EDA & Data Visualization

Siti Aminah*, Dinial Utami

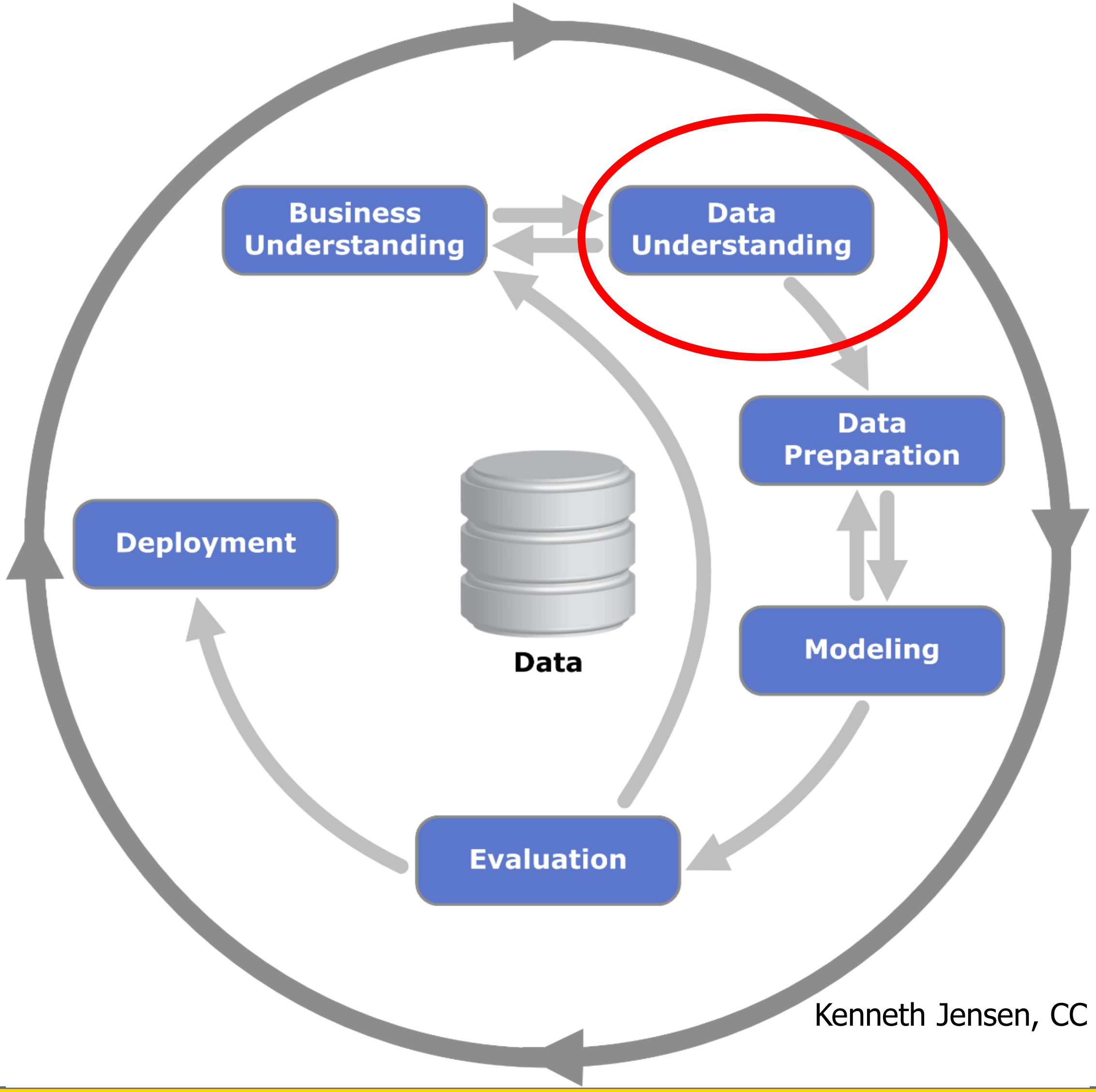
CSGE603130 - Kecerdasan Artifisial dan Sains Data Dasar
Semester Genap 2022/2023

UNIVERSITAS INDONESIA
Veritas, Probitas, Iustitia
FACULTY OF COMPUTER SCIENCE

Outline

- Exploratory Data Analysis
- Descriptive Statistics
- Data Visualization
  - Data Visualization Principles
  - Basic Visualization Tools
  - Specialized Visualization Tools
  - Advanced Visualization Tools

CRISP-DM: Cross-industry standard process for data mining



Kenneth Jensen, CC BY-SA 3.0, via Wikimedia Commons

Exploratory Data Analysis

EDA: Take a peek at data

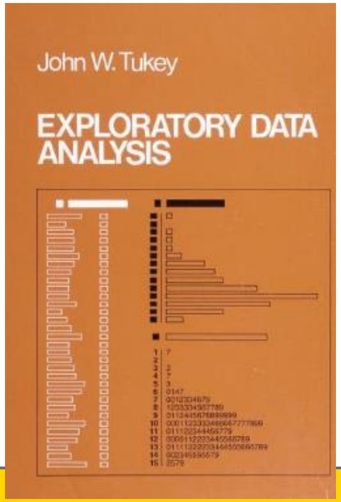
- EDA is a term for an initial analysis done with datasets.
- It's basically taking a peek at the data to understand more about what it represents and how to use it.
- It's often a precursor to more advanced data analytics techniques.

Exploratory data analysis (EDA)

Exploratory data analysis (EDA) is an approach:

- to **analyzing datasets**
- by **summarizing their main characteristics**
- often with **visual methods**.

The term EDA was coined by John W. Tukey in the book "Exploratory Data Analysis" in **1977.**



Why EDA?

- We need to familiarize with a new dataset: How does it look like?
  - How many attributes, and of what kind?
  - Are there any missing values?
  - How are the values distributed?
  - Is our dataset imbalanced? (= if left untreated, our model can be biased)

- Hunting for something interesting: What catches your eyes?
  - Are there any outliers?
  - Are there any correlations between attributes?
  - How do the distributions compare between different samples?

