1st Meet Up

Introduction to Data Visualization

Anton Suhartono

DQLab

AGENDA

Review

- *Why Data Visualization*
- *Definition Data Visualization*
- *Step Creating Data Visualization*
- *Type of Data Visualization*
- *Choosing the right Chart*
- *Tools for Data Visualization*
- *Introduction to Pandas*
- *Introduction to SQL Database*

Practice

- *How to Manipulate data using SQL & Python*
- *Understanding & Importing Data*
- *Selecting Data based on Criteria*
- *Grouping & Aggregation*
- *Creating new Column based on Criteria*

Definition of Data Visualization

Why need Visualization

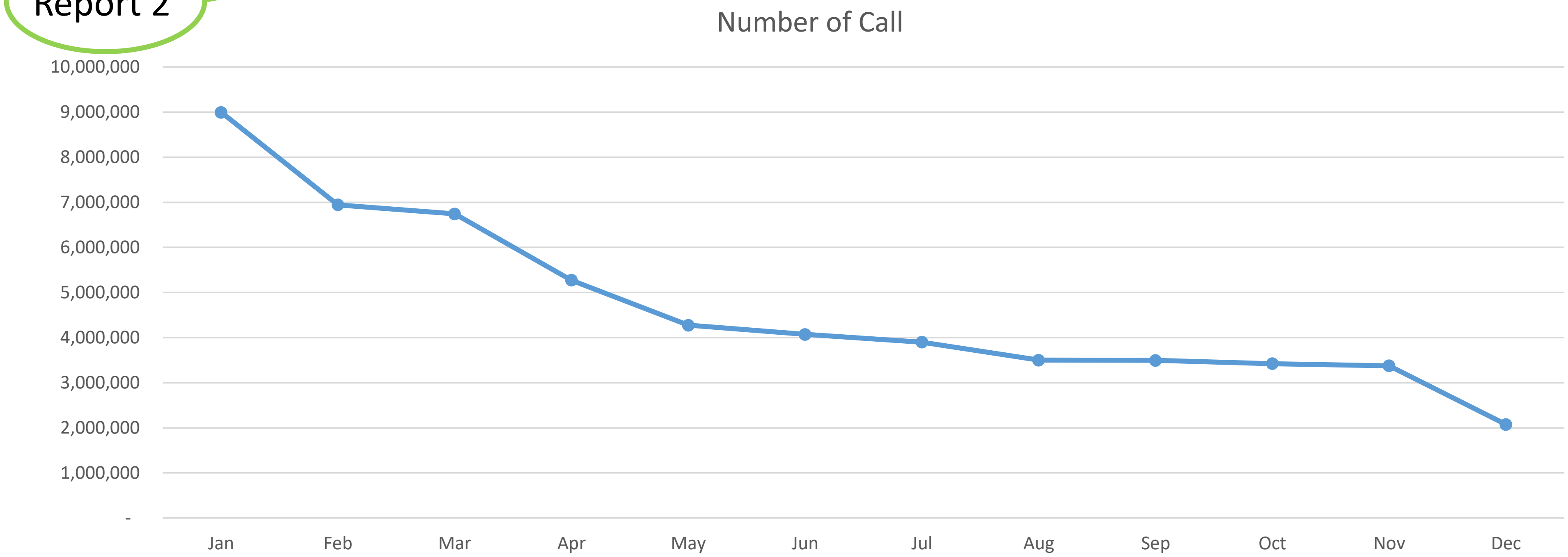
Illustration 1

Report 1

Month	January	February	March	April	May	June	July	August	September	October	November	December
Number of Call	8,994,827	6,942,827	6,742,927	5,273,429	4,275,429	4,070,429	3,900,029	3,500,029	3,495,029	3,422,220	3,375,429	2,075,429

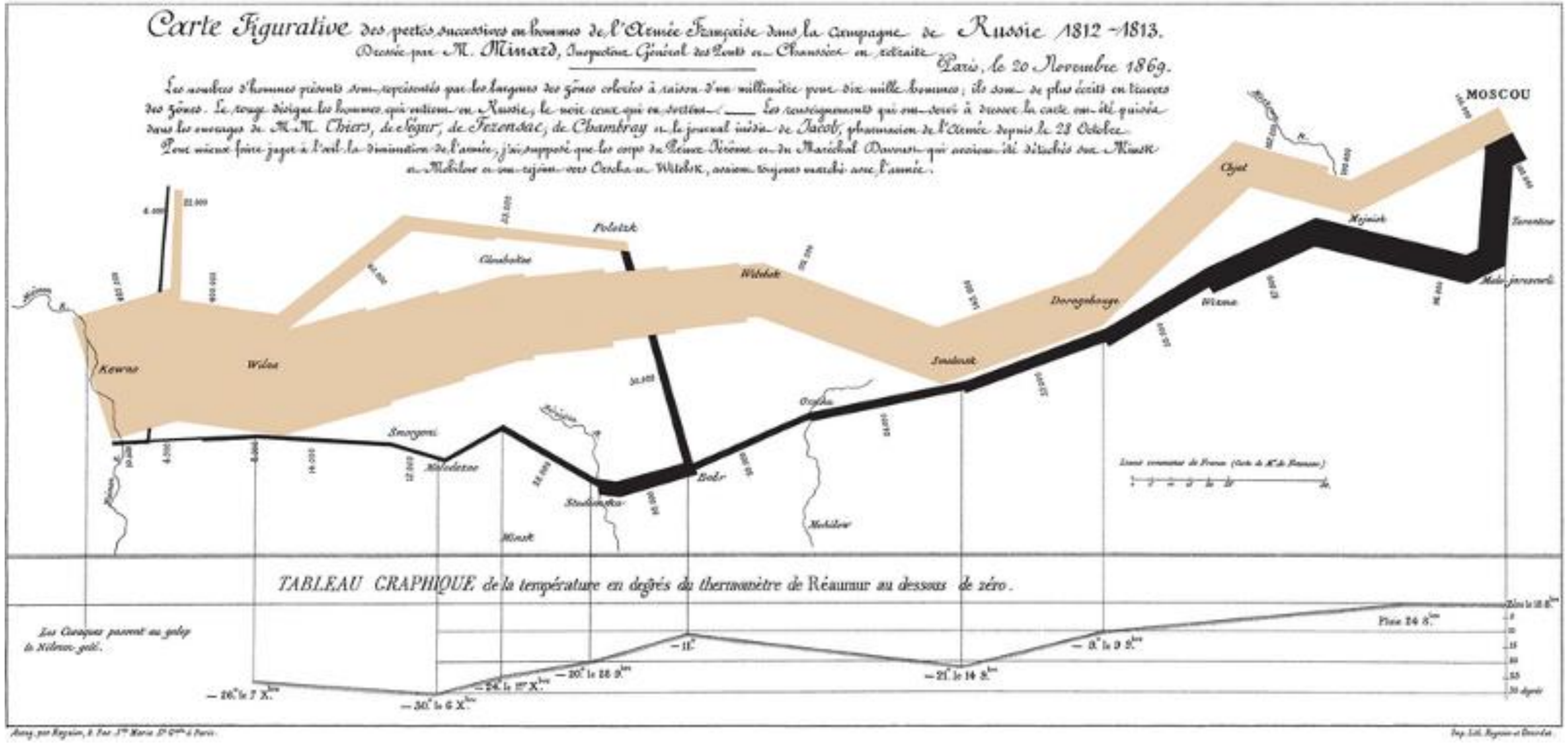
Report 2

Easy to understand



Key Figures in the History of Data Visualization

Illustration 2

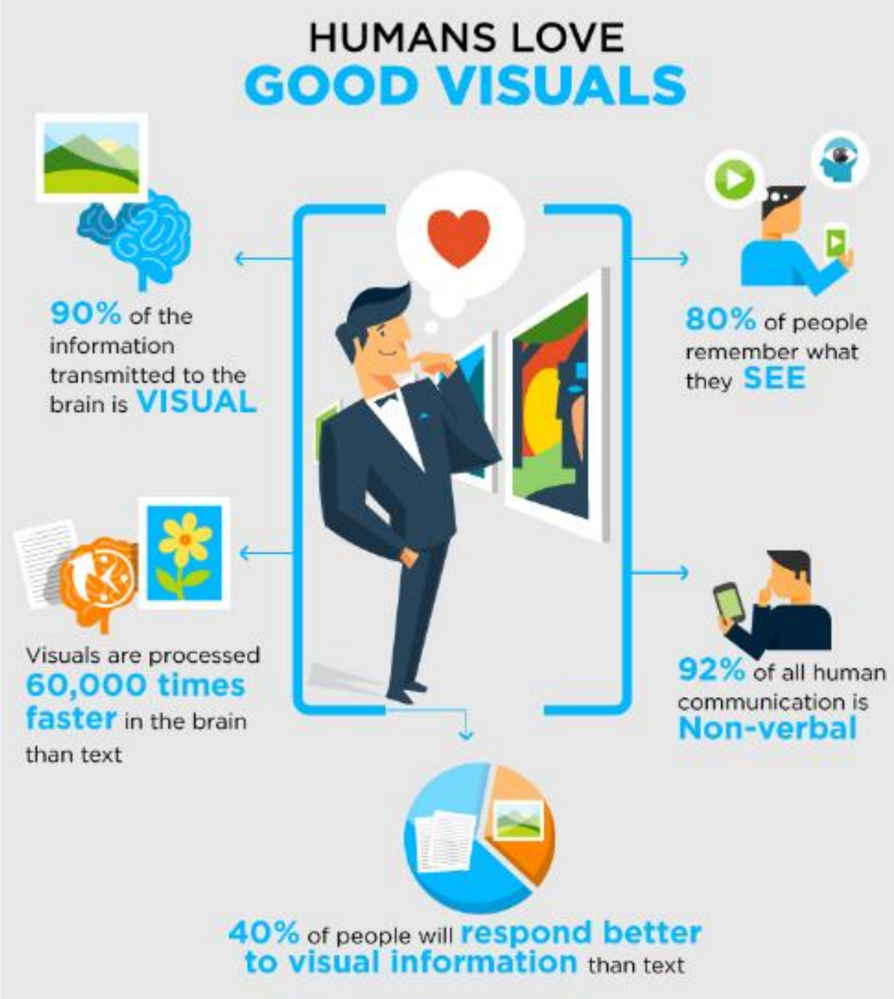


Charles Joseph Minard (1781–1870)

Charles Joseph Minard was a French civil engineer famous for his representation of numerical data on maps. His most famous work is the map of Napoleon's Russian campaign of 1812 illustrating the dramatic loss of his army over the advance on Moscow and the following retreat. This classic lithograph dates back to 1869, displaying the number of men in Napoleon's 1812 Russian army, their movements, and the temperatures they encountered along their way. It has been called one of the "best statistical drawings ever created."

Human Perspective on Visualization

Illustration 1


HUMANS LOVE
GOOD VISUALS
90% of the information transmitted to the brain is VISUAL
80% of people remember what they SEE
Visuals are processed 60,000 times faster in the brain than text
92% of all human communication is Non-verbal
40% of people will respond better to visual information than text

1. To convey **information** through **visual** representation
2. Produces(interactives) **visual representations** of abstract data **to reinforce human cognition**; thus enabling the viewer to gain knowledge about the internal structure of the data and causal relationships in it

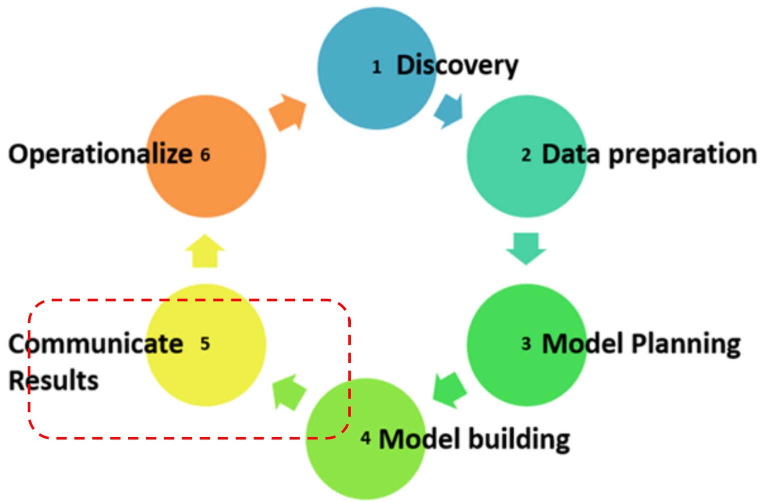
Purpose Of Data Visualization

3 Questions of Data Visualization

Are You Exploring Data ?
Used for exploratory Data Analysis (EDA), affirmation of hypothesis, etc

Are You Formatting it for Decision Making ?
Are you presenting a neutral case so your audience can use the info to make their own decision

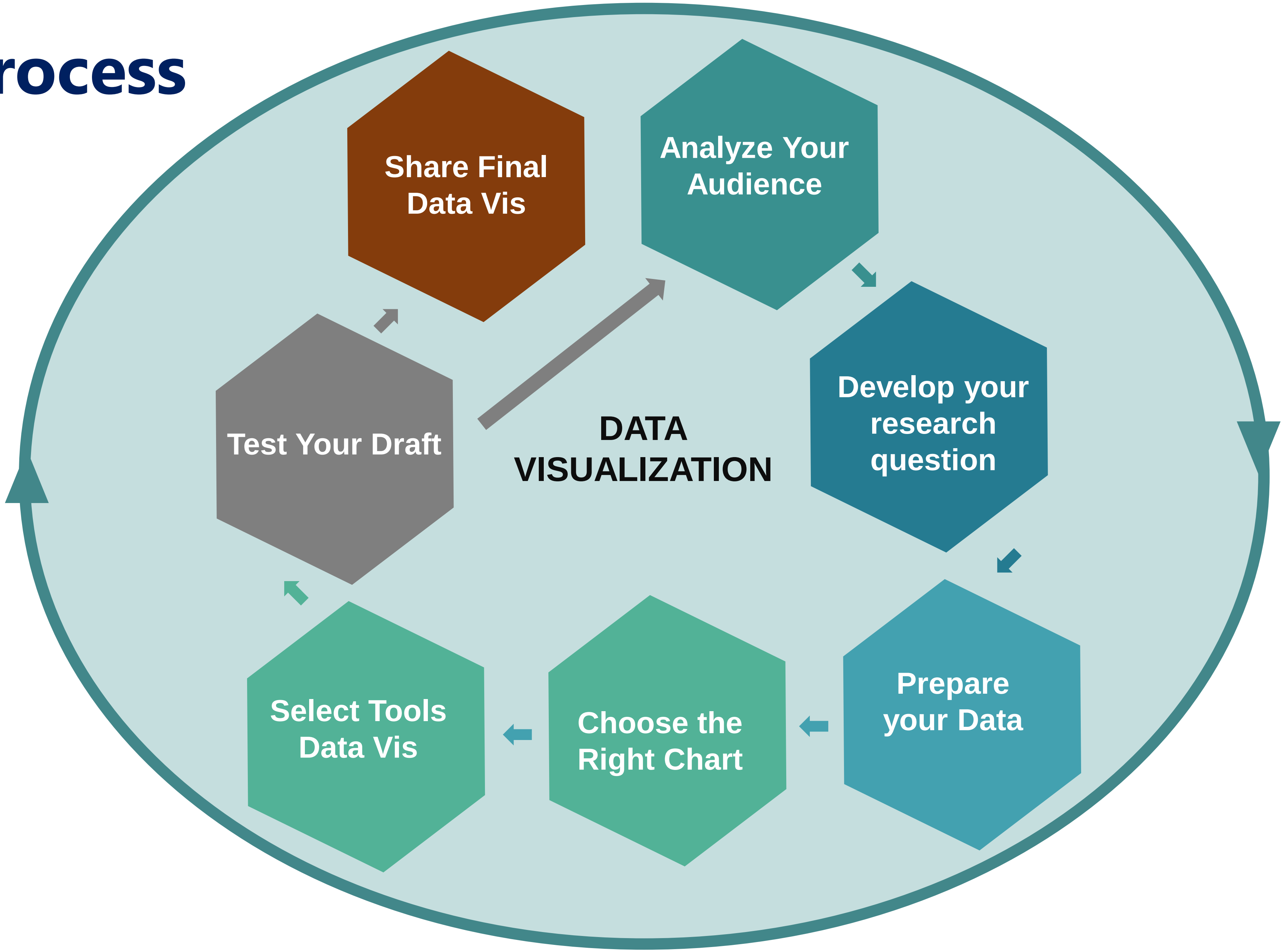
Are You telling Story ?
Used for affirmation of opinion



What is Data Visualization ?

