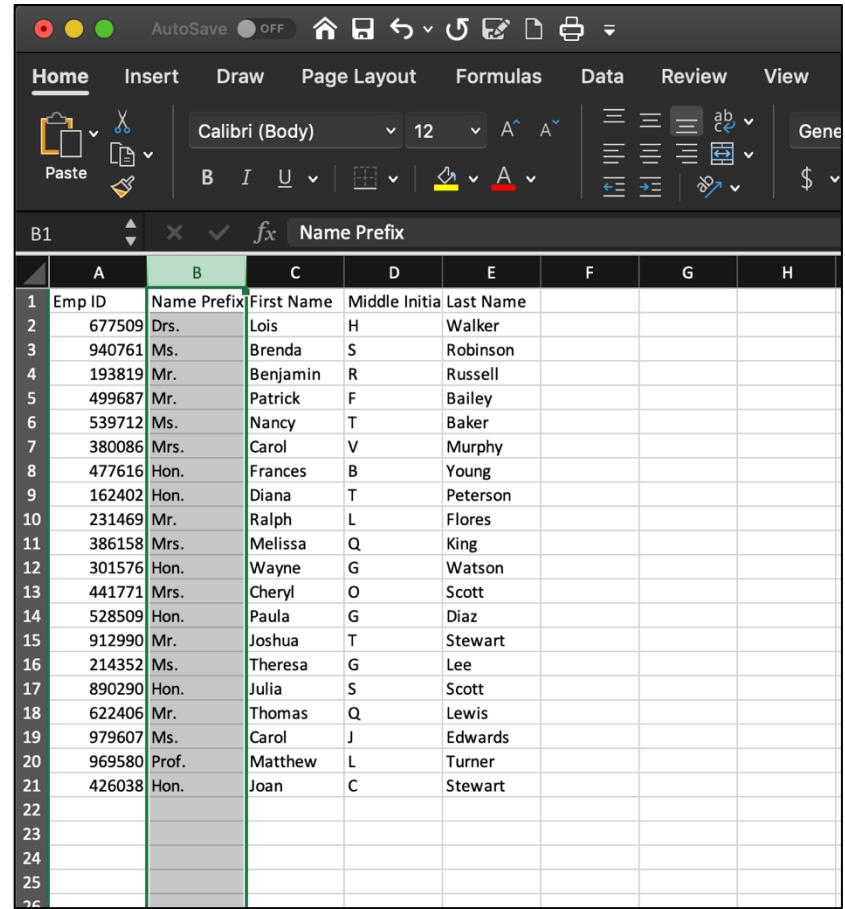
# Data manipulation steps

Deleting Columns - This process can be handled simply by selecting the column that is not needed, right clicking and selecting “Delete”.



A screenshot of Microsoft Excel showing a context menu for column B. The menu options include Cut, Copy, Paste, Paste Special, Insert, Delete, Clear Contents, Format Cells..., Column Width..., Hide, Unhide, and Import Image. The 'Delete' option is highlighted with a red box.

Emp ID	Middle Initial	Last Name
677509		Walker
940761		Robinson
193819		Russell
499687		Bailey
539712		Baker
380086		Murphy
477616		Young
162402		Peterson
231469		Flores
386158		King
301576		Watson
441771		Scott
528509		Diaz
912990		Stewart
214352		Lee
890290	Hon.	Scott
622406	Julia	S
Mr.	Thomas	Q
Ms.	Carol	J
Prof.	Matthew	L
Hon.	Joan	C



A screenshot of Microsoft Excel showing the same data as the first screenshot, but with column B removed. The columns are now labeled A, C, D, E, F, G, H. The data has shifted to the left.

Emp ID	Name Prefix	First Name	Middle Initial	Last Name
677509	Drs.	Lois	H	Walker
940761	Ms.	Brenda	S	Robinson
193819	Mr.	Benjamin	R	Russell
499687	Mr.	Patrick	F	Bailey
539712	Ms.	Nancy	T	Baker
380086	Mrs.	Carol	V	Murphy
477616	Hon.	Frances	B	Young
162402	Hon.	Diana	T	Peterson
231469	Mr.	Ralph	L	Flores
386158	Mrs.	Melissa	Q	King
301576	Hon.	Wayne	G	Watson
441771	Mrs.	Cheryl	O	Scott
528509	Hon.	Paula	G	Diaz
912990	Mr.	Joshua	T	Stewart
214352	Ms.	Theresa	G	Lee
890290	Hon.	Julia	S	Scott
622406	Thomas	Q		Lewis
979607	Ms.	Carol	J	Edwards
969580	Prof.	Matthew	L	Turner
426038	Hon.	Joan	C	Stewart



# Data manipulation steps

Moving Columns – Select the column you wish to move, right click, select “Cut”, select the column to the right of where you want the column to go, right click, and select “Paste”.