EDA approach

- **Descriptive statistics**
  - Central tendency
  - Measure of variation
  - Skewness & kurtosis
  - Correlations

- **Data visualizations**
  - Single attribute (univariate analysis):
    Barcharts, histogram, pie charts, donut charts
  - Multiple attributes (multivariate analysis):
    Scatter plots, bubble charts, line charts, heat maps

Descriptive Statistics

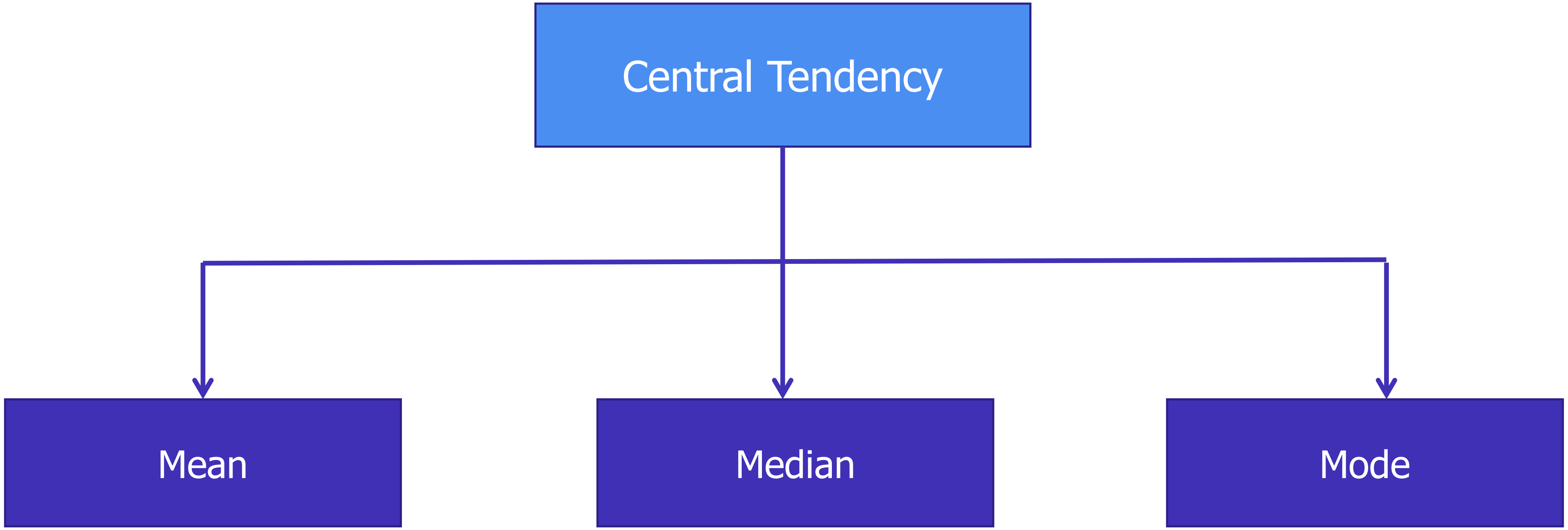
Kenali Data Anda

- Kategorikal vs Numerik

Kenali Data Anda

- Nominal, Ordinal, Interval, Rasio

Central Tendency



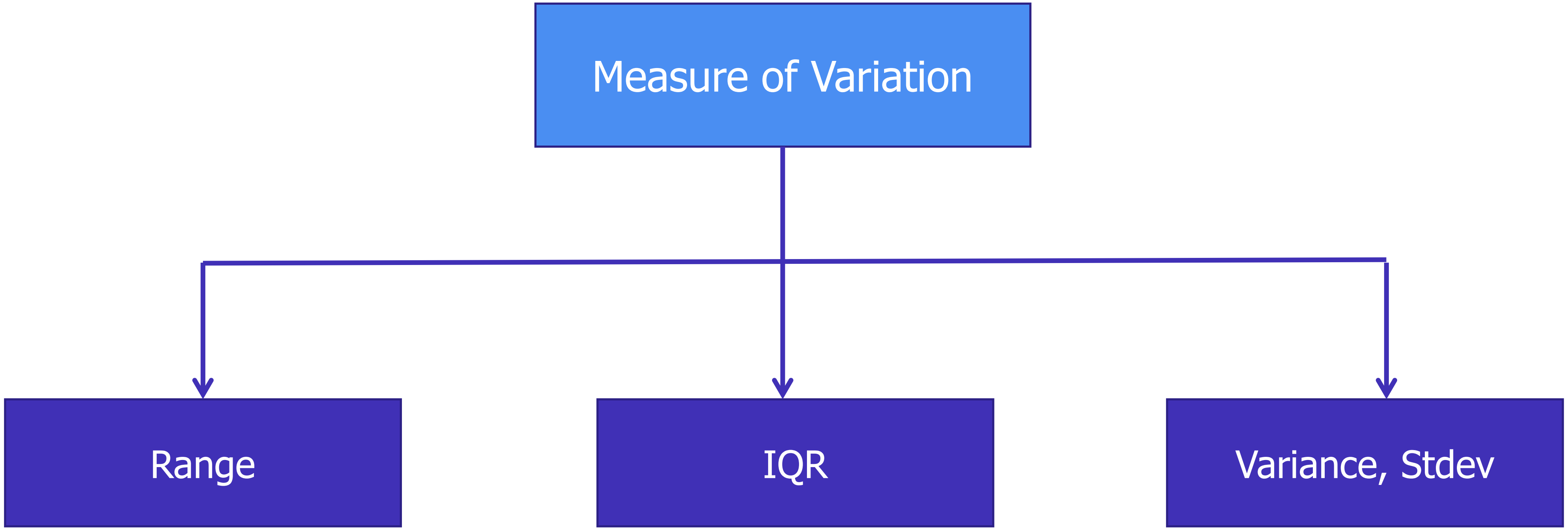
Central Tendency

Mean

Median

Mode

Measure of Variation



Measure of Variation

Range

IQR

Variance, Stdev

Measure of Variation

Range

Range = max - min

The simplest measure of variation, often denoted by indicating the largest and smallest values separately.

Measure of Variation

Inter-Quartile Range (IQR)

Divides a dataset into quartiles:

- Q1 (lower quartile): 25th percentile
  - Median of lower half
- Q2 (median): 50th percentile
- Q3 (upper quartile): 75th percentile
  - Median of upper half

IQR = Q3-Q1

Measure of Variation

Inter-Quartile Range (IQR)

- From the data (n = 7):
  5, 7, 4, 4, 6, 2, 8

- Q1 = ?
- Q2 = ?
- Q3 = ?