Data visualization is a graphic representation that expresses the significance of data. It **reveals insights** and patterns that **are not** immediately **visible** in **the raw data**. It is an art through which information, numbers, and measurements can be made more understandable.

7 Steps of Data Visualization Process



DATA VISUALIZATION

Analyze Your Audience

Develop your research question

Prepare your Data

Choose the Right Chart

Select Tools Data Vis

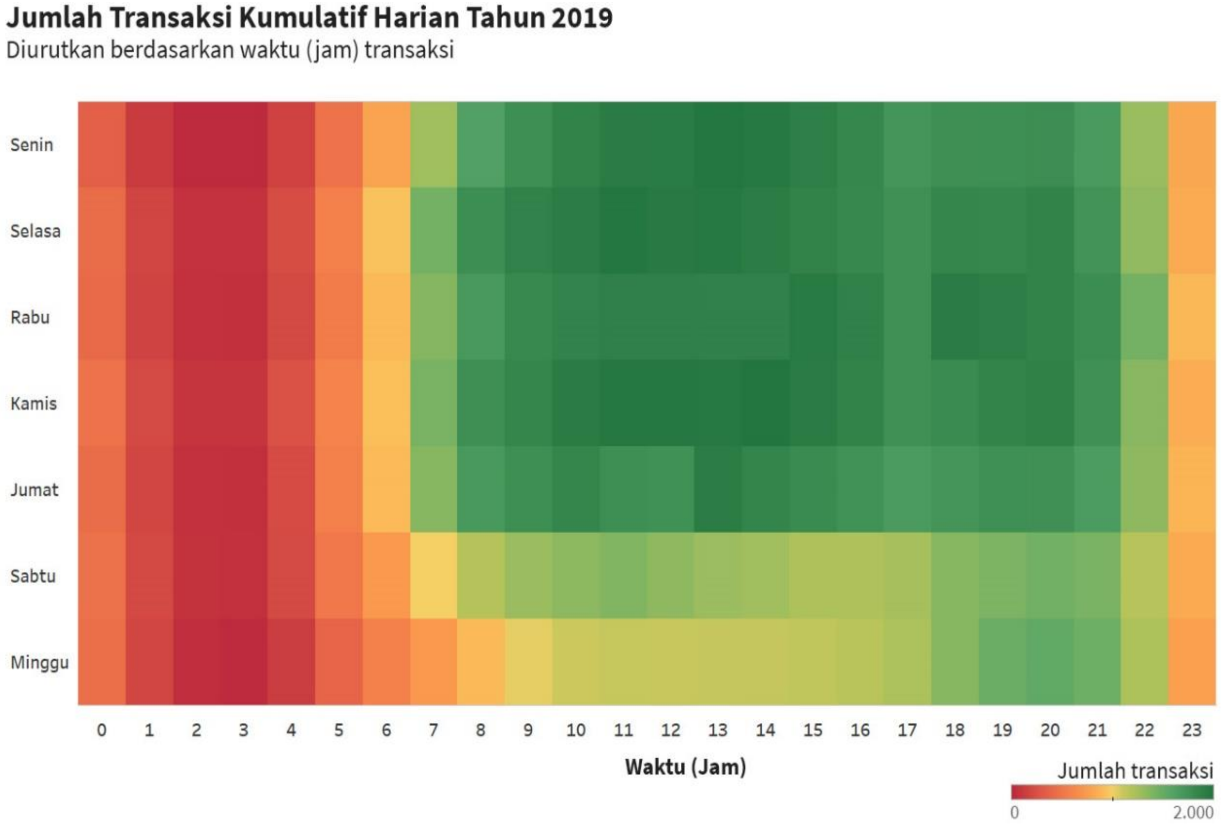
Test Your Draft

Share Final Data Vis

Type of Data Visualization

Narrative vs Explorative

Narrative Visual


Jumlah Transaksi Kumulatif Harian Tahun 2019

Diurutkan berdasarkan waktu (jam) transaksi

Gambar 1: Heatmap menggambarkan transaksi kumulatif harian selama satu tahun. Grafik ini tidak menampilkan data secara detail karena tujuan utamanya adalah memperlihatkan pada jam berapa transaksi tertinggi dan terendah terjadi.

Narrative Visual

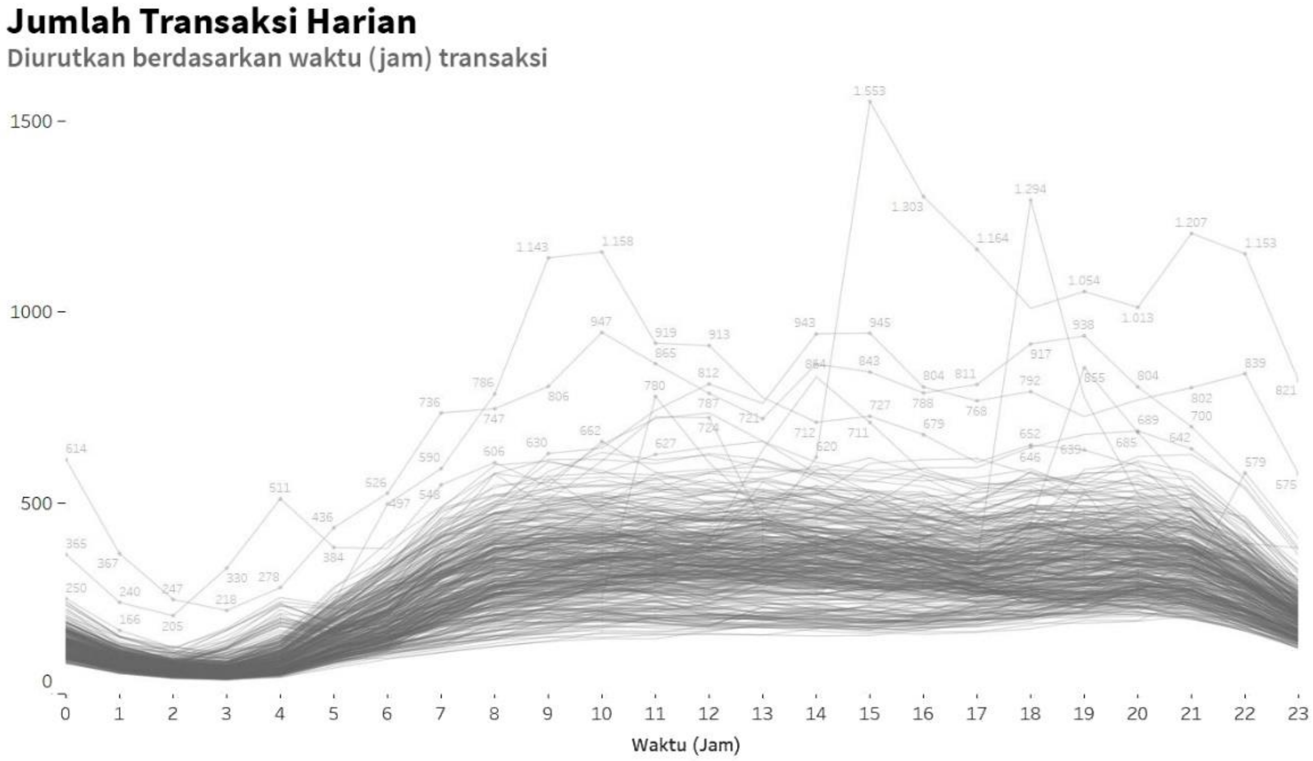
- *Usually used to explain the final results or conclusions of the analyst*
- *Static*
- *Using Visual Beauty*
- *Explanation not Detail*
- *Easy to Understand*

Narrative vs Explorative

Explorative Visual

Explorative Visual

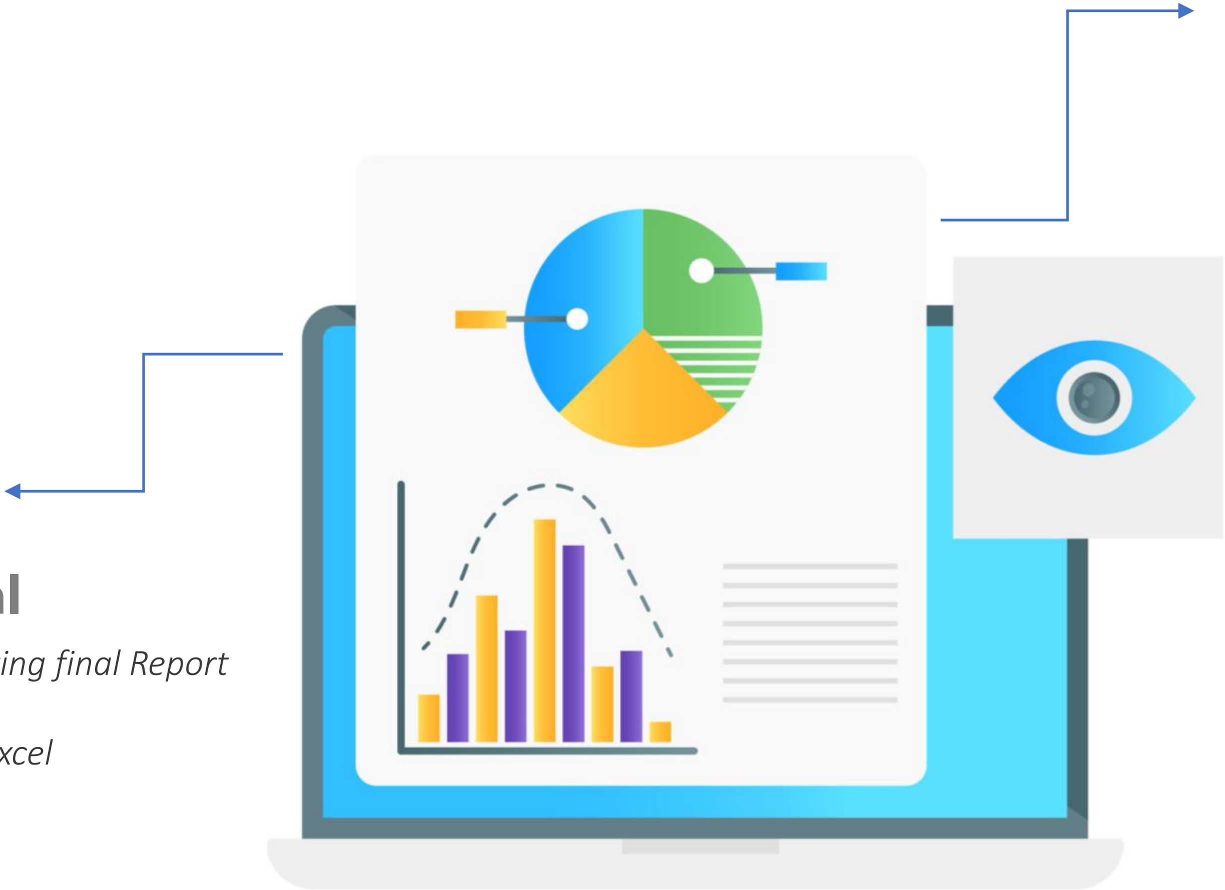
- *Describes the process carried out to get the right end result*
- *Complex & Detail*
- *Selective Audience*



Gambar 2: Grafik transaksi harian selama satu tahun. Grafik menggunakan elemen secara detail untuk memperlihatkan performa per jam setiap hari.

Static & Dynamic

Definition



Dynamic Visual

Usually for presenting Report Periodically
- tableau, d3, plotly-dash, etc

Static Visual

Usually for presenting final Report or exploring data
- matplotlib, ppt, excel

Choosing the Right Data Visualization

Choosing Chart

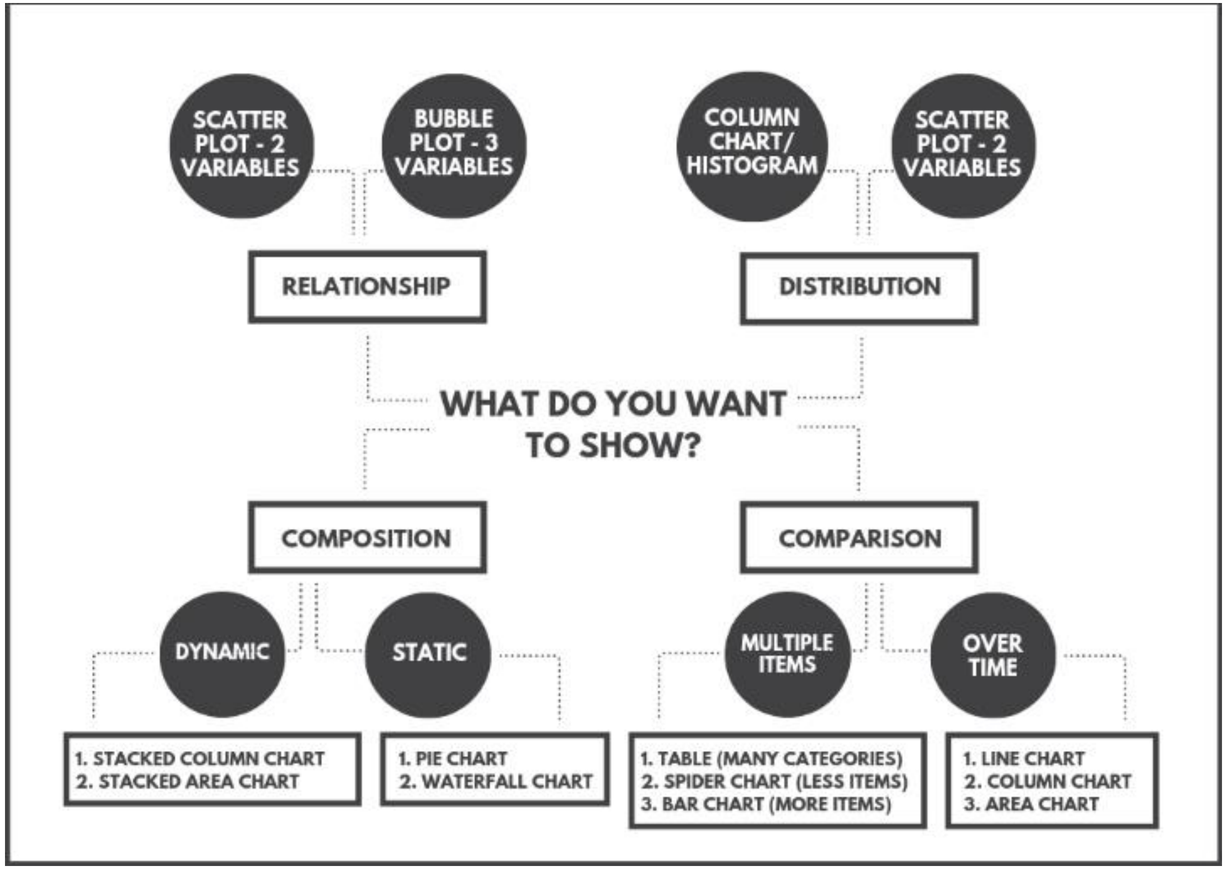
Goals

The graph guide breaks up your options into 4 paths:

1. Comparison
2. Relationship
3. Distribution
4. Composition

Every data visualization project or initiative is slightly different, which means that different data visualization chart types will suit varying goals, aims, or topics.

Chart to select based on what kind of data you need to show



WHAT DO YOU WANT TO SHOW?