The figure consists of three screenshots of Microsoft Excel demonstrating the process of moving a column:

- Screenshot 1:** Shows the "Name Prefix" column (Column F) selected. A context menu is open at cell F1 with the "Cut" option highlighted.
- Screenshot 2:** Shows the "Name Prefix" column moved to the right, occupying the position of the previous "Last Name" column (Column E). A context menu is open at cell F1 with the "Paste" option highlighted.
- Screenshot 3:** The final state where the "Name Prefix" column is now the fourth column (Column F), and the original "Last Name" column has been moved to the end of the sheet (Column G).



# Data manipulation steps

Copy the target format field names into a header record (the first row) on the worksheet. You may have to use Copy & Paste Special and select **Transpose** in order to copy field names listed vertically to be listed across the worksheet.

The screenshot shows two Excel windows. The left window displays a list of employee records with columns for Emp ID, Name, and various suffixes like Drs., Ms., Mr., Hon., etc. The first four rows are selected and have a red border. The right window shows a context menu over the selected rows. The 'Copy' option is highlighted with a red box. A large red arrow points from the 'Copy' option to the 'Transpose' option in the 'Paste Special' submenu of the context menu. The 'Transpose' option is also highlighted with a red box.

H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	AA	AB
Emp ID	677509	940761	193819	499687	539712	380086	477616	162402	231469	386158	301576	441771	528509	912990	214352	890290	622406	979607	969580	426038
Name Prefix	Drs.	Ms.	Mr.	Mr.	Ms.	Mrs.	Hon.	Hon.	Mr.	Mrs.	Hon.	Mr.	Hon.	Mr.	Ms.	Hon.	Mr.	Ms.	Prof.	Hon.
First Name	Lois	Brenda	Benjamin	Patrick	Nancy	Carol	Frances	Diana	Ralph	Melissa	Wayne	Cheryl	Paula	Joshua	Theresa	Julia	Thomas	Carol	Matthew	Joan

# Try it now



Copy the columns from **Exercise** Worksheet, and Paste in **Feeder** Worksheet, the sequence of columns must follow the list below.

1. CustomerID
2. Name\_Final
3. Gender
4. MaritalStatus
5. Age
6. State
7. NumberofDependents
8. NumberBankAccts
9. Savings
10. LastTransactionDate
11. LastTransactionAmt
12. MonthlySalary
13. CreditScore
14. Year
15. Month
16. Day
17. AmountInDollar

# Try it now



Create a new worksheet named “**Charts**”, and select a suitable chart to visualise the data in **Feeder** worksheet:

1. What are the number of male and female customers?
2. What are the number of married and single customers?
3. What are the number of customers from each state?
4. What are the number of male and female customers from each state?
5. What are the number of customers for each year?
6. What is the yearly trend of the number of customer?
7. What are the total transaction for each month?
8. What are the yearly trend for total transactions?

(Hint: determine in each question what would you like to show, either comparison, relationship, composition or distribution; then you can select the right charts accordingly to answer your question and represent your data.)

# Contents

- Overview of training datasets
  - Managing multiple worksheets
  - Data understanding
  - Conditional formatting and logical operators
  - Lookup and reference functions
  - A review of essential functions
  - Fundamentals of descriptive statistics
  - Selecting the right charts to visualize your analysis result
- Dashboard dos and don'ts**