- **IQR = ?** **Range = ?**

Measure of Variation

Inter-Quartile Range (IQR)

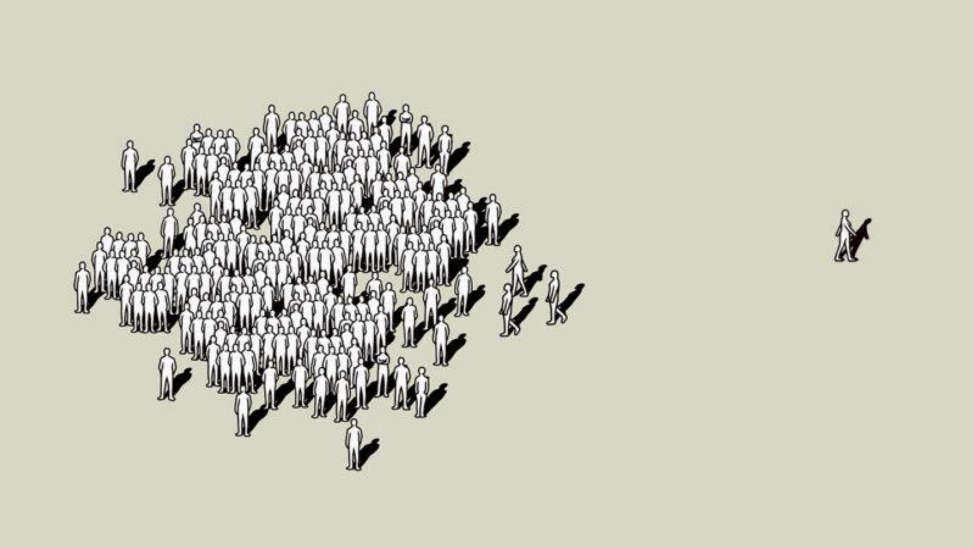
- From the data (n = 7):
  5, 7, 4, 4, 6, 2, 8 -> Sorted: **2, 4, 4, 5, 6, 7, 8**

- Q1 = median of lower half = 4
- Q2 = 5
- Q3 = median of upper half = 7

- **IQR = Q3-Q1 = 3 Range = 6**

Outliers

An outlier is an observation that lies an abnormal distance from other values in a random sample from a population.



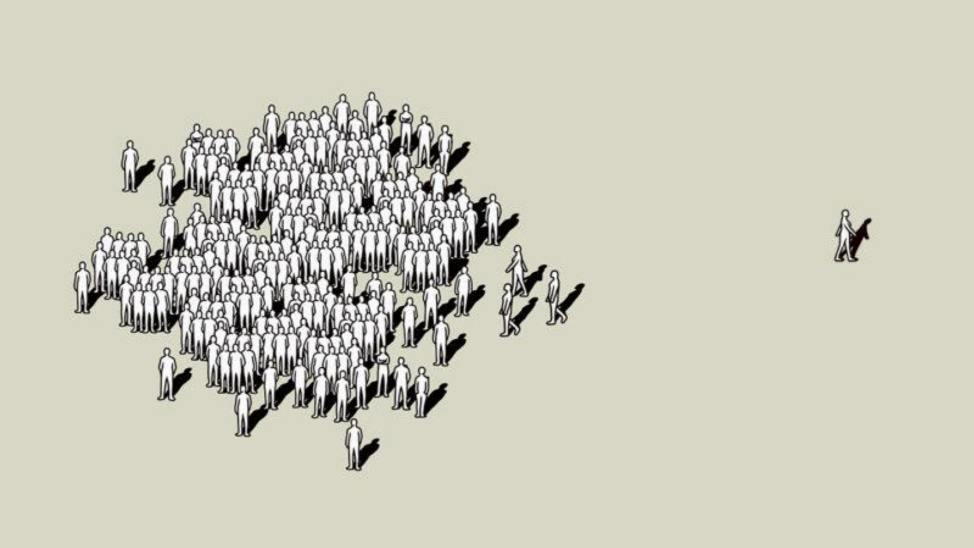
Before abnormal observations can be singled out, it is necessary to characterize normal observations.

<u>Outliers, according to IQR</u>, are data points whose values are:

- less than Q1-1.5*IQR, or
- more than Q3+1.5*IQR

Outliers

An outlier is an observation that lies an abnormal distance from other values in a random sample from a population.



Before abnormal observations can be singled out, it is necessary to characterize normal observations.

Outliers, according to IQR, are data points whose values are:

- less than Q1-1.5*IQR, or
- more than Q3+1.5*IQR

Kapan menggunakan rumus ini ? Untuk atribut yang karakteristiknya bagaimana?

Variance & Standard Deviation

Variance = Average of the squared deviation of the observations from the mean

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$$

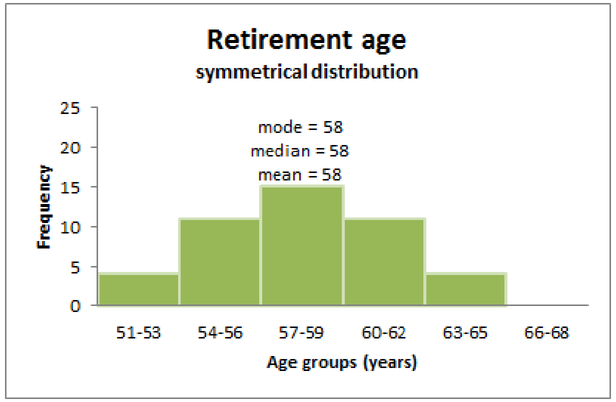
Standard deviation s = Square root of the variance

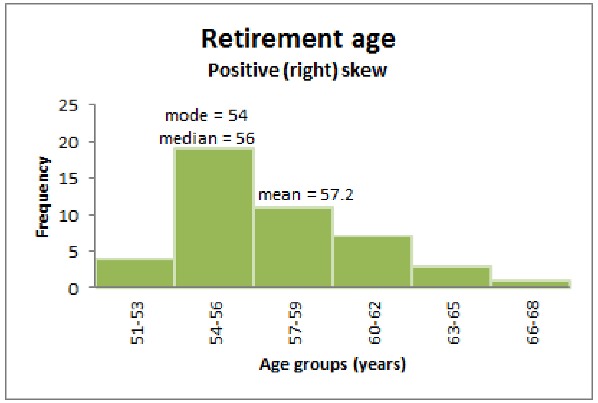
Skewness & Kurtosis

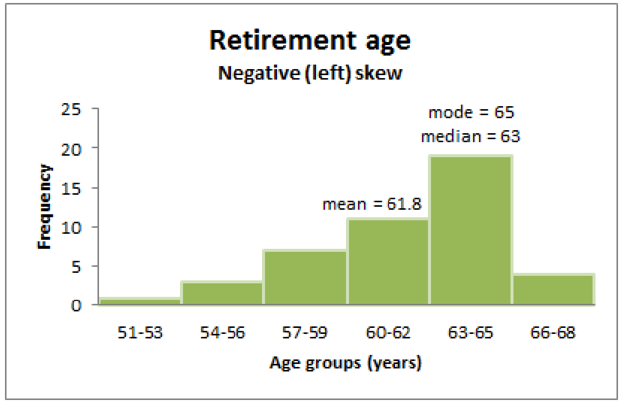
- **Skewness**
  A measure of asymmetry

- **Kurtosis**
  A measure of outliers

Skewness







- Skewness is a measure of asymmetry of the data around the mean.

- When a **distribution is skewed**, the mode remains the most commonly occurring value, the median remains the middle value in the distribution, but **the mean is generally ‘pulled’ in the direction of the tails**.