RELATIONSHIP

SCATTER PLOT - 2 VARIABLES

BUBBLE PLOT - 3 VARIABLES

DISTRIBUTION

COLUMN CHART/ HISTOGRAM

SCATTER PLOT - 2 VARIABLES

COMPOSITION

DYNAMIC

1. STACKED COLUMN CHART
2. STACKED AREA CHART

STATIC

1. PIE CHART
2. WATERFALL CHART

COMPARISON

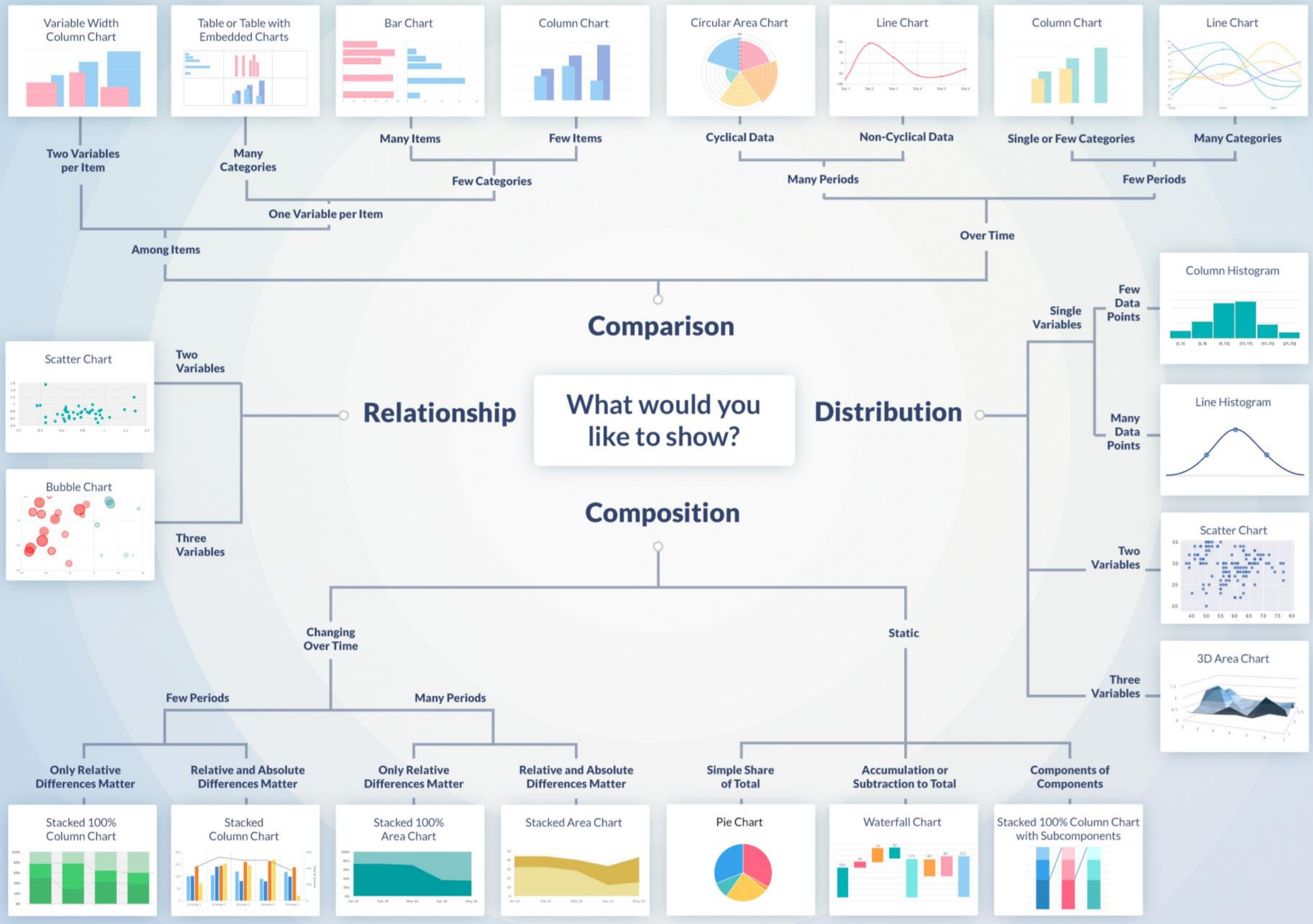
MULTIPLE ITEMS

1. TABLE (MANY CATEGORIES)
2. SPIDER CHART (LESS ITEMS)
3. BAR CHART (MORE ITEMS)

OVER TIME

1. LINE CHART
2. COLUMN CHART
3. AREA CHART

Guided Visualizations for Charts and Graphs



What would you like to show?

Comparison

Among Items

Two Variables per Item

Variable Width Column Chart

One Variable per Item

Many Categories

Table or Table with Embedded Charts

Few Categories

Many Items

Bar Chart

Few Items

Column Chart

Over Time

Many Periods

Cyclical Data

Circular Area Chart

Non-Cyclical Data

Line Chart

Few Periods

Single or Few Categories

Column Chart

Many Categories

Line Chart

Relationship

Two Variables

Scatter Chart

Three Variables

Bubble Chart

Distribution

Single Variables

Few Data Points

Column Histogram

Many Data Points

Line Histogram

Two Variables

Scatter Chart

Three Variables

3D Area Chart

Composition

Changing Over Time

Few Periods

Only Relative Differences Matter

Stacked 100% Column Chart

Relative and Absolute Differences Matter

Stacked Column Chart

Many Periods

Only Relative Differences Matter

Stacked 100% Area Chart

Relative and Absolute Differences Matter

Stacked Area Chart

Static

Simple Share of Total

Pie Chart

Accumulation or Subtraction to Total

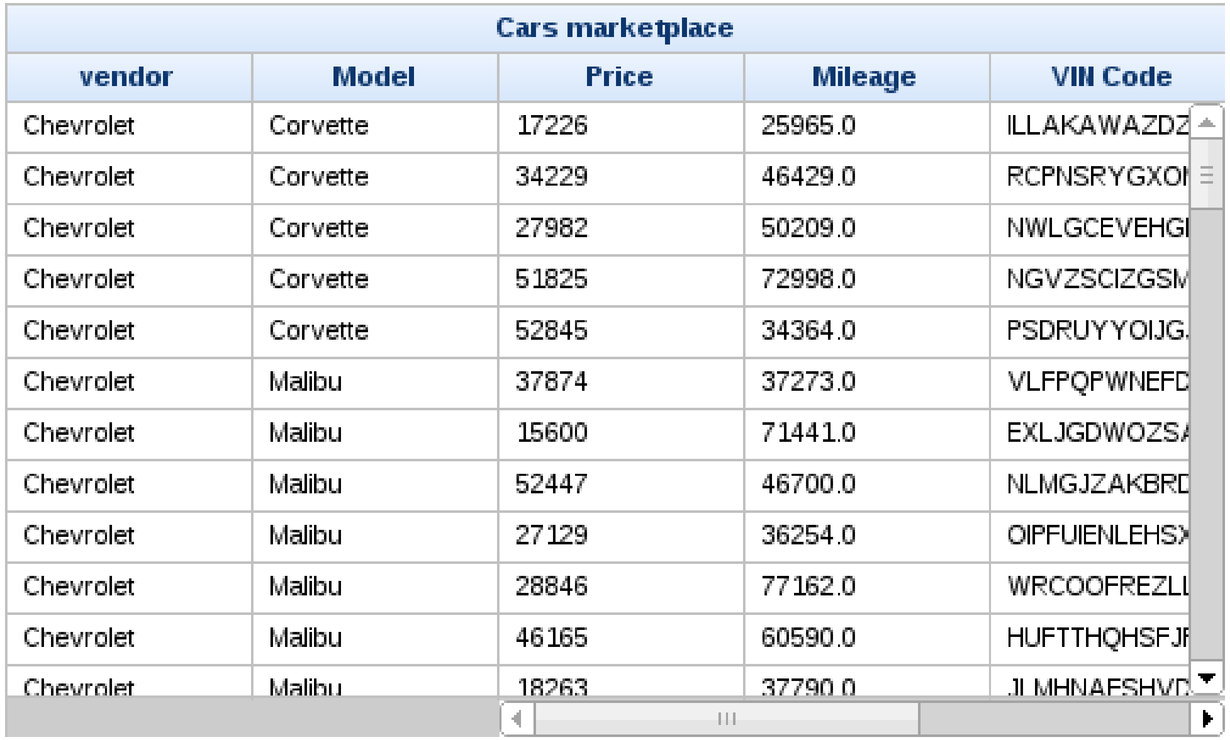
Waterfall Chart

Components of Components

Stacked 100% Column Chart with Subcomponents

Table

Definition, Usage, Tips & Tricks



Cars marketplace				
vendor	Model	Price	Mileage	VIN Code
Chevrolet	Corvette	17226	25965.0	ILLAKAWAZDZ
Chevrolet	Corvette	34229	46429.0	RCPNSRYGXON
Chevrolet	Corvette	27982	50209.0	NWLGCEVEHGI
Chevrolet	Corvette	51825	72998.0	NGVZSCIZGSM
Chevrolet	Corvette	52845	34364.0	PSDRUYYOIJG.
Chevrolet	Malibu	37874	37273.0	VLFPQPWNEFD
Chevrolet	Malibu	15600	71441.0	EXLJGDWOZSA
Chevrolet	Malibu	52447	46700.0	NLMGJZAKBRD
Chevrolet	Malibu	27129	36254.0	OIPFUIENLEHSX
Chevrolet	Malibu	28846	77162.0	WRCOOFREZLL
Chevrolet	Malibu	46165	60590.0	HUFTTHQHSFJF
Chevrolet	Malibu	18263	37790.0	JLMHNAESHVD

Definition:

Data tables display information in a grid-like format of rows and columns.

Visual Dimensions:

Columns, Value of Data

Usage:

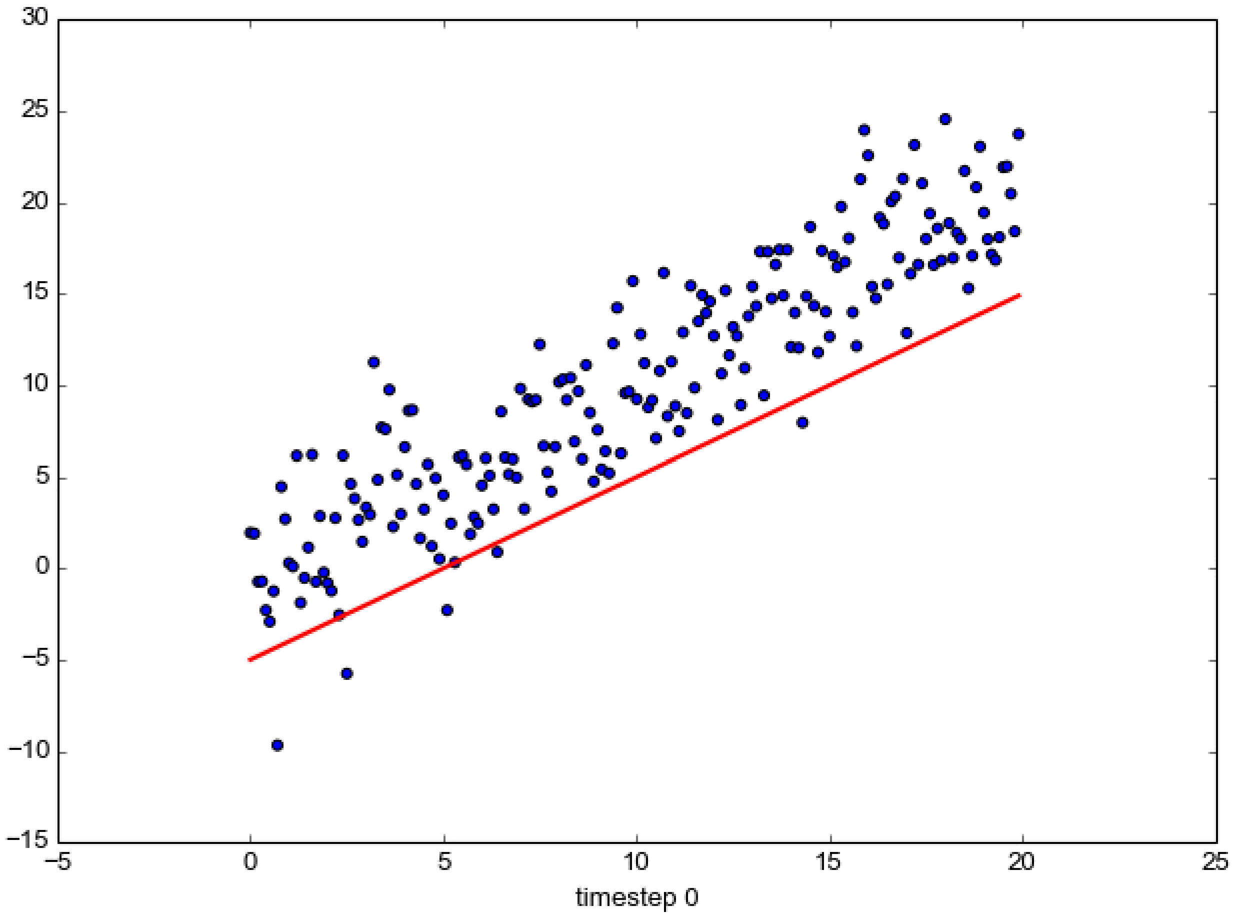
Detail Observation

Scatter Plot

Definition, Usage, Tips & Tricks

Definition:
This graph is used to describe the relationship between two variables. The X axis represents abstract values that are independent of other variables, so they are called independent variables. The value of Y is the dependent variable and is placed on the vertical axis.

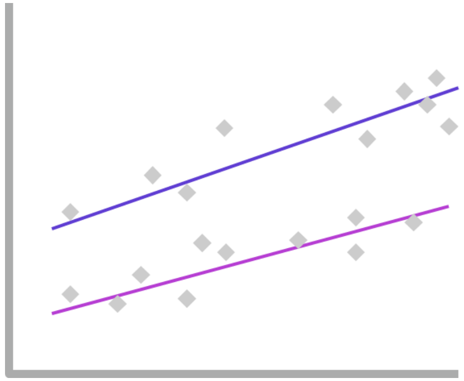
Visual Dimensions:
Length, line, dot
Usage:
- Correlation two variables
- Perfect to use for large data sets such as population or epidemiology studies.



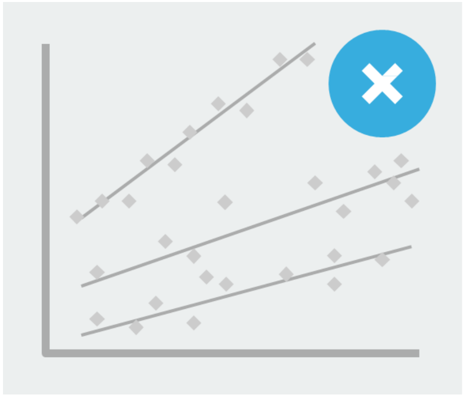
Scatter Plot

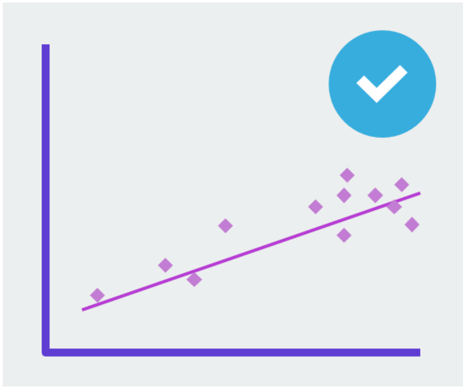
Definition, Usage, Tips & Tricks

Use **lines** to show trends & relationships.