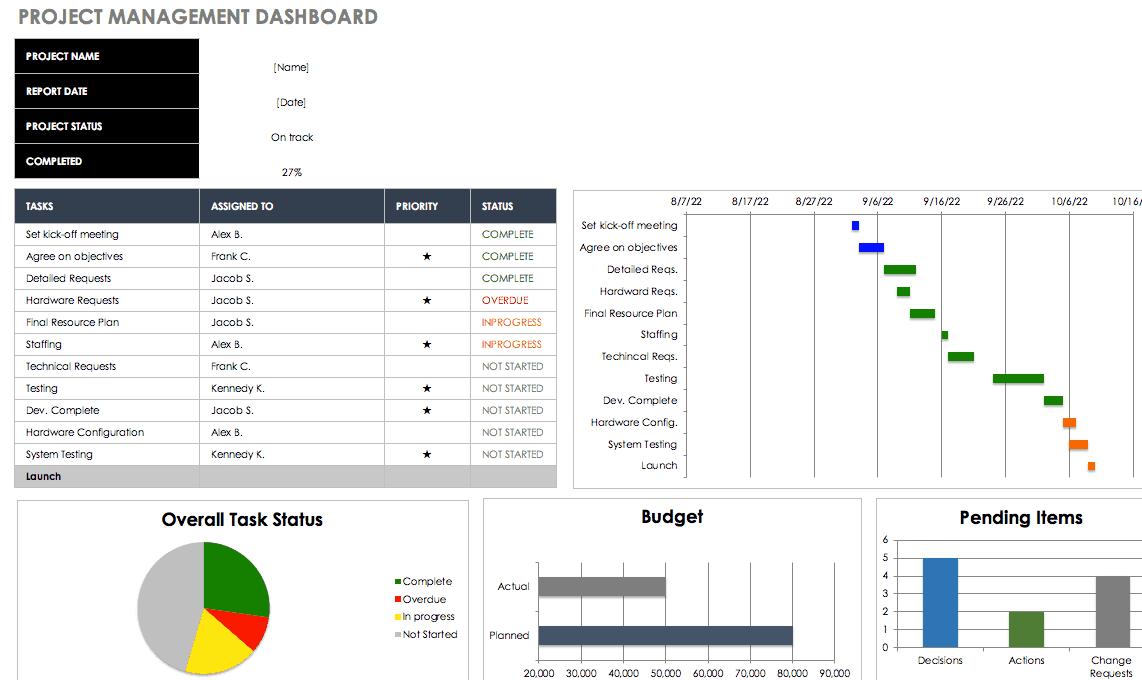
NATURAL INTELLIGENCE SOLUTIONS TECHNOLOGY

# Dashboards



Dashboards track KPIs, metrics, and other data points in one visual, central place. They give you a high-level view of work, helping you make quick decisions and keeping everyone up to date. A dashboard's visual nature simplifies complex data and provides an at-a-glance view of current status or performance in real time.

Dashboards are made up of tables, charts, gauges, and numbers. They can be used in any industry, for almost any purpose. For example, you need to build a customers demographic dashboard for your boss to provide him an insight of the existing customers profile so that he can identify a potential market to target the customers to sell the loan products.



# Questions to ask yourself when dashboarding



Before you start building a dashboard, there are 3 factors you must determine.

- 1. Your purpose:** Is this dashboard for a single project, multiple projects, or a general information hub?

Back to our example:

We need to prepare a dashboard to provide customer demographic to identify target customer group to sell the loan product to increase the revenue.

- 2. Your audience:** Are you building a dashboard for yourself, your team, middle management, or a C-level executive?

Back to our example:

Our dashboard is for the new boss.

- 3. Your data:** What data and information will be most useful for your audience?

Back to our example:

The datasets we have and need to build the dashboard:

Customers personal info & Financial info.

# Dashboard Dos



Here are some general dashboard advice and tips:

1. Keep it simple: A simple, easy-to-understand dashboard is much more effective than a “pretty” dashboard. Avoid the temptation to add 3D effects, gradients, extra shapes, and other bells and whistles. Instead, use magazine formatting. Look at the charts and tables from business magazines and adopt their simplicity and readability.
2. Use freeze panes: You can use freeze panes to make your dashboard more easily navigable. You can freeze all your rows so that viewers can see your whole dashboard in one view, without scrolling. You can also freeze headings in the top rows so that the headings are visible, even when you scroll down the dashboard.
3. Add alerts for important information: If you want to highlight a certain aspect of your dashboard or quickly convey status, add alerts (or red, yellow, green balls). You can use these traffic light symbols with conditional formatting, so they automatically update based on the values in your table.
4. Use shapes and charts together: The layout of your dashboard can get repetitive if you have multiple charts all in the same box layout. You can add visual interest and hierarchy to your dashboard by inserting a shape (like a rectangle, circle, etc) and then putting your chart on top of that shape.
5. Use different tabs: Keep your dashboard organized by using different sheets for different things. For example, one tab will hold the dashboard itself and another tab will hold the raw data that populates the dashboard.