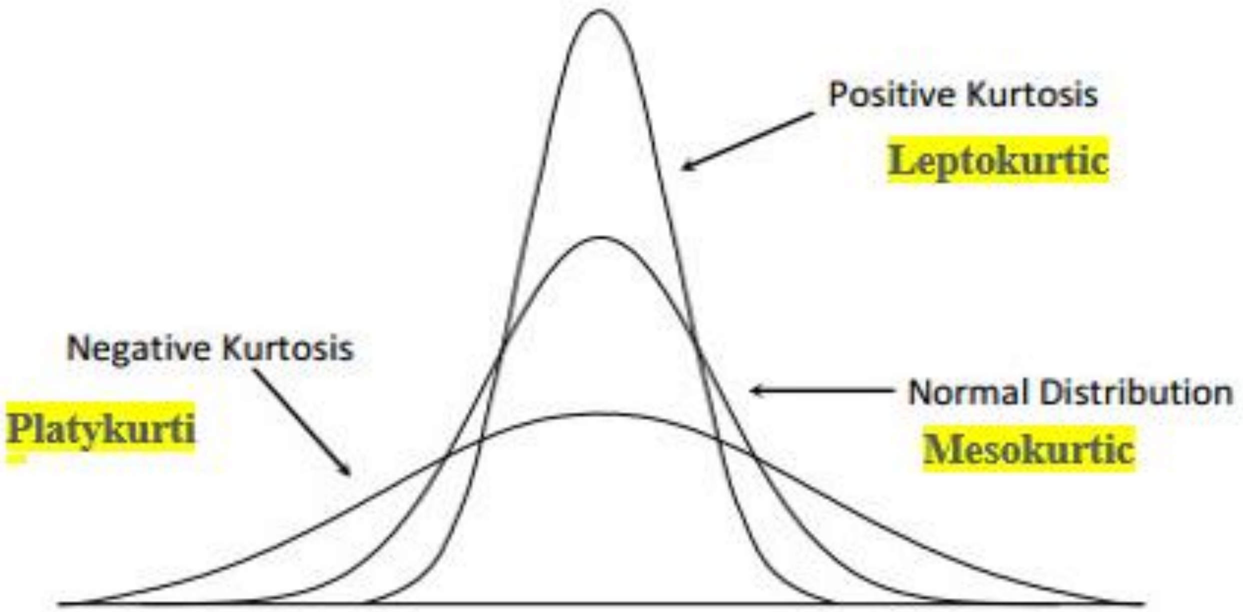
Skewness

$$\sum \frac{(x_i - \bar{x})^3}{ns^3}$$

- where x_1 is each data point, $\bar{x}$ is the arithmetic mean, n is the size of the data , and s is the standard deviation.

- The skewness for a normal distribution is zero, and any symmetric data should have skewness near zero. Negative values for the skewness indicate data that are skewed left and positive values for the skewness indicate data that are skewed right.

Kurtosis



- High kurtosis indicates the presence of outliers!

https://www.analyticsvidhya.com/blog/2021/05/shape-of-data-skewness-and-kurtosis/

Kurtosis

$$\text{Kurtosis} = \frac{\sum(x_i - \bar{x})^4}{nS^4}$$

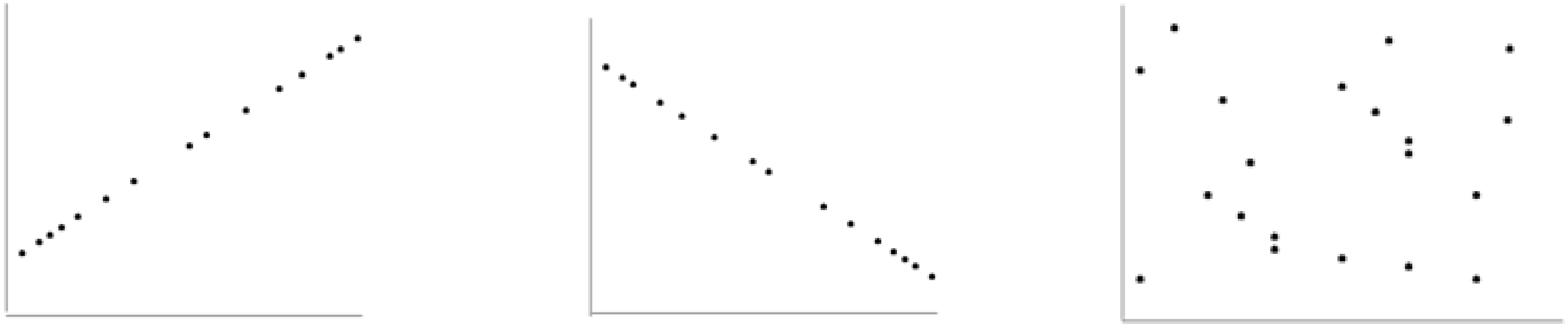
- where x_1 is each data point, $\bar{x}$ is the arithmetic mean, n is the size of the data , and s is the standard deviation.

- A normal distribution has kurtosis exactly 3 (mesokurtic).
- A distribution with kurtosis<3 is called platykurtic.
- A distribution with kurtosis>3 is called leptopkurtic

(Pearson) Correlation

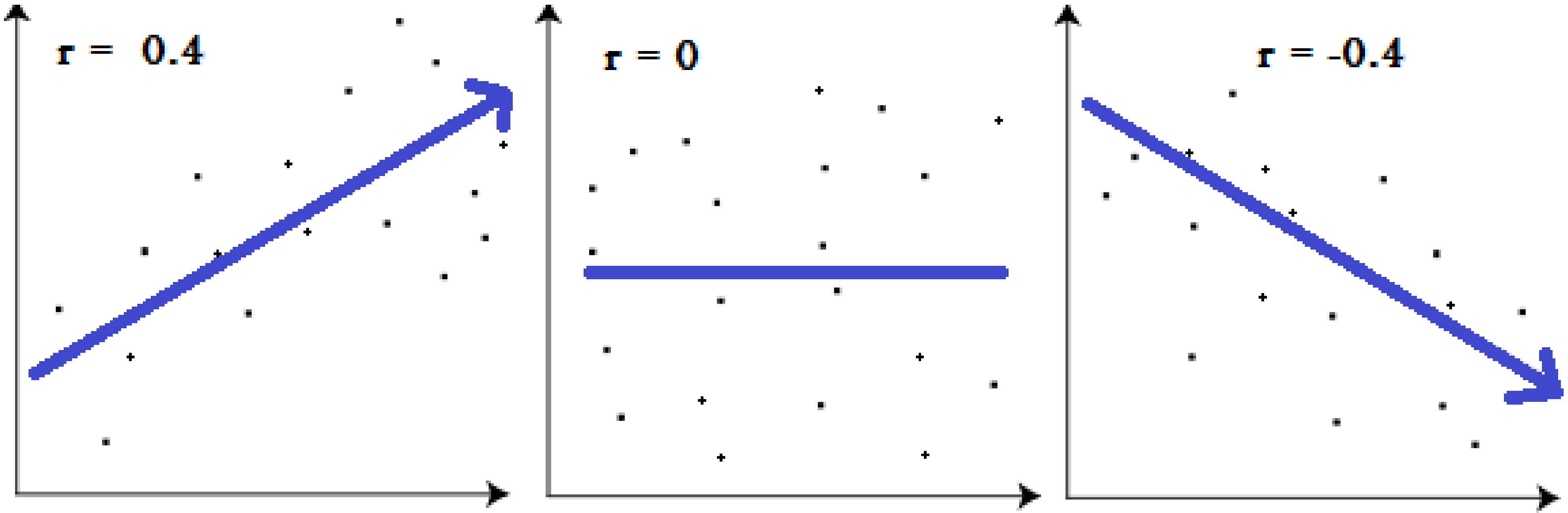
- It is a technique to investigate the relationship between two variables: that is, measures the strength of the association between the two variables
- Pearson's correlation coefficient (r) is a type of correlation coefficient
- Correlation coefficient returns a value between -1 and 1
  - -1 denotes strongest negative correlation
  - 0 denotes no correlation
  - 1 denotes strongest positive correlation