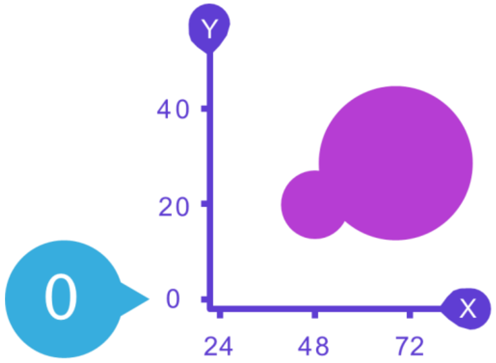


Use as **few lines** as possible





Always **start** with the Y-axis at **0.**

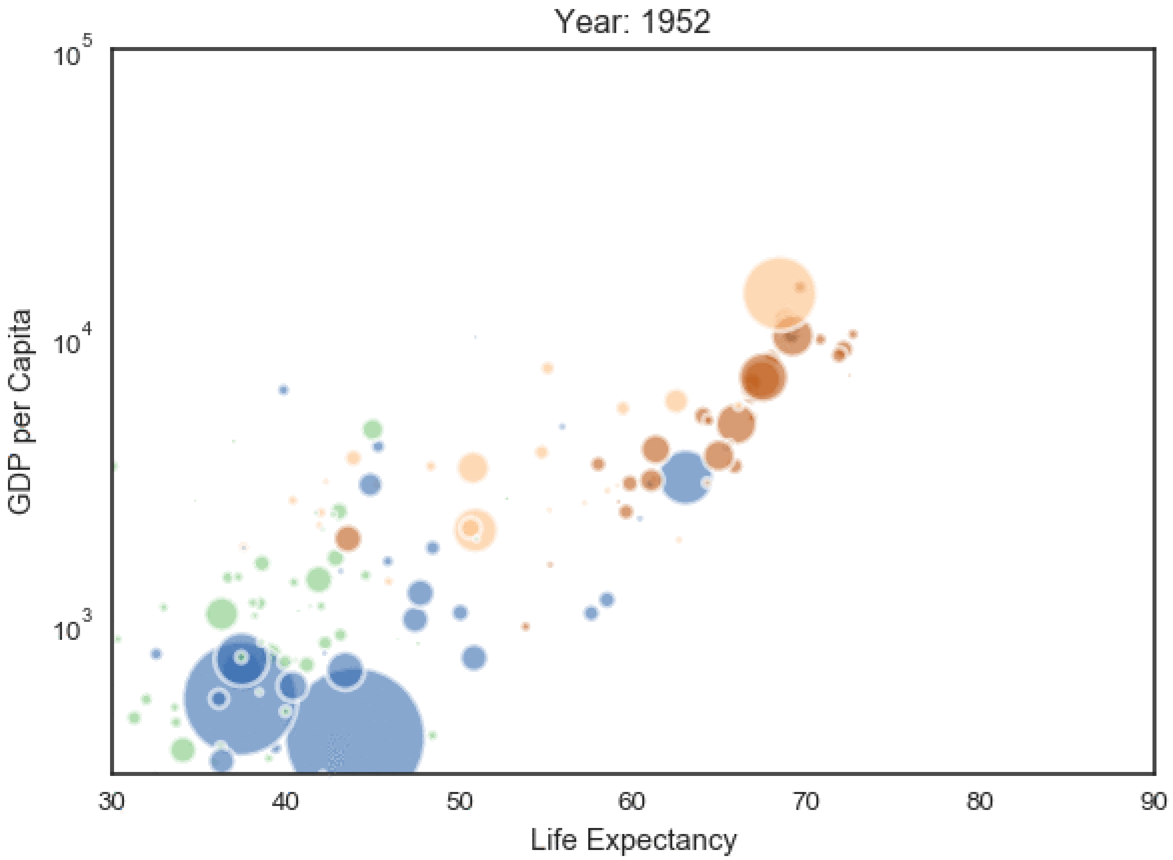


Bubble Chart

Definition, Usage, Tips & Tricks

Definition:
is a variation of a scatter chart in which the data points are replaced with bubbles, and an additional dimension of the data is represented in the size of the bubbles.

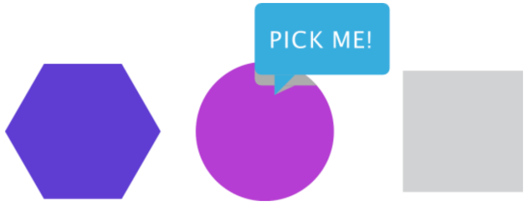
Visual Dimensions:
Length, line, dot, size, color
Usage:
Correlation two variables in dimension



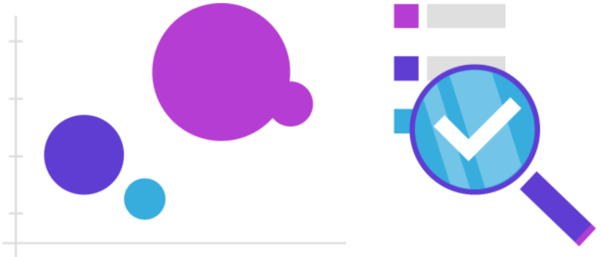
Bubble Chart

Definition, Usage, Tips & Tricks

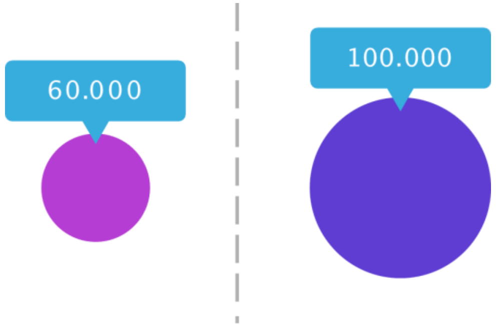
Use **simple** shapes.
Circles work best.



Use **clear** and visible labels.

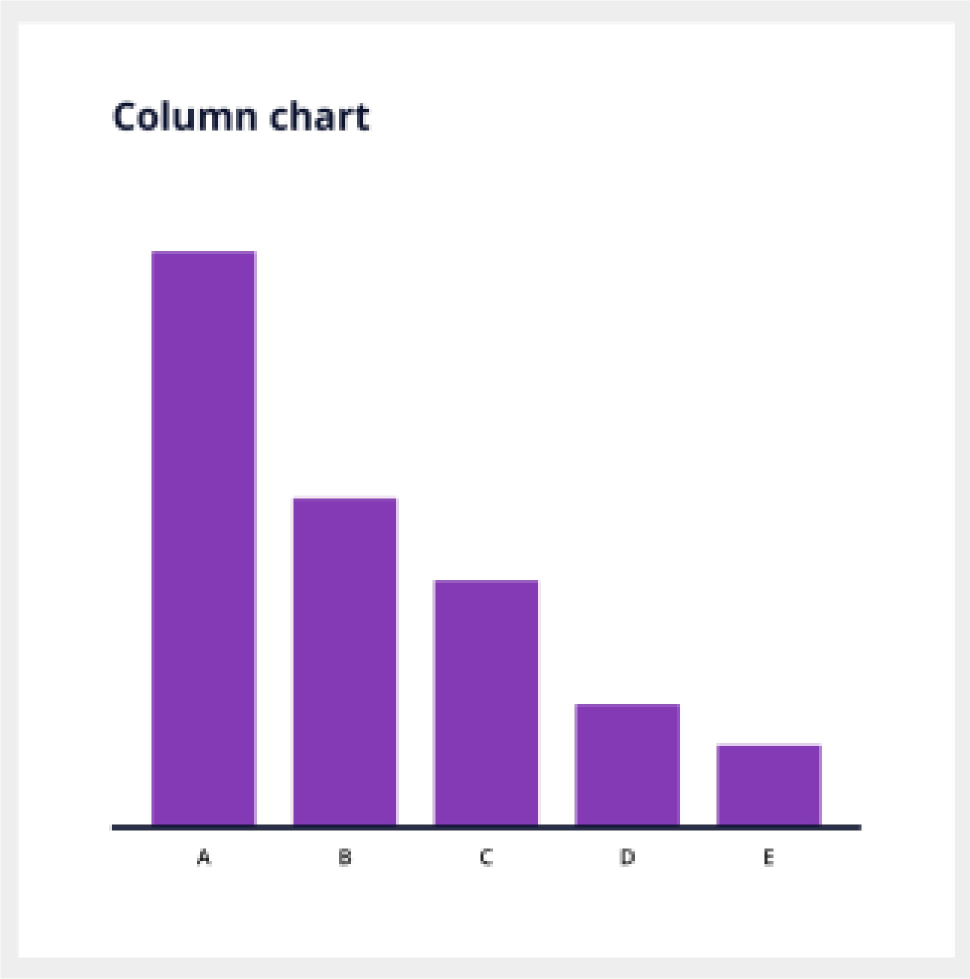


Size bubbles appropriately.



Column Chart

Definition, Usage, Tips & Tricks



Definition:
Column charts or vertical charts can be used to compare a number of categories and/or their changes in a certain time period (trend). When used to display trends, they function the same as line charts.

Visual Dimensions:
Length, category, color

Usage:
compare a number of categories and/or their changes in a certain time period (trend)

Tips & Tricks:

- Multiple categories, use a different color for each category, or use the darker color the more prominent.
- This graph will be difficult to read if it contains too many categories.
- Always use zero baseline or zero point on the Y axis.
- Use a consistent scale.

Bar Chart

Definition, Usage, Tips & Tricks

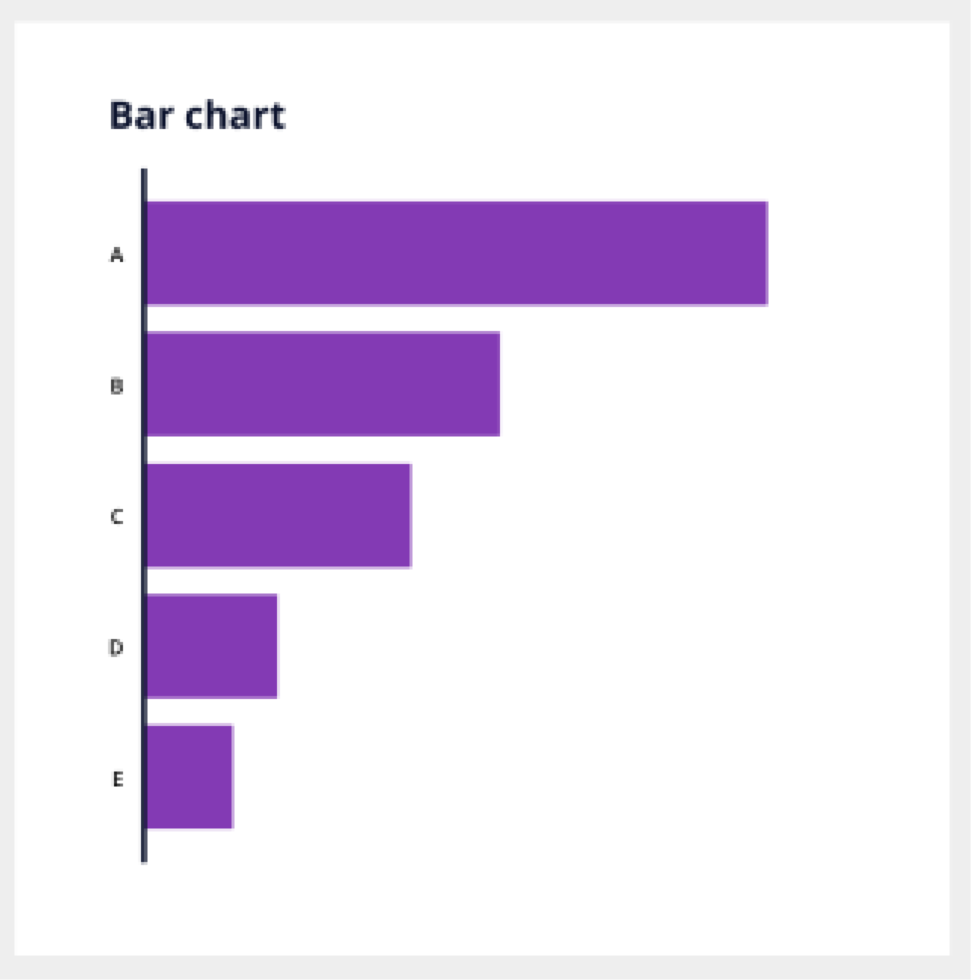
Definition:
Bar charts use horizontal bars to display data and are used to compare values across categories. The lengths of the bars are proportional to the values they represent.

Visual Dimensions:
Length, category, color

Usage:
Best suited for data comparisons with multiple categories or data series (data series)

Tips & Tricks:

- For ease of reading data, you can sort categories based on their value, for example from the highest to the lowest value
- It is different with data series, where data is distributed based on tiered categories, for example the population based on age range or education level.



Histogram

Definition, Usage, Tips & Tricks

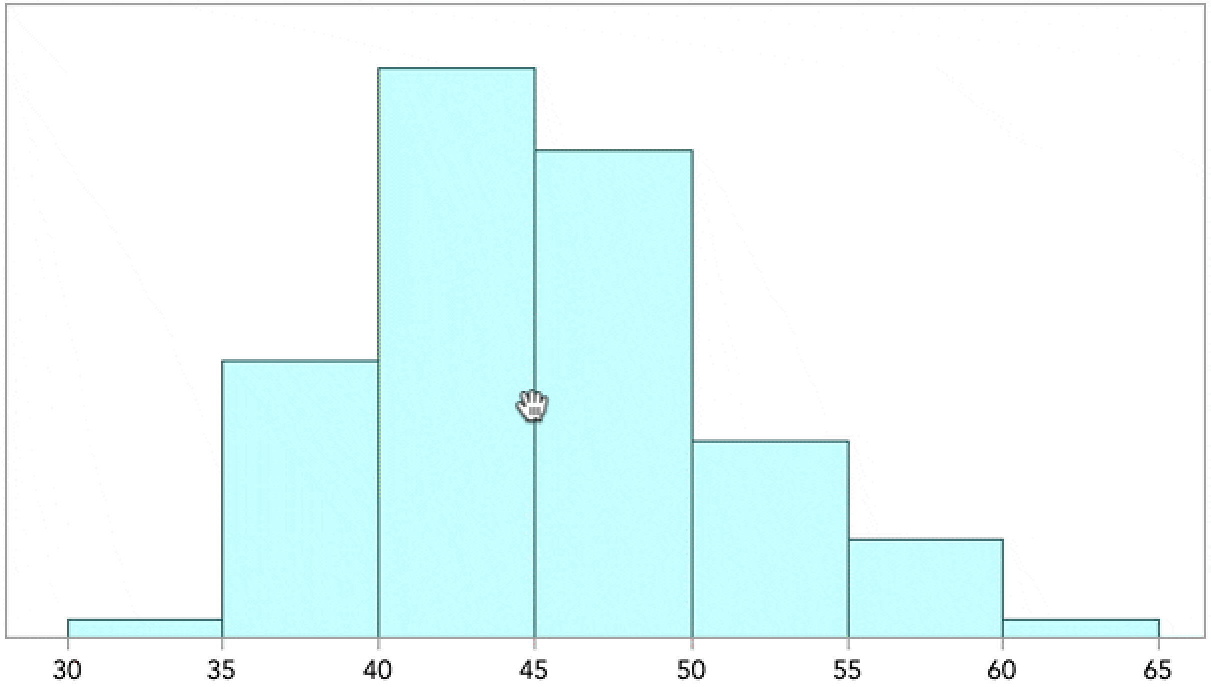
Definition:
A graphical display of data using bars of different heights. At first glance this chart is similar to a bar/column chart. However, there is actually a fundamental difference between a histogram and a bar graph. The distance between the columns / rods is made as close as possible, even sticking. From a visual perspective, this narrow distance will bring the reader's eye to connect groups of data and sort them based on certain criteria.