# Dashboard Don'ts



Here are some general dashboard advice and tips that you should avoid:

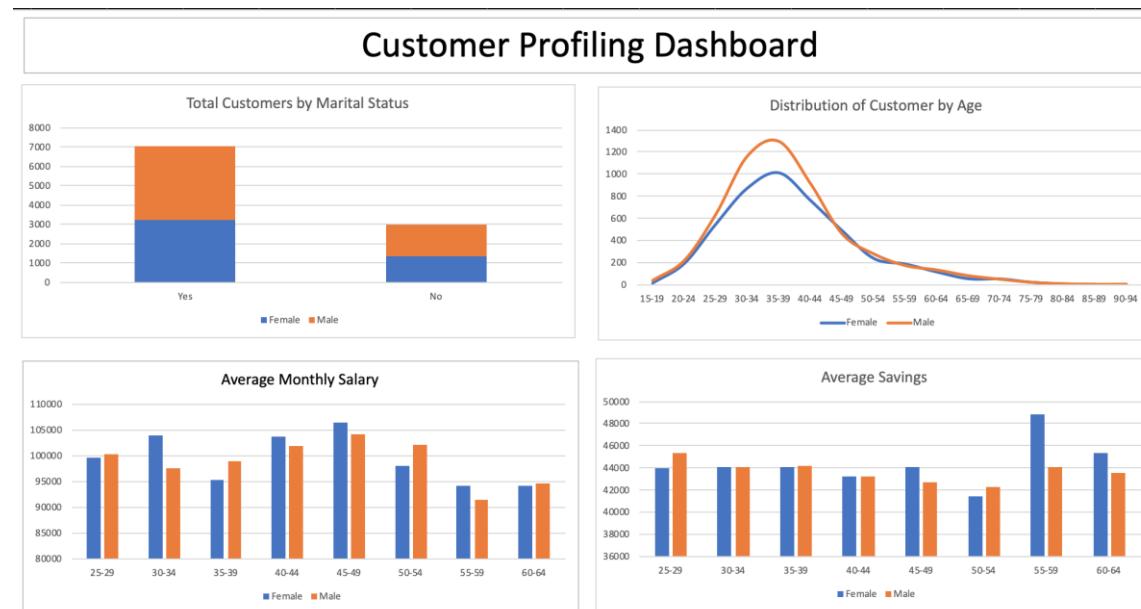
1. Bold, bright colours: Don't add a rainbow-themed colour palette to your dashboard in hopes of making it look more "fun." These bright colours distract from the important information. Instead, use muted colours and only add stronger colours to highlight key items.
2. Crowded layout: Don't include every possible data set or chart to your dashboard. Too much data will overwhelm the viewer and end up hiding the really important information. If you end up with a crowded dashboard, take a step back and re-evaluate if everything is necessary. All the data should support the single purpose of your dashboard.
3. Lack of focus: A crowded layout and lack of focus usually go hand-in-hand. Make sure all your charts are supporting the same purpose or hypothesis and cut out all the extras. The data should tell the same story.

# Try it now



Select the right chart to visualise the result of descriptive analysis you conducted previously in **Dashboard** worksheet:

1. total customers by marital status and by gender
2. the distribution of customers by gender with different age group e.g. age 15-19, 20-24, 25-29, etc
3. average monthly salary by age group
4. average saving by age group





## MALAYSIA

NATURAL INTELLIGENCE SOLUTIONS TECHNOLOGY SDN BHD  
16<sup>th</sup> Floor, Office Block Plaza Berjaya, 12 Jalan Imbi, 55100 Kuala Lumpur

## SINGAPORE

NATURAL INTELLIGENCE SOLUTIONS PTE LTD  
133 Cecil Street, #17-01B Keck Seng Tower  
Singapore 069535

## PHILLIPINES

NATURAL INTELLIGENCE SOLUTIONS PTE LTD  
A Clean & Well-Lighted Place, 27F Jollibee Plaza  
F. Ortigas Jr. Road, Ortigas Center, Pasig, Philippines 1605



NATURAL INTELLIGENCE SOLUTIONS TECHNOLOGY