(Pearson) Correlation



Berapa nilai korelasi (Pearson r) masing-masing gambar ini?

(Pearson) Correlation



Positive Correlation

No correlation

Negative

(Pearson) Correlation

- Kapan kita menggunakan Pearson Correlation?

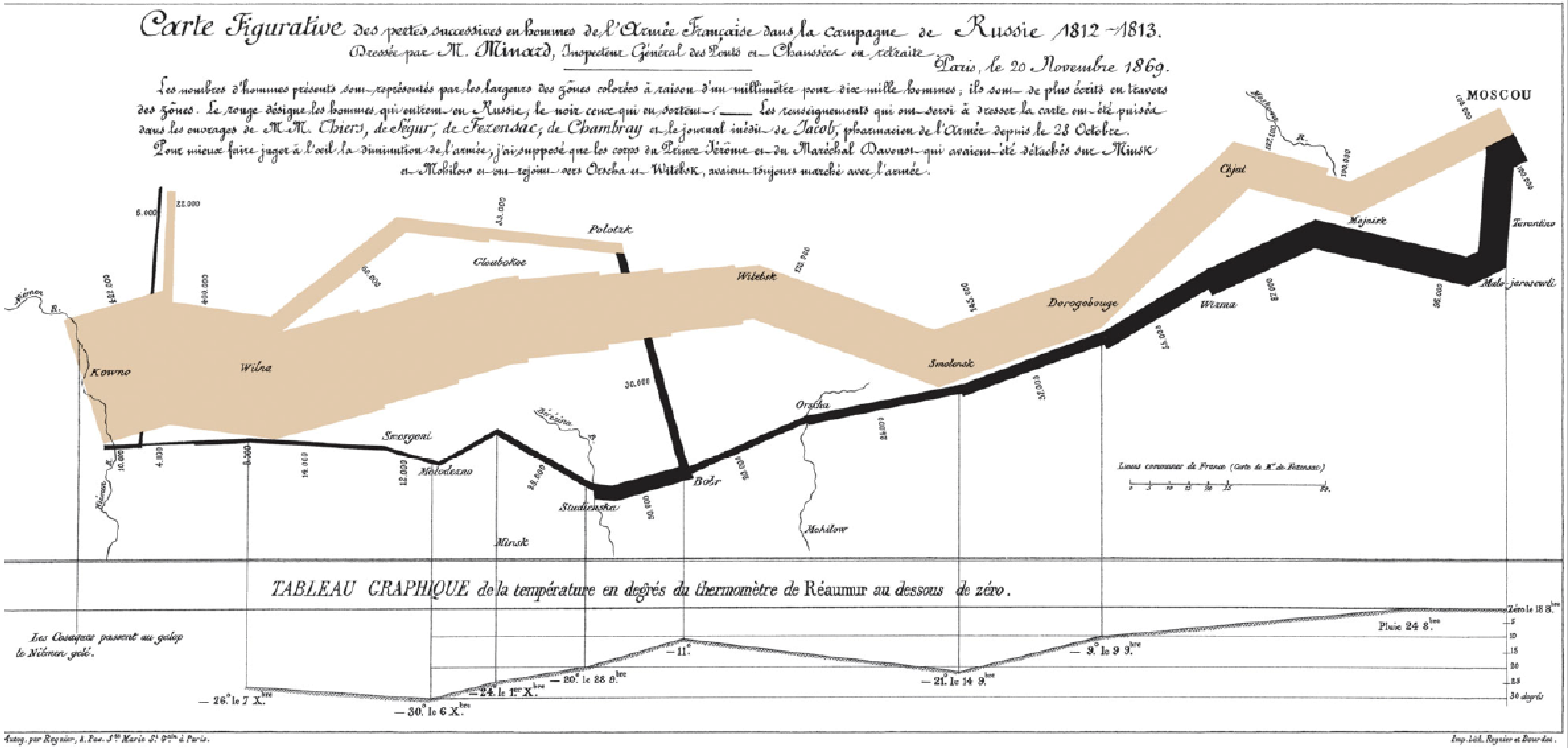
(Pearson) Correlation

- Kapan kita menggunakan Pearson Correlation?

- Untuk atribut-atribut yang tidak cocok dihitung korelasinya dengan Pearson Correlation, rumus apa yang bisa dipakai?

Data Visualization

Charles Joseph Minard 1869
Napoleon’s March



Carte Figurative des pertes successives en hommes de l'Armée Française dans la campagne de Russie 1812–1813.
Dressée par M. Minard, Inspecteur Général des Ponts et Chaussées en retraite.
Paris, le 20 Novembre 1869.

Les nombres d'hommes présents sont représentés par les largeurs des zônes colorées à raison d'un millimètre pour dix mille hommes; ils sont de plus écrits en travers des zônes. Le rouge désigne les hommes qui entrent en Russie, le noir ceux qui en sortent. —— Les renseignements qui ont servi à dresser la carte ont été puisés dans les ouvrages de M.M. Thiers, de Ségur, de Fezensac, de Chambray et le journal inédit de Jacob, pharmacien de l'Armée depuis le 28 Octobre.
Pour mieux faire juger à l'œil la diminution de l'armée, j'ai supposé que les corps du Prince Jérôme et du Maréchal Davoust qui avaient été détachés sur Minsk et Mobilow et ont rejoint vers Orscha et Witebsk, avaient toujours marché avec l'armée.