Visual Dimensions:
Length, category, color

Usage:
displays the shape and spread of continuous sample data

Tips & Tricks:

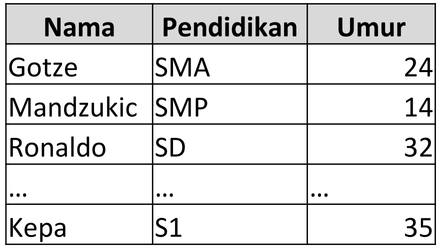
- Always use zero baseline or zero point on the Y axis.
- No space between categories



Column Chart vs Histogram

Definition, Usage, Tips & Tricks

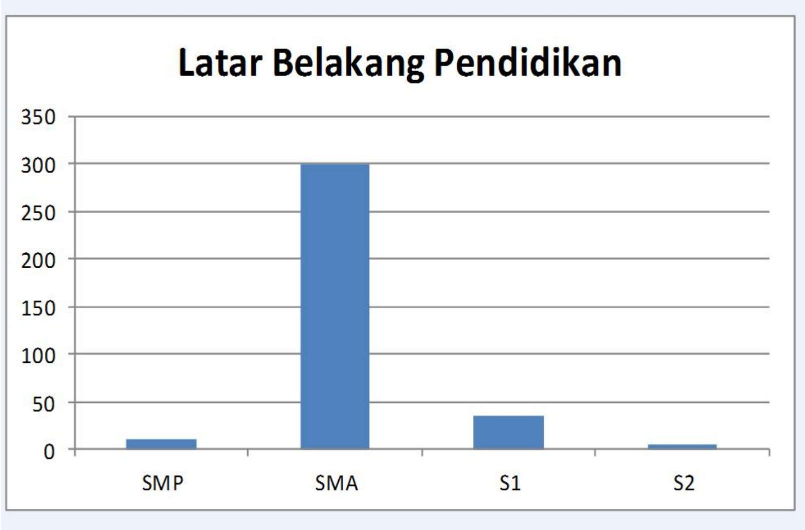
For example Variables in Data:



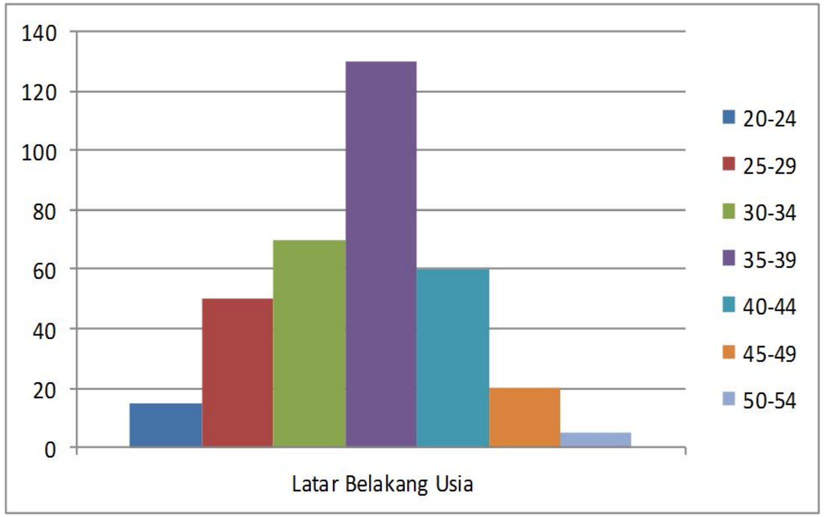
Nama	Pendidikan	Umur
Gotze	SMA	24
Mandzukic	SMP	14
Ronaldo	SD	32
…	…	…
Kepa	S1	35

Maka:

Column Chart



Histogram



Box Plot

Definition, Usage, Tips & Tricks

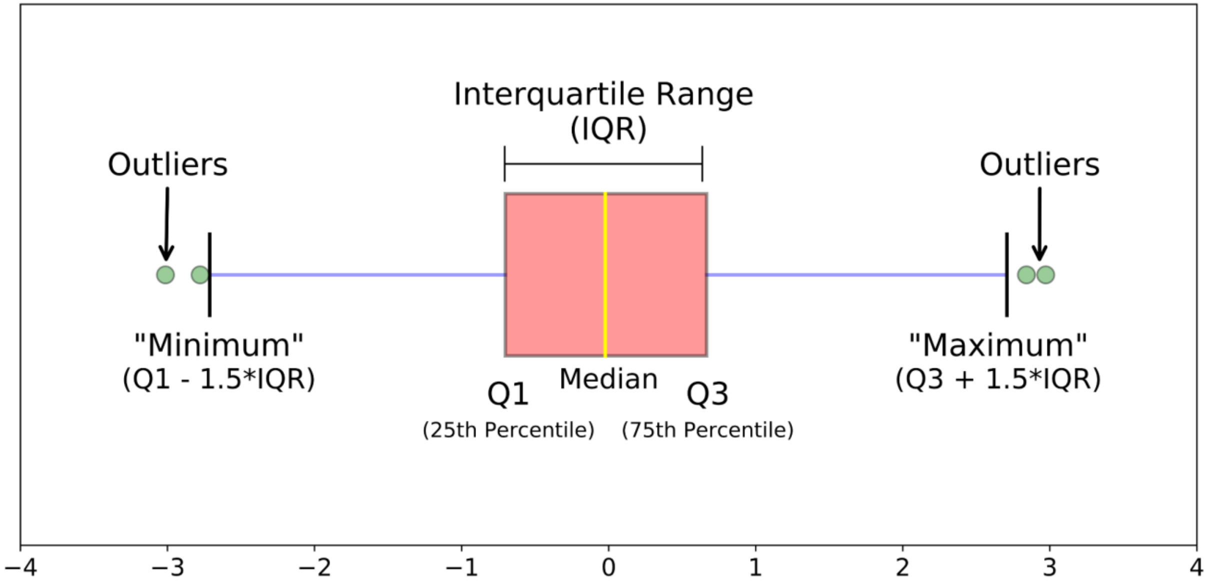
Definition:

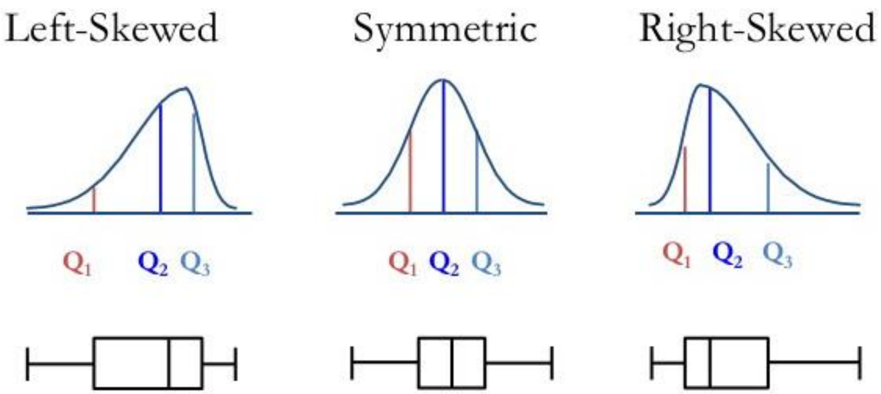
Box plots visually show the distribution of numerical data and skewness through displaying the data quartiles (or percentiles) and averages.

Usage:

Show distribution of data

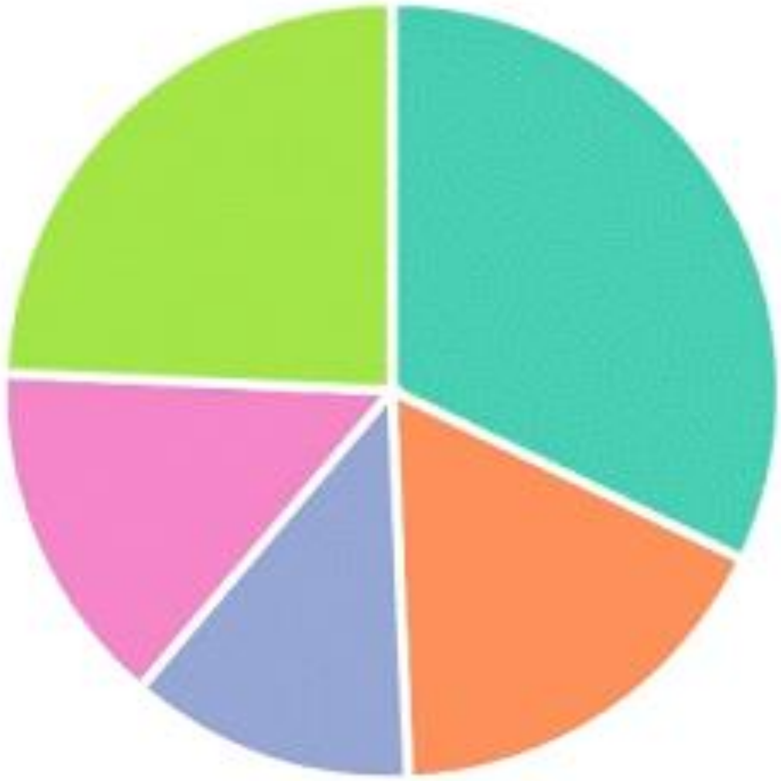
outlier





Pie Chart

Definition, Usage, Tips & Tricks



Definition:
used to describe the composition between parts of a unified whole. This part is usually represented in percent so that if all the parts are added up, the result equals one hundred percent.

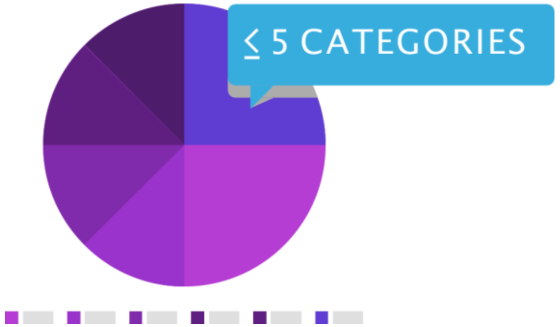
Visual Dimensions:
Proportion/Percentage, Category, Color

Usage:
Percentage of categories in a data

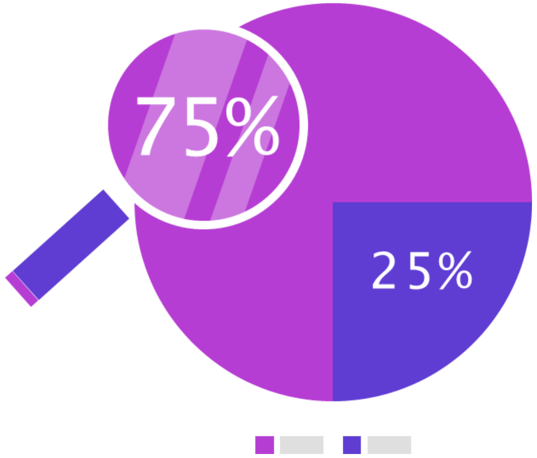
Pie Chart

Definition, Usage, Tips & Tricks

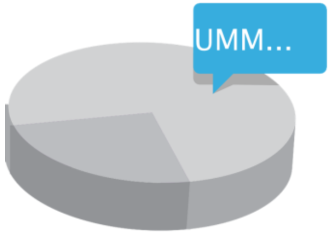
Less is more.
No more than 5 categories

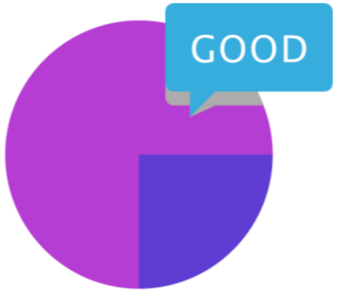


Clearly label percentages to avoid misinterpretation of the segment sizes

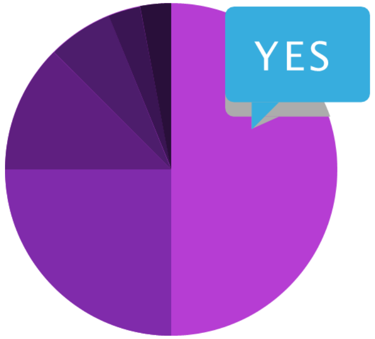


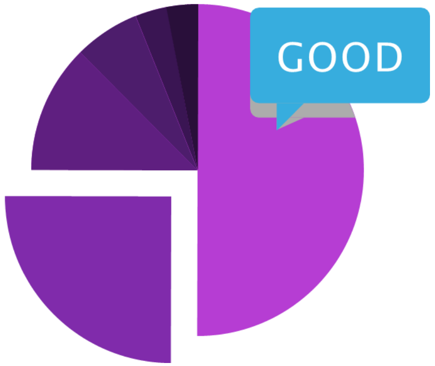
Avoid the use of **3D pie charts**, they make the data more difficult to understand





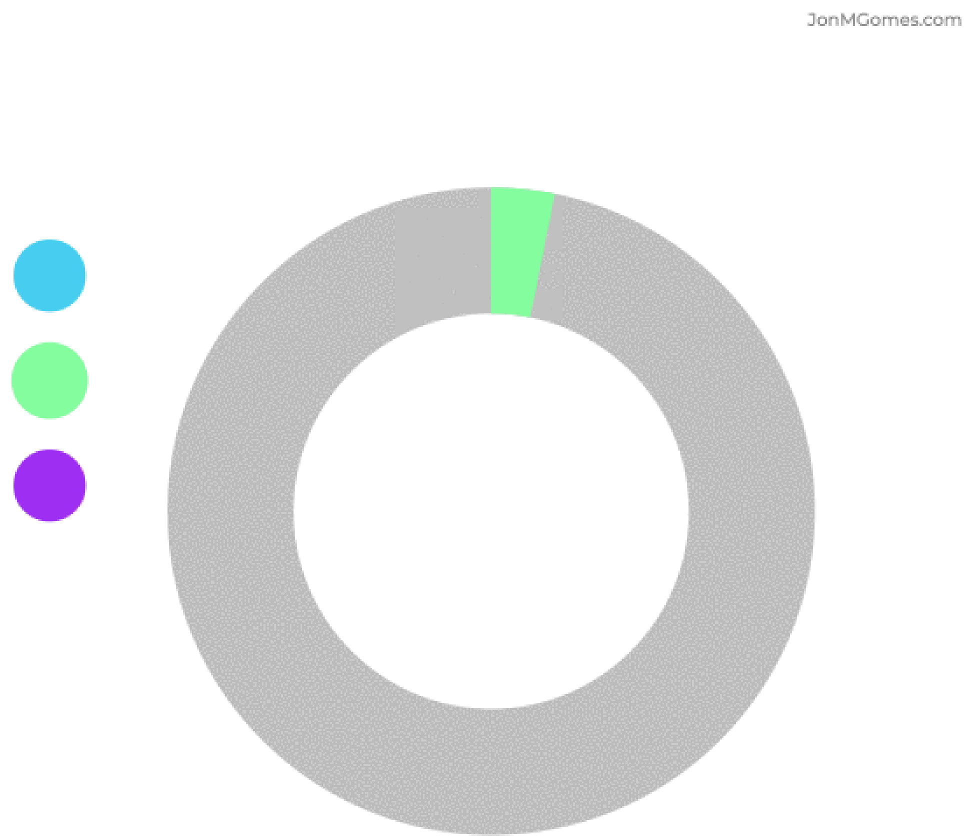
Order slices so that they are quickly understood





Donut Chart

Definition, Usage, Tips & Tricks



Definition:

This graph is another form of pie chart, its function also represents the proportion or composition between parts. The total number of parts was one hundred percent.

Because it looks simpler, this graph is also often modified into a semicircle

Visual Dimensions:

Proportion/Percentage, Category, Color

Usage:

Percentage of categories in a data

Text & Number

Definition, Usage, Tips & Tricks

Definition:
Data does not have to be presented in graphical form. Can use text and numbers only, with a note that only 1-2 data you want to display. Give bold or color to the number or text that you want to highlight so that the reader's attention is focused on that part.

Visual Dimensions:
text

Usage:
Summarizing data

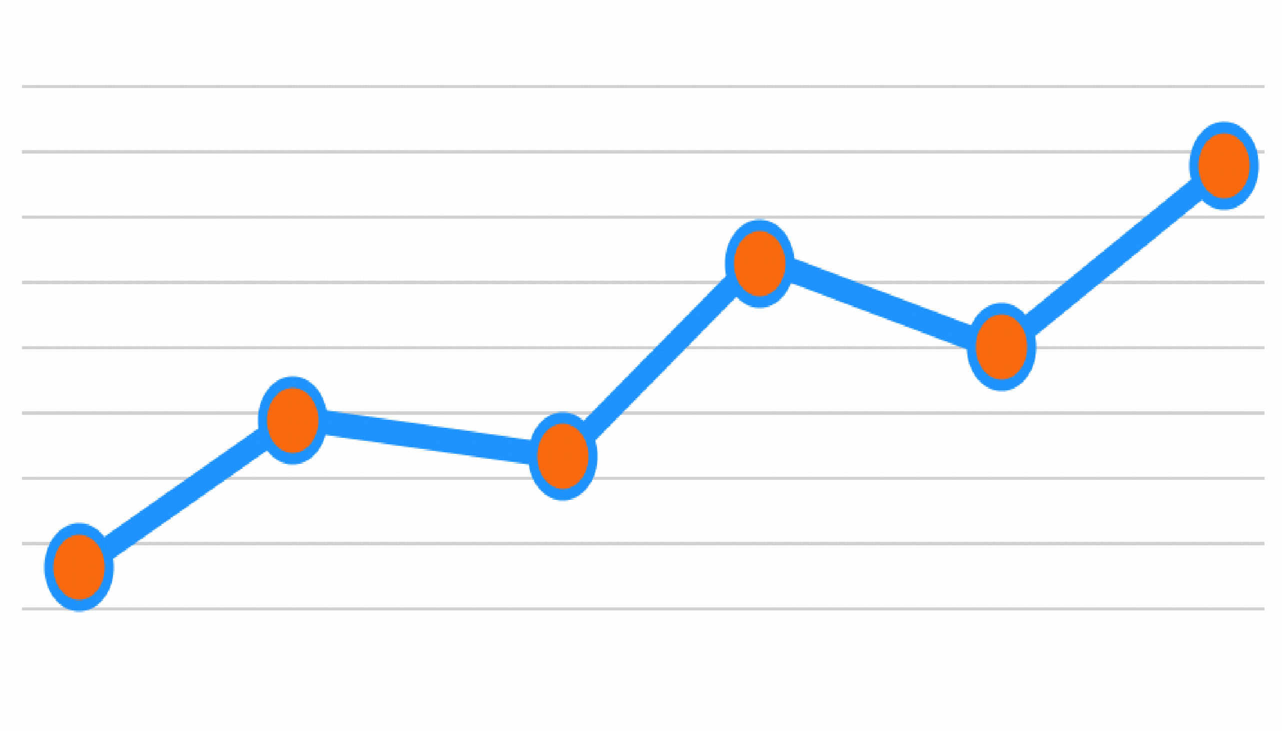
Tips & Tricks:

- Clear text

67%
responden
setuju

Line Chart

Definition, Usage, Tips & Tricks



Definition:

a type of chart which displays information as a series of data points called 'markers' connected by straight line segments. The X axis usually represents the time period, the Y axis represents the value/quantity.