According to Tufte: “It may well be the best statistical graphic ever drawn.”
5 variables: Army Size, location, dates, direction, temperature during retreat

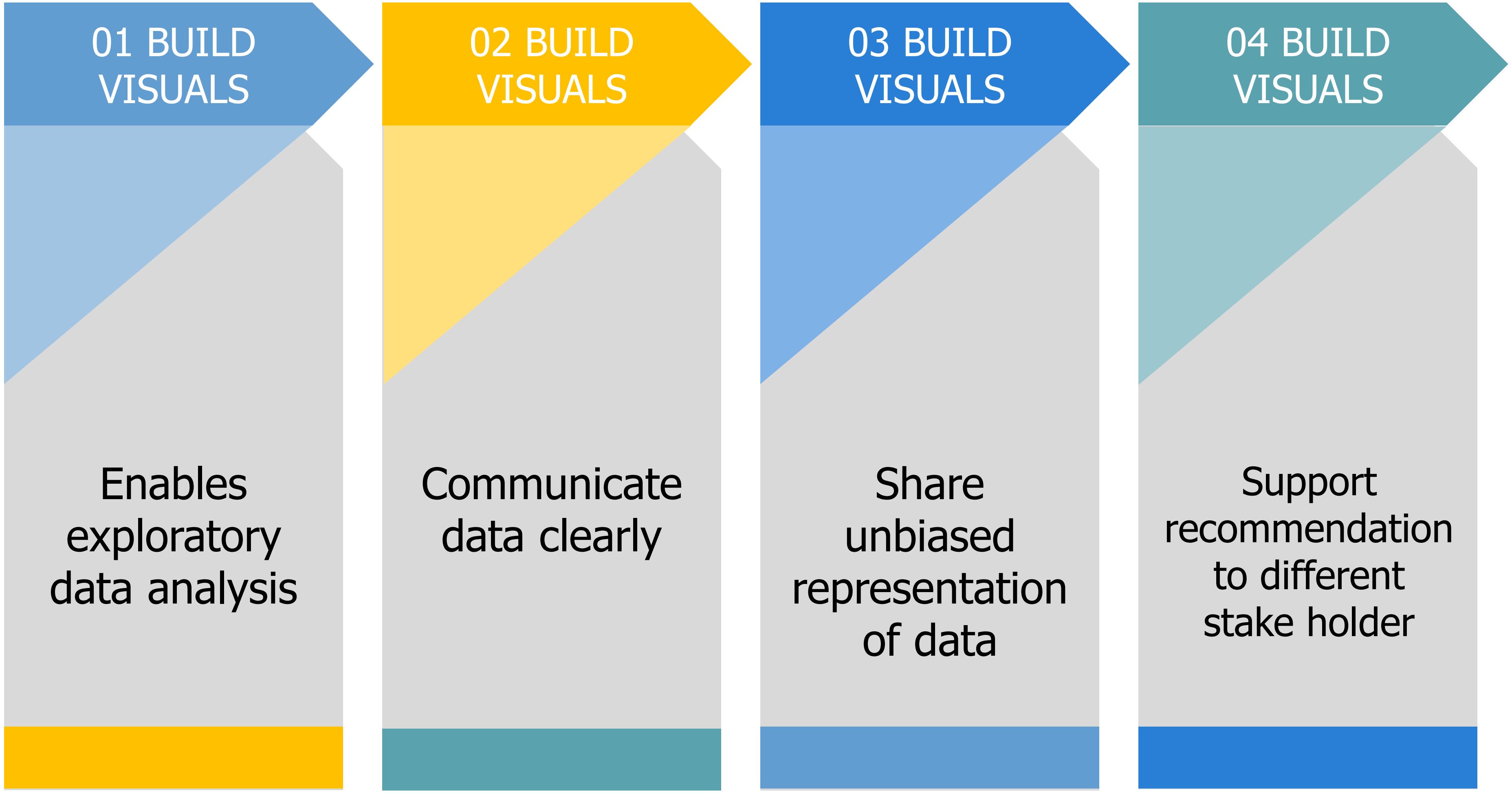
More Examples

- The famous GapminderVideo, Hans Rosling: 200 Countries, 200 Years, 4 Minutes: https://www.youtube.com/watch?feature=player_embedded&v=jbkSRLYSojo

- NY Times Interactive Visualizations (e.g., 2013 Federal Budget)

http://www.nytimes.com/interactive/2012/02/13/us/politics/2013-budget-proposal-graphic.html

Why Data Visualization?



01 BUILD VISUALS

Enables exploratory data analysis

02 BUILD VISUALS

Communicate data clearly

03 BUILD VISUALS

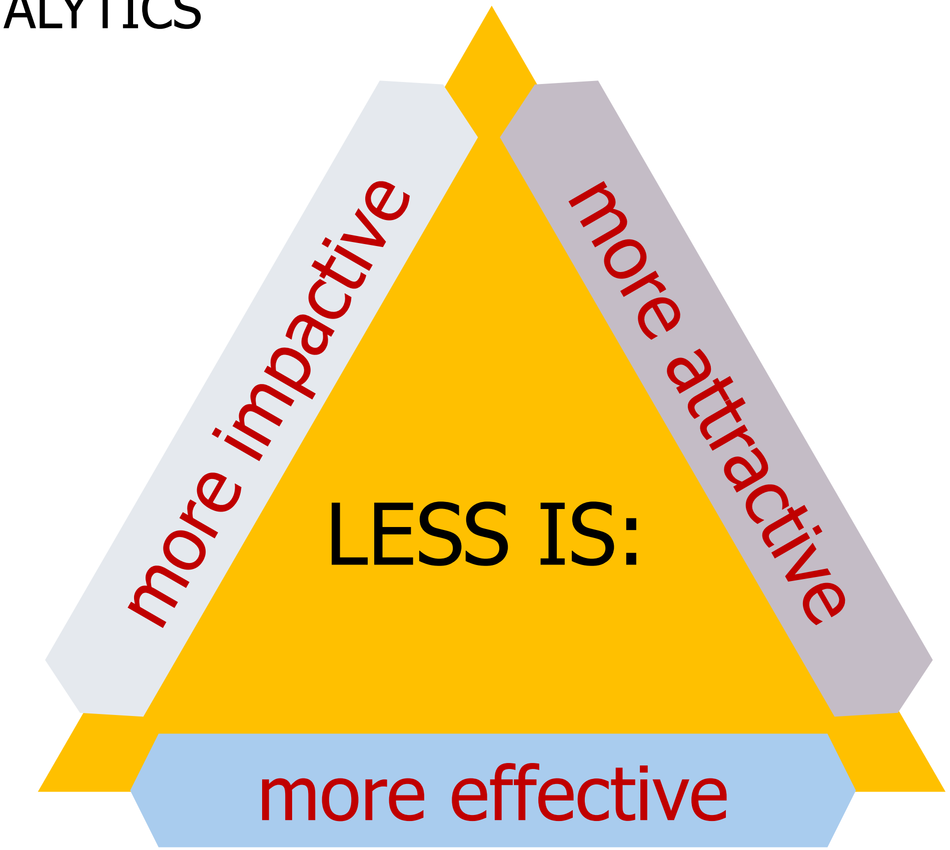
Share unbiased representation of data

04 BUILD VISUALS

Support recommendation to different stake holder

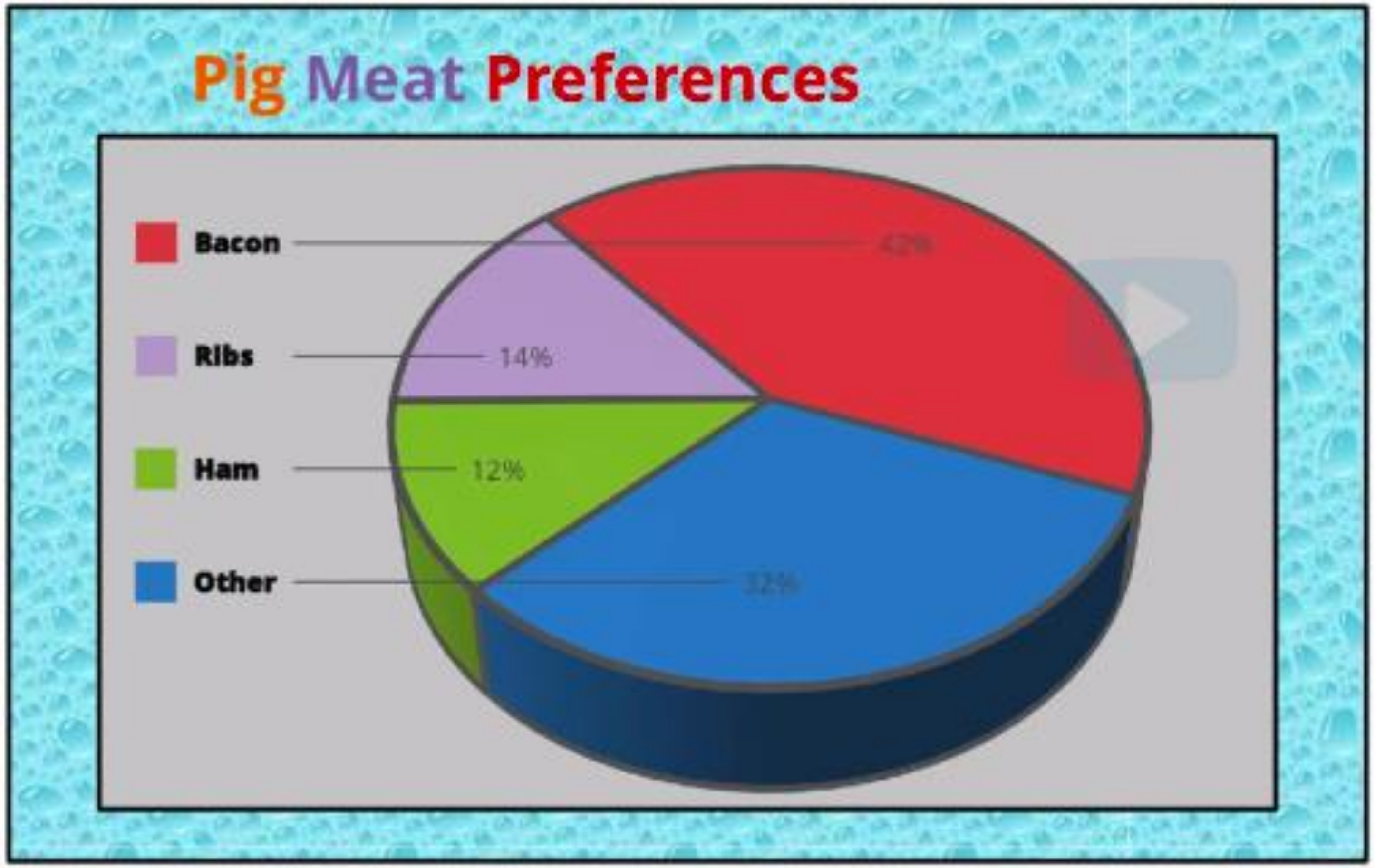
Three Key Points of Build Visuals