Visual Dimensions:

Length, Series of time, Line

Usage:

Time series Data

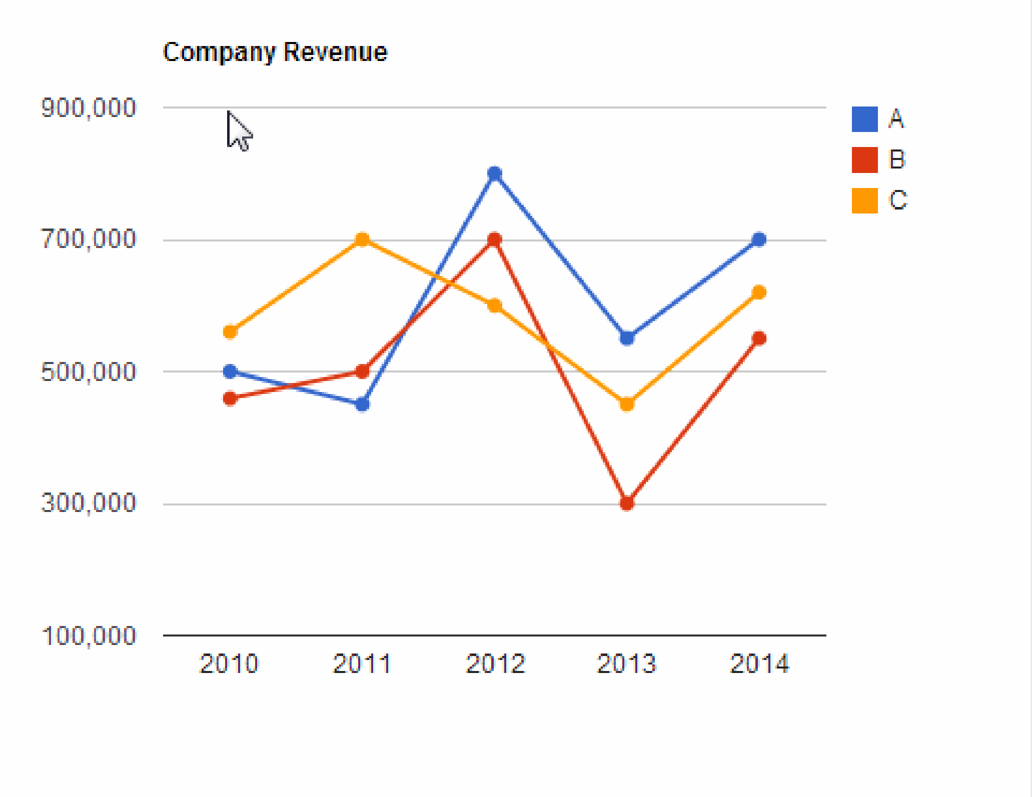
MultiLine Chart

Definition, Usage, Tips & Tricks

Definition:
is a basic line chart with one or more additional lines that represent comparison trends.

Visual Dimensions:
Length, Series of time, Line, color

Usage:
Comparison Time series Data

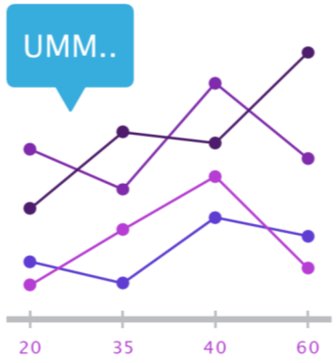


MultiLine Chart

Definition, Usage, Tips & Tricks

Use a maximum of **4 lines** when comparing

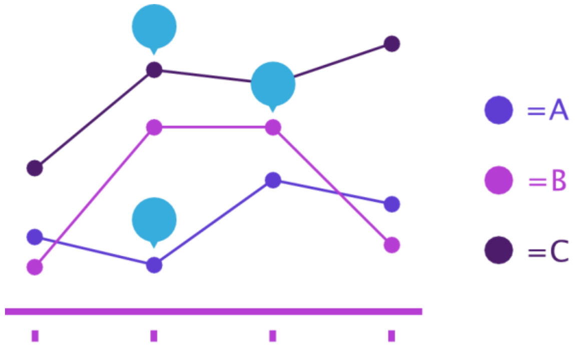
Use as **few lines** as possible



GOOD
20
40
60
80

Use **solid** line instead

Label each line separately



Area Chart

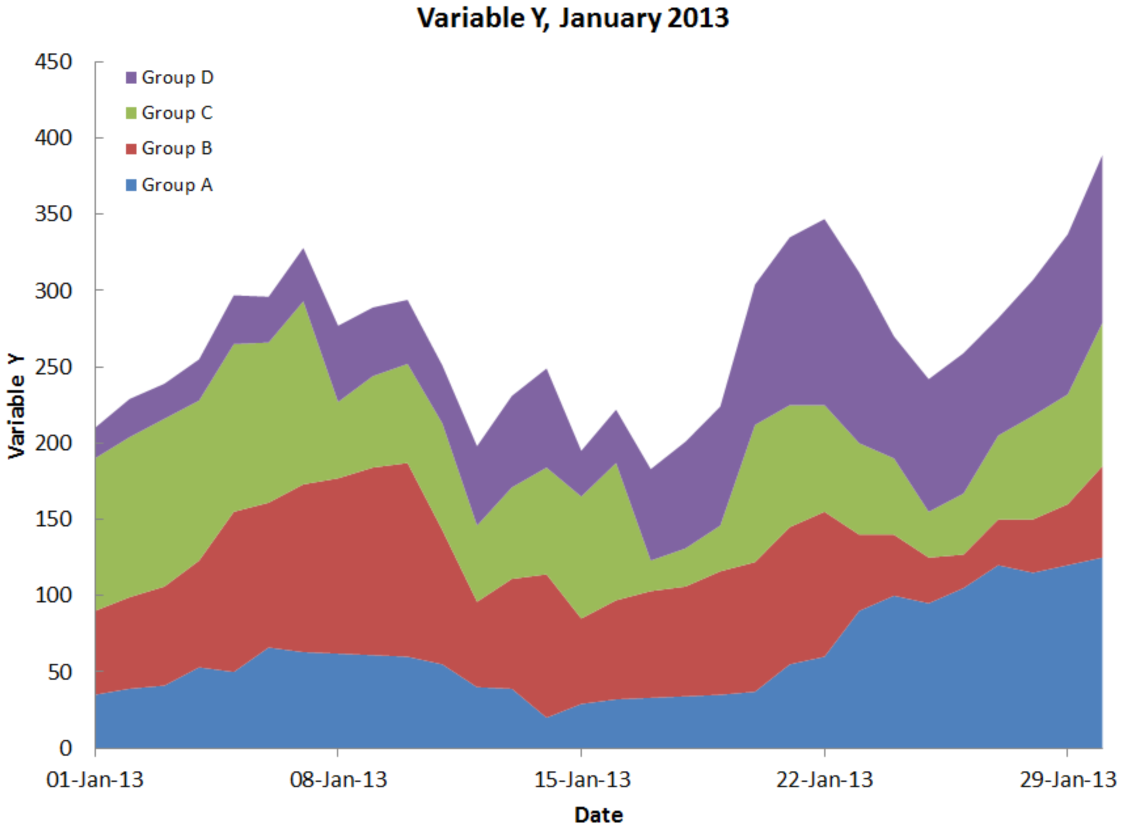
Definition, Usage, Tips & Tricks

Definition:
displays graphically quantitative data. It is based on the line chart. The area between axis and line are commonly emphasized with colors, textures and hatchings.

Visual Dimensions:
Length, Category, Area, Color

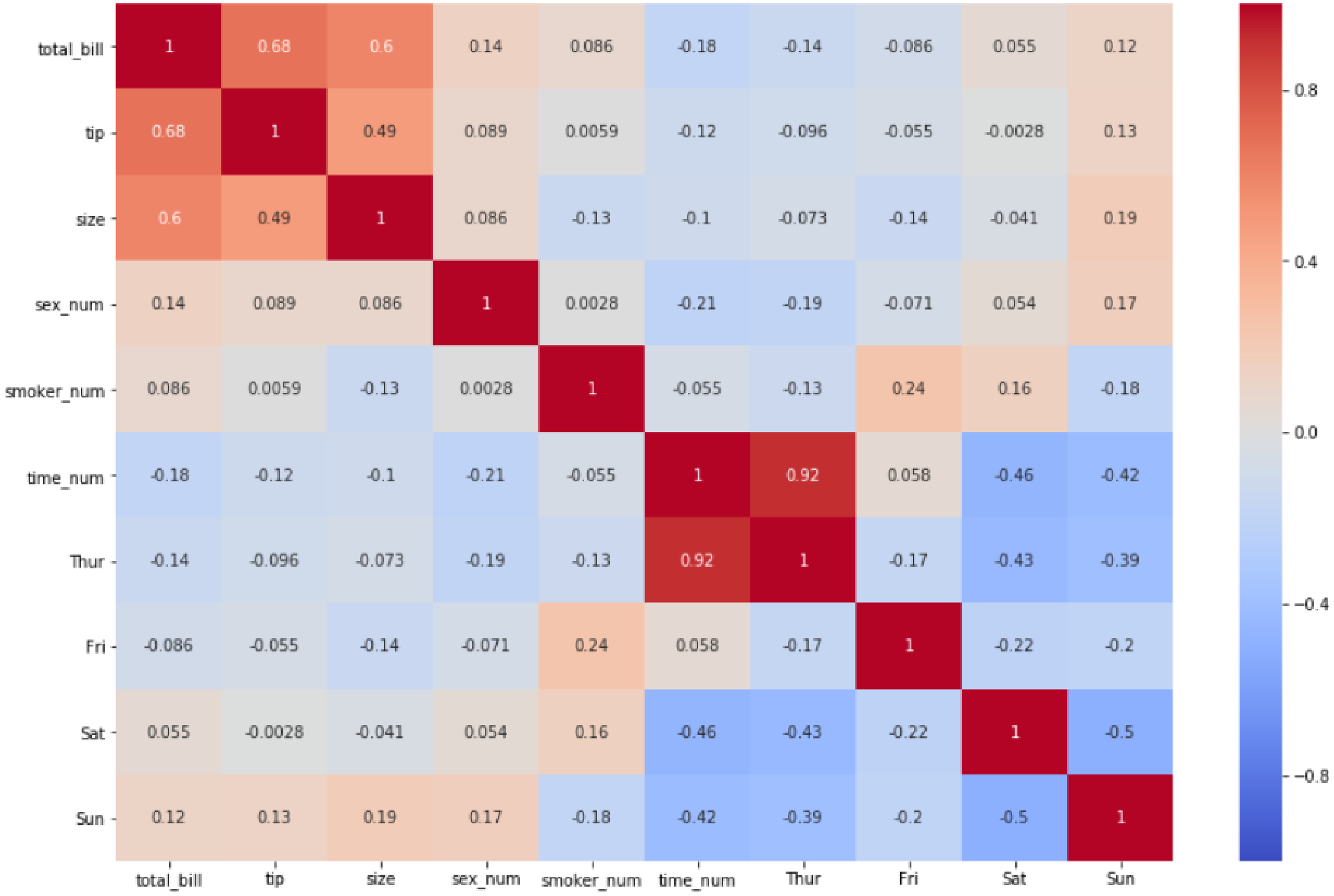
Usage:
used to illustrate total values in numbers or percentages over time

Tips & Tricks:
Don't let any area cover other areas.



Heat Map

Definition, Usage, Tips & Tricks



	total_bill	tip	size	sex_num	smoker_num	time_num	Thur	Fri	Sat	Sun
total_bill	1	0.68	0.6	0.14	0.086	-0.18	-0.14	-0.086	0.055	0.12
tip	0.68	1	0.49	0.089	0.0059	-0.12	-0.096	-0.055	-0.0028	0.13
size	0.6	0.49	1	0.086	-0.13	-0.1	-0.073	-0.14	-0.041	0.19
sex_num	0.14	0.089	0.086	1	0.0028	-0.21	-0.19	-0.071	0.054	0.17
smoker_num	0.086	0.0059	-0.13	0.0028	1	-0.055	-0.13	0.24	0.16	-0.18
time_num	-0.18	-0.12	-0.1	-0.21	-0.055	1	0.92	0.058	-0.46	-0.42
Thur	-0.14	-0.096	-0.073	-0.19	-0.13	0.92	1	-0.17	-0.43	-0.39
Fri	-0.086	-0.055	-0.14	-0.071	0.24	0.058	-0.17	1	-0.22	-0.2
Sat	0.055	-0.0028	-0.041	0.054	0.16	-0.46	-0.43	-0.22	1	-0.5
Sun	0.12	0.13	0.19	0.17	-0.18	-0.42	-0.39	-0.2	-0.5	1

Definition:
to show relationships between two variables, one plotted on each axis. By observing how cell colors change across each axis, you can observe if there are any patterns in value for one or both variables

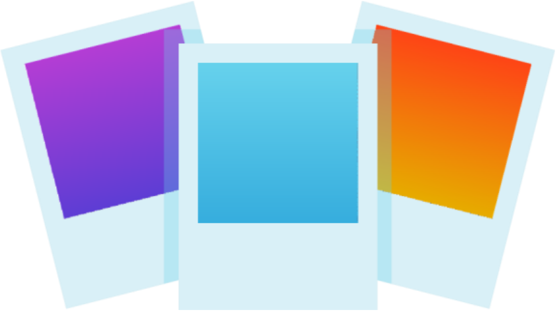
Visual Dimensions:
Color, Variables

Usage:
show relationships between two variables

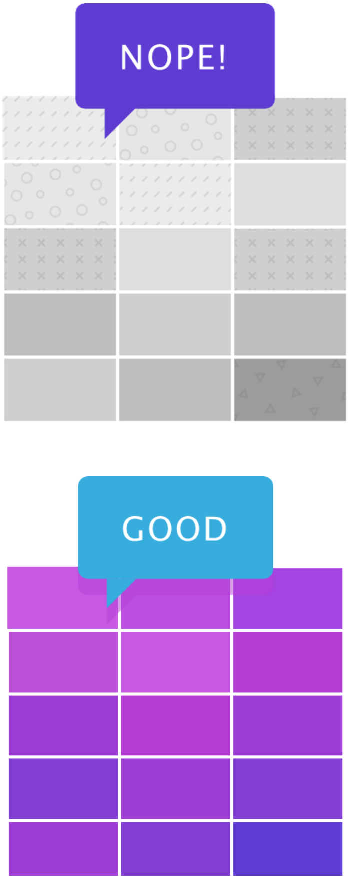
Heat Map

Definition, Usage, Tips & Tricks

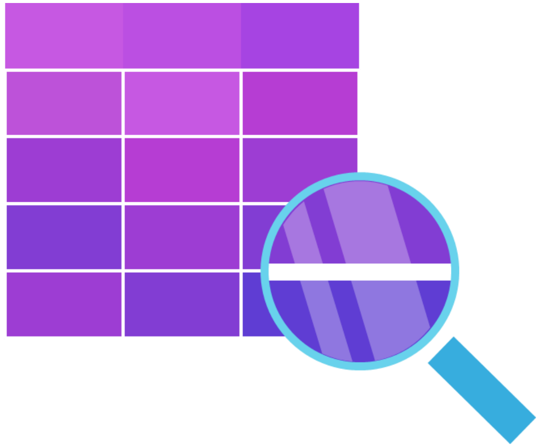
Use **Simple color** gradients



Keep patterns to a **minimal**

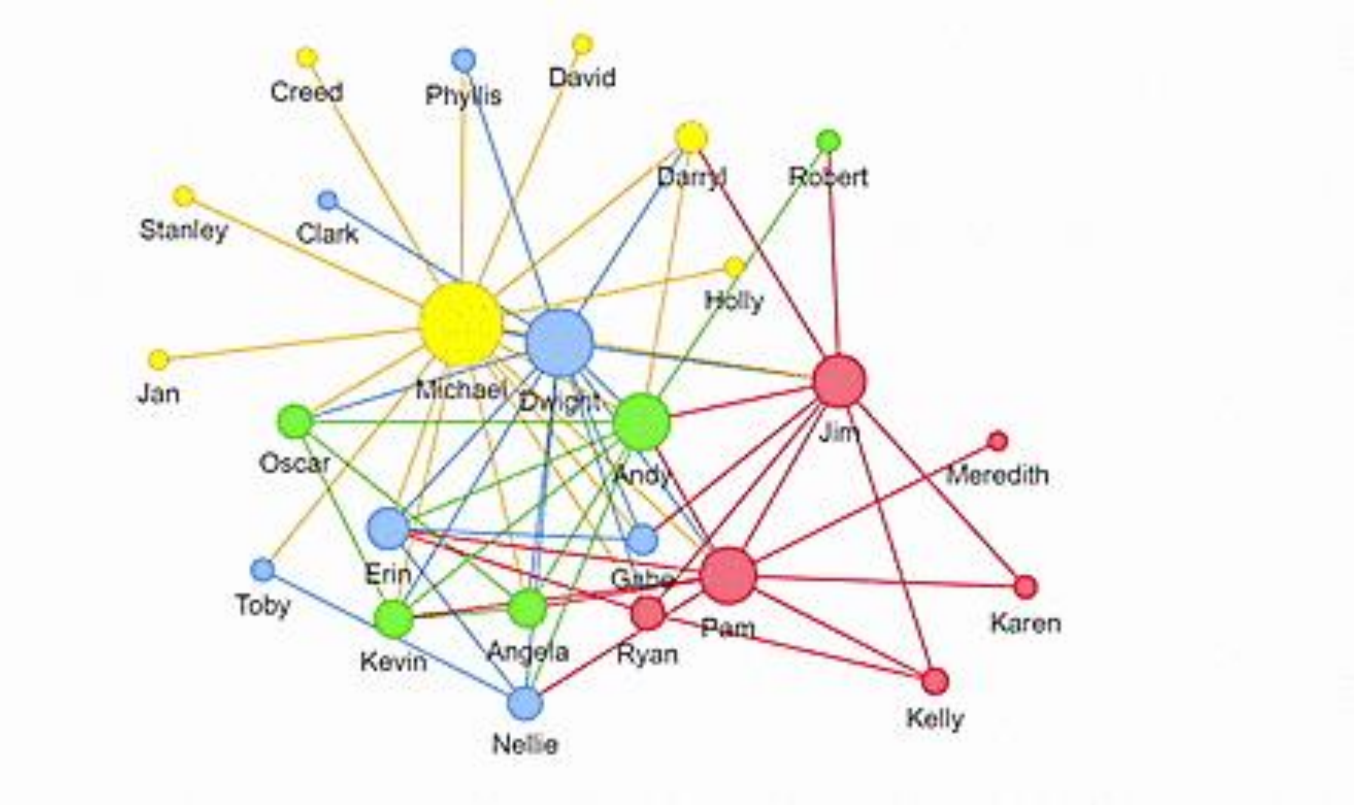


Use **Clear** map **boundaries**



Social Network

Definition, Usage, Tips & Tricks



Definition:
A social network diagram visually displays the relationships and interactions between people, groups, computers and other information entities. It maps out the nodes (individuals or groups) and the links (relationships or interactions) that connect them.

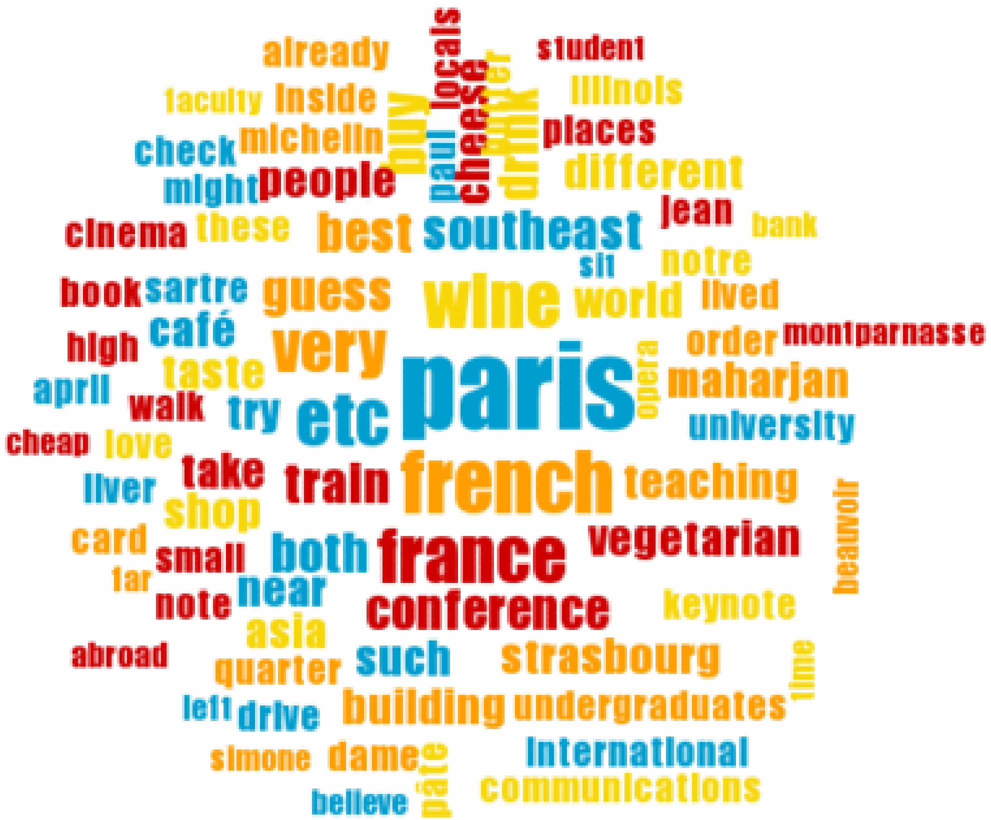
Visual Dimensions:
Dot, size, line

Usage:
Transaction of money, social media interaction

Word Cloud

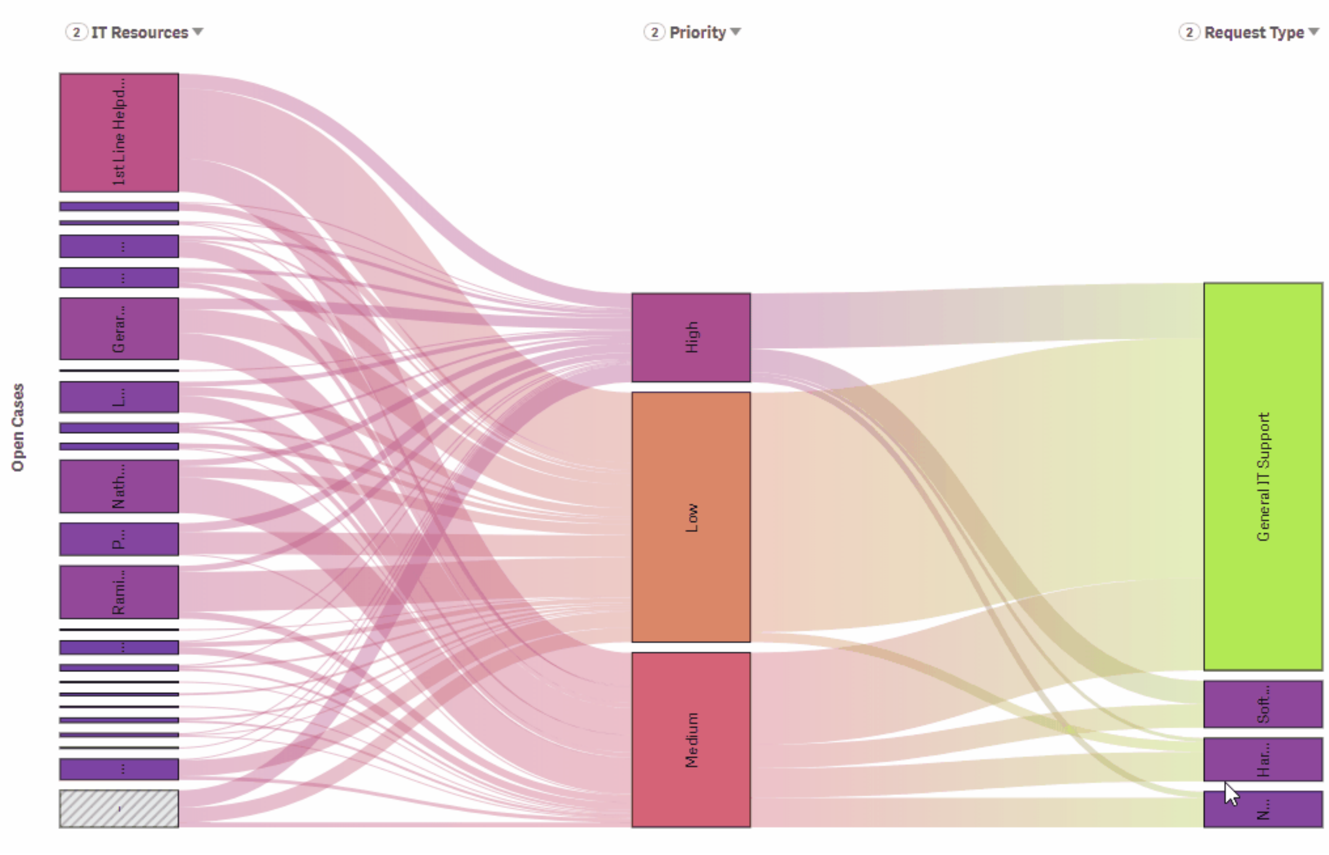
Definition, Usage, Tips & Tricks

Definition:
can be used to highlight popular values or show the frequency of text data using font size and color. In a word cloud chart, more prominent values are displayed with a larger font size than the less prominent values.
Visual Dimensions:
Text, size
Usage:
Popular topic in social media or text
Tips & Tricks:
Filter unnecessary word, prefix, etc.



Sankey Chart

Definition, Usage, Tips & Tricks



Definition:

a visualization used to depict a flow from one set of values to another. The things being connected are called nodes and the connections are called links.

Visual Dimensions:

Nodes, link

Usage:

- **a many-to-many mapping between two domains**
- multiple paths through a set of stages
- (for instance, Google Analytics uses sankeys to show how traffic flows from pages to other pages on your web site).

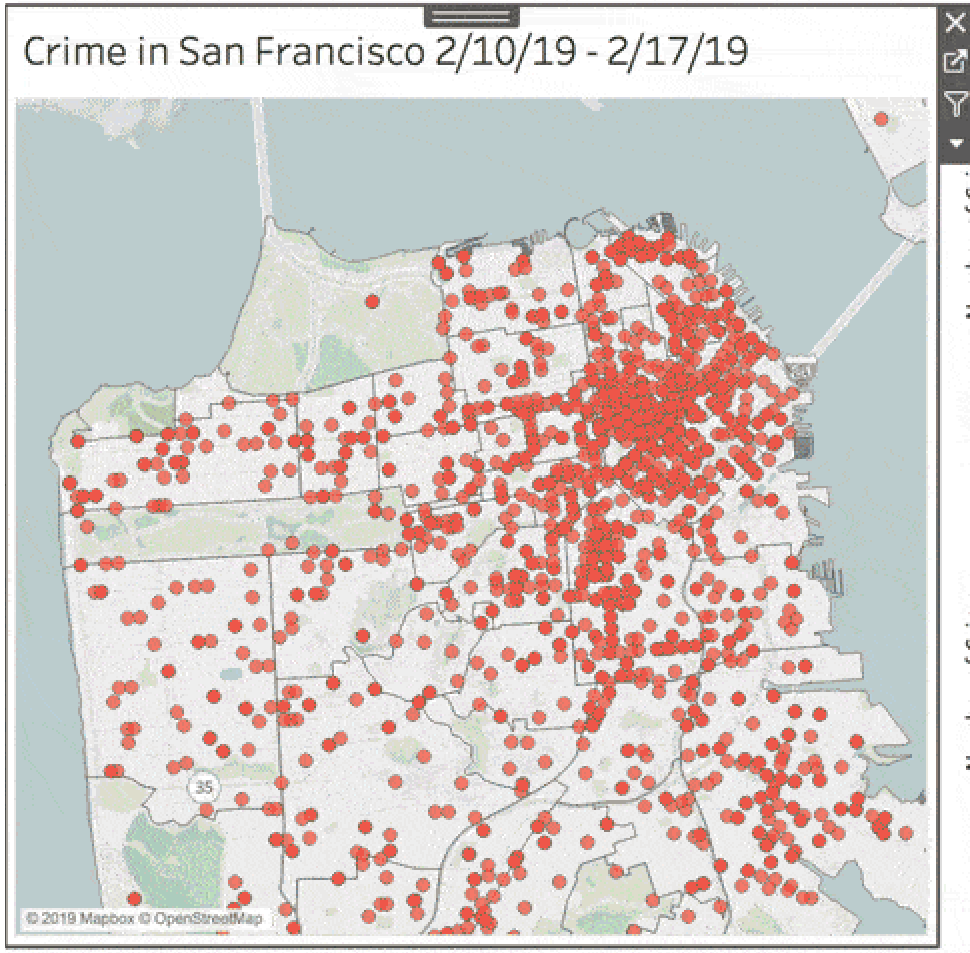
Map Chart

Definition, Usage, Tips & Tricks

Definition:
Map charts allow you to position your data in a context, often geographical, using different layers. The layers can be either data layers, such as marker layers or feature layers, or reference layers such as map layers.