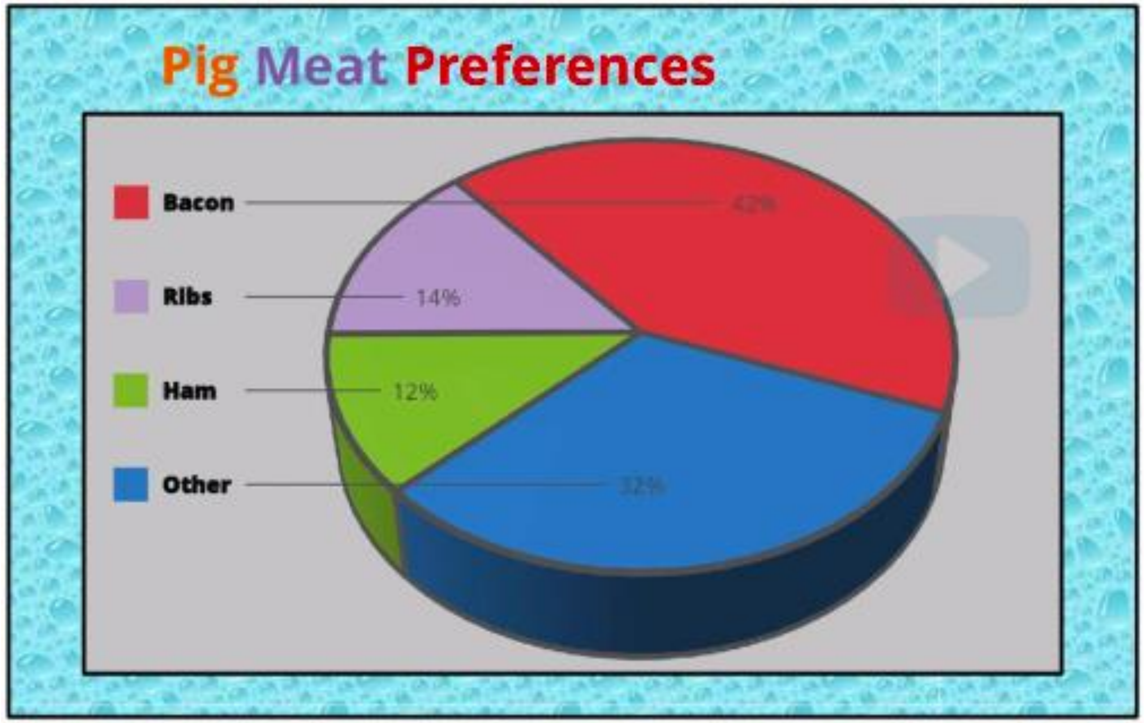
by DARKHORSE ANALYTICS



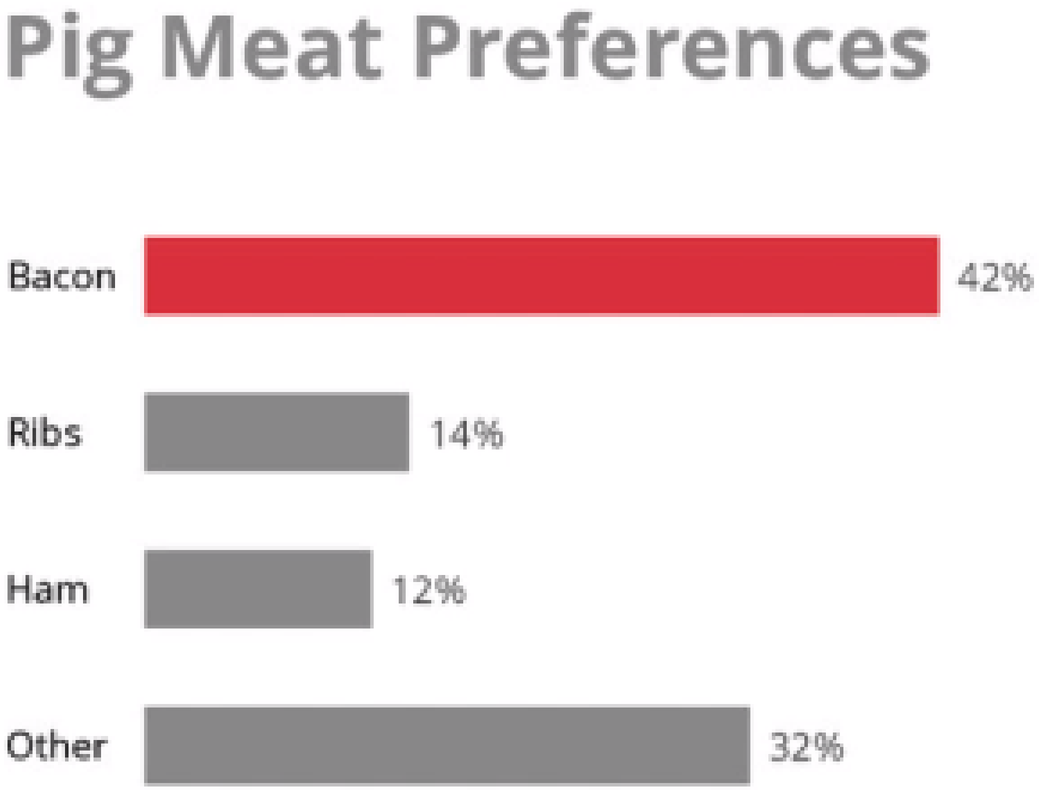
Any **feature or design** you incorporate in your plot to make it more attractive or pleasing should **support the message** that the plot **is meant** to get across and **not distract from it**.

Look at this figure

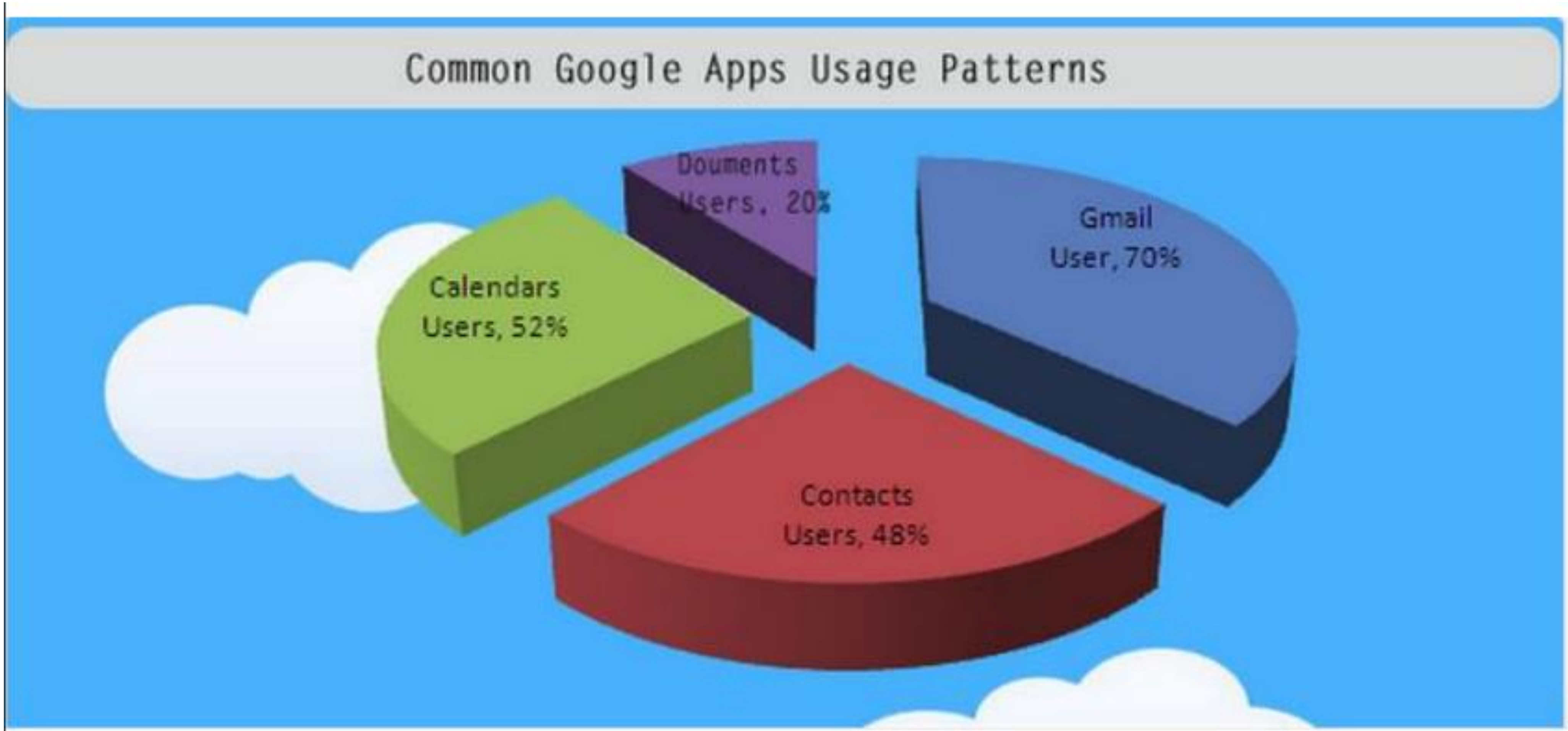