Visual Dimensions:
marker, map, data

Usage:
Knowing characteristic data in selected region


Crime in San Francisco 2/10/19 - 2/17/19
35
© 2019 Mapbox © OpenStreetMap

Tools for Data Visualization

Type of Tools Data Visualization

Definition, Usage, Tips & Tricks

CODE BASED

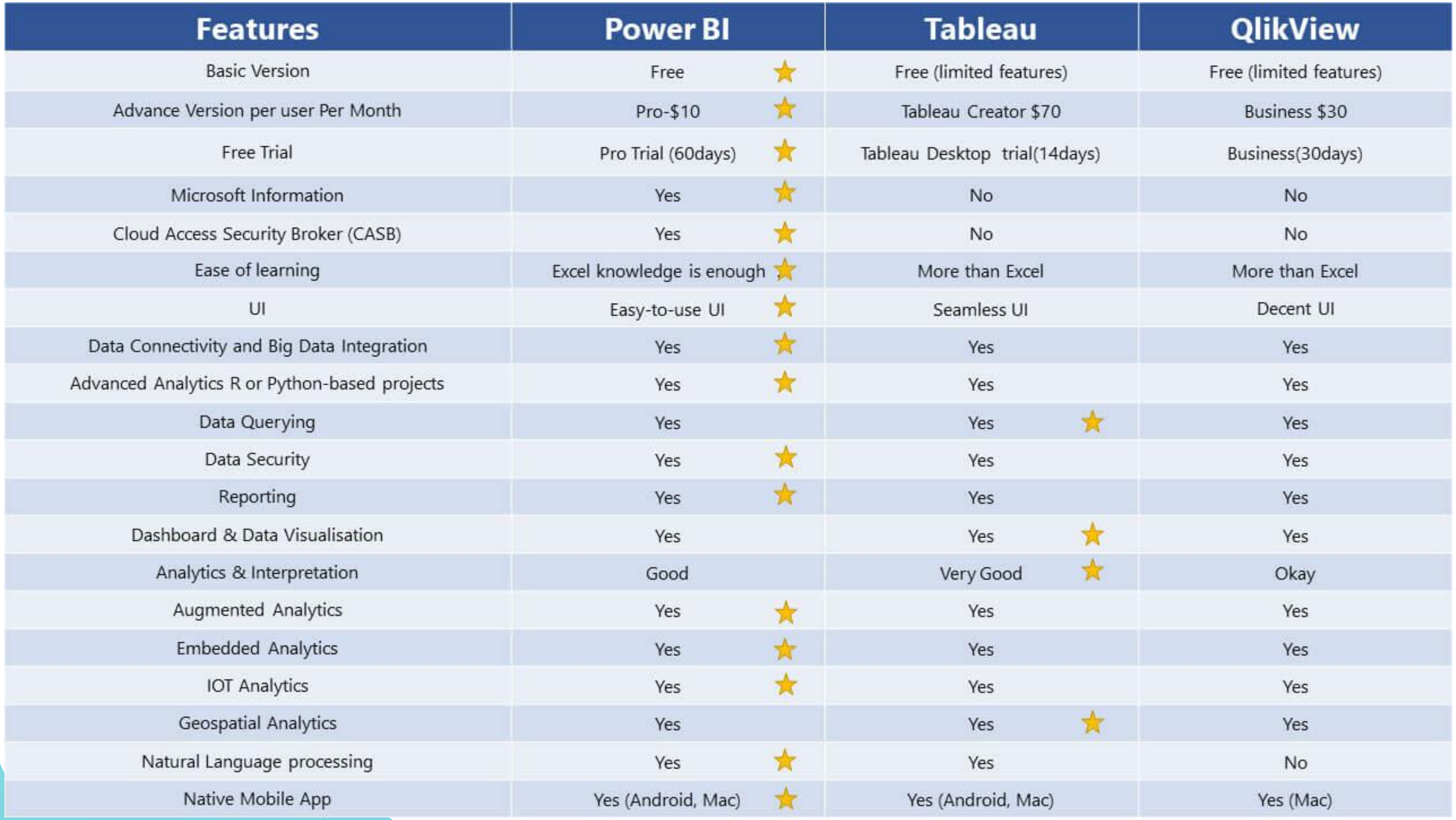
- Chart.js
- R
- D3
- python™

GUI BASED

- tableau
- Google Data Studio
- Qlik
- Power BI
- Gephi makes graphs handy
- redash

Power BI, Tableau & QlikView

Comparison based on Combay Consultant



Features	Power BI	Tableau	QlikView
Basic Version	Free ★	Free (limited features)	Free (limited features)
Advance Version per user Per Month	Pro-$10 ★	Tableau Creator $70	Business $30
Free Trial	Pro Trial (60days) ★	Tableau Desktop trial(14days)	Business(30days)
Microsoft Information	Yes ★	No	No
Cloud Access Security Broker (CASB)	Yes ★	No	No
Ease of learning	Excel knowledge is enough ★	More than Excel	More than Excel
UI	Easy-to-use UI ★	Seamless UI	Decent UI
Data Connectivity and Big Data Integration	Yes ★	Yes	Yes
Advanced Analytics R or Python-based projects	Yes ★	Yes	Yes
Data Querying	Yes	Yes ★	Yes
Data Security	Yes ★	Yes	Yes
Reporting	Yes ★	Yes	Yes
Dashboard & Data Visualisation	Yes	Yes ★	Yes
Analytics & Interpretation	Good	Very Good ★	Okay
Augmented Analytics	Yes ★	Yes	Yes
Embedded Analytics	Yes ★	Yes	Yes
IOT Analytics	Yes ★	Yes	Yes
Geospatial Analytics	Yes	Yes ★	Yes
Natural Language processing	Yes ★	Yes	No
Native Mobile App	Yes (Android, Mac) ★	Yes (Android, Mac)	Yes (Mac)

Libraries in Python

Common used Libraries in Python about Data Visualization

matplotlib

Pros:

- Easy to see the property of the data
- Can Plot anything

Cons:

- may be complex to plot non-basic plots

seaborn

Pros:

- Less code
- Make common-used plots prettier

Cons:

- more constrained and does not have as wide a collection as matplotlib

plotly

Pros:

- gives you the same quality plots like in R
- Easy to create interactive plots
- Complex plots made easy

Cons:

- Not suitable for static Report

Folium

Pros:

- Easy to create a map with markers
- Add potential location
- Plugins

Cons:

- Not so good Google Maps

Data Preparation & Manipulation

Introduction to Pandas

Definition, Usage, Type

Pandas is a Python library used for working with data sets.
It has functions for **analyzing**, **cleaning**, **exploring**, and **manipulating data**.
The name "Pandas" has a reference to both "Panel Data", and "Python Data Analysis" and was created by Wes McKinney in 2008.

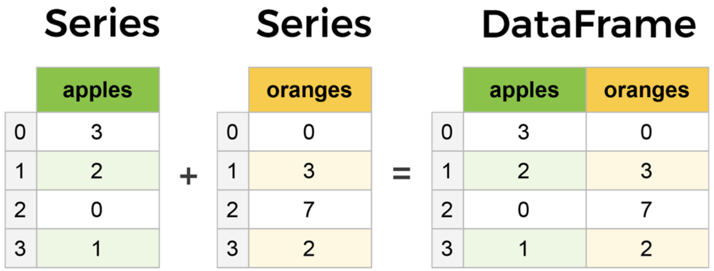
There are two types of data structures in pandas:

- Series

A pandas Series is a one-dimensional data structure (“a one-dimensional ndarray”) that can store values — and for every value, it holds a unique index, too.

- DataFrames

a two (or more) dimensional data structure – basically a table with rows and columns. The columns have names and the rows have indexes.



Series

	apples
0	3
1	2
2	0
3	1

+

Series

	oranges
0	0
1	3
2	7
3	2

=

DataFrame

	apples	oranges
0	3	0
1	2	3
2	0	7
3	1	2

Reading Data using Pandas

Pandas functions for reading the contents of files are named using the pattern .**read_<file-type>()**, where **<file-type>** indicates the type of the file to read.

- CSV FILES

A CSV (comma-separated values) file is a text file that has a specific format which allows data to be saved in a table structured format.

```
# reading csv file
data =  pd.read_csv(‘data.csv’, sep=“,”)
```

- EXCEL FILE

Xlsx extension is used for files saved as Microsoft Excel worksheets.

```
# reading Excel file
data =  pd.read_excel(‘data.xlsx’, sheet_name=“sheet_name”)
```

- TABLE FROM DATABASE

A SQL database is a collection of tables that stores a specific set of structured data

```
# reading Table Database
data = pd.read_sql(‘table_data', 'postgres:///db_name')
```

- JSON FILE

JSON stands for JavaScript Object Notation. JSON is a lightweight format for storing and transporting data.

```
# reading Json File
data = pd.read_json(‘files/sample_file.json’, orient=“index”)
```

- PICKLE FILE

Python pickle files are the binary files that keep the data and hierarchy of Python objects. They usually have the extension .pickle or .pkl.

```
# reading Pickle File
data = pd.read_pickle(“./dummy.pkl”)
```

More info here

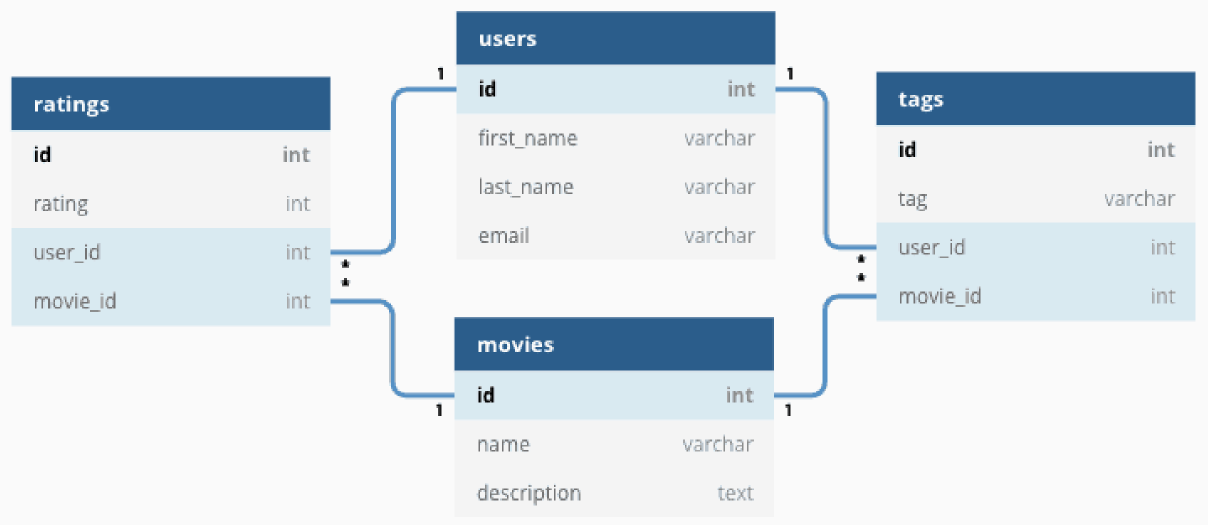
Introduction SQL

Definition, RDBMS

- Structured Query Language (SQL) is a programming language that is typically used in relational database management systems (RDBMS).
- We use SQL to be able to communicate with databases directly.
- It is capable to perform tasks such as creating, reading, updating, and deleting tables in a database.

RDBMS (Relational Database Management System)

- A relational database refers to a database that stores data in a structured format, using rows and columns. This makes it easy to locate and access specific values within the database. It is "relational" because the values within each table are related to each other.



ratings	
id	int
rating	int
user_id	int
movie_id	int

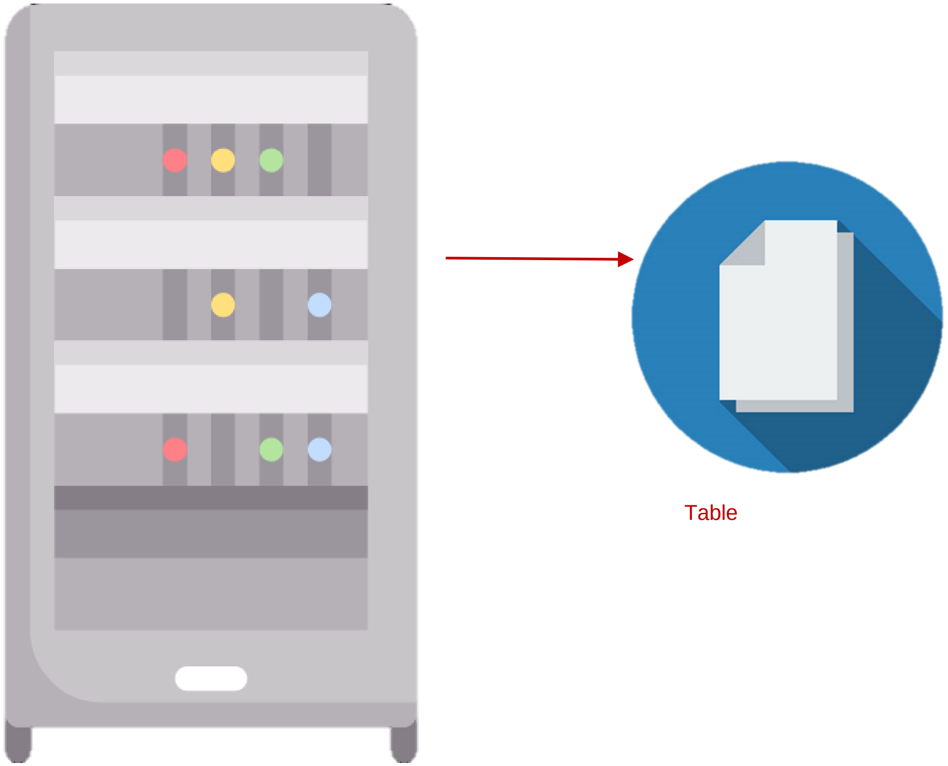
users	
id	int
first_name	varchar
last_name	varchar
email	varchar

movies	
id	int
name	varchar
description	text

tags	
id	int
tag	varchar
user_id	int
movie_id	int

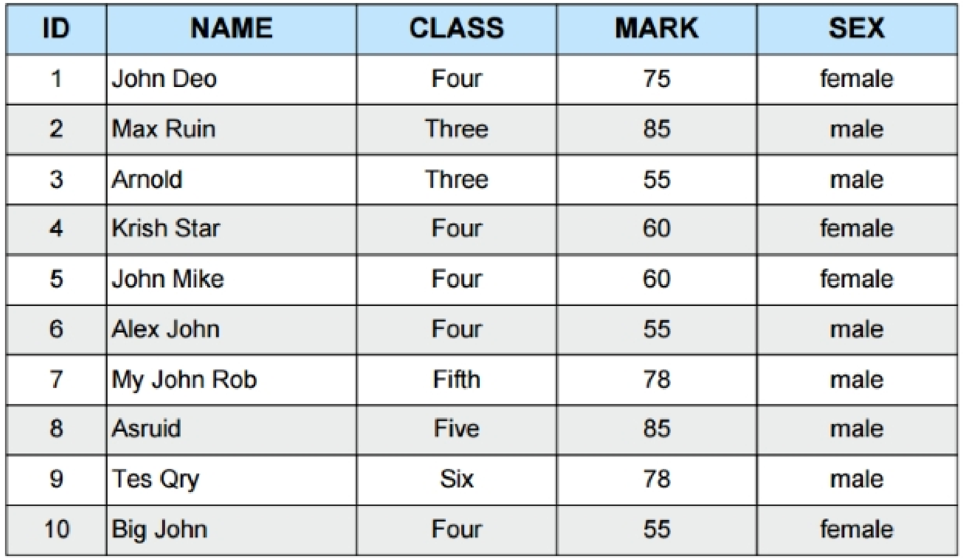
Table & Database

A **table** is a collection of related data held in a table format within a database.
A **Schema/database** consist of many tables.



Schema/Database

Table



ID	NAME	CLASS	MARK	SEX
1	John Deo	Four	75	female
2	Max Ruin	Three	85	male
3	Arnold	Three	55	male
4	Krish Star	Four	60	female
5	John Mike	Four	60	female
6	Alex John	Four	55	male
7	My John Rob	Fifth	78	male
8	Asruid	Five	85	male
9	Tes Qry	Six	78	male
10	Big John	Four	55	female

Command in SQL

DDL, DML, DCL

Data Definition Language (DDL)

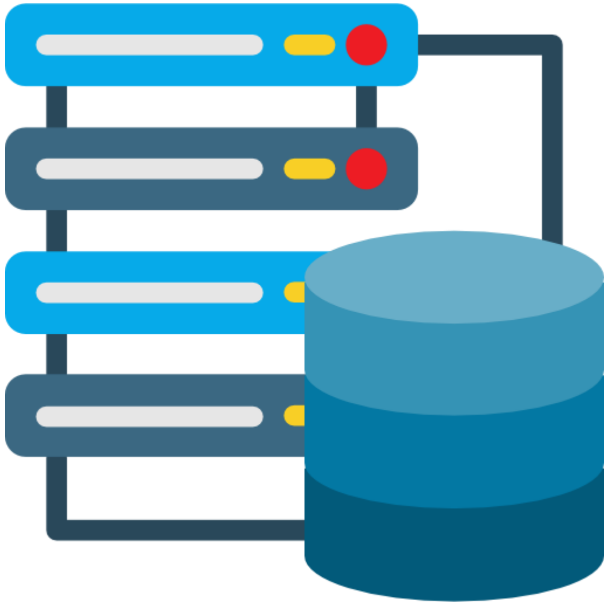
Actually consists of the SQL commands that can be used to define the database schema. It simply deals with descriptions of the database schema and is used to create and modify the structure of database objects in the database.

Ex: Create, Drop, Alter, Truncate

Data Manipulation Language (DML)

The SQL commands that deals with the manipulation of data present in the database belong to DML or Data Manipulation Language and this includes most of the SQL statements.

Ex: Select, Insert, Delete, Update



Data Control Language (DCL)

which includes commands such as GRANT and mostly concerned with rights, permissions and other controls of the database system.

Ex: Grant, Revoke

Materi:
s.id/1bUrY

Thank You

Ikan hiu makan kecap
See you at next meet up !

DQLab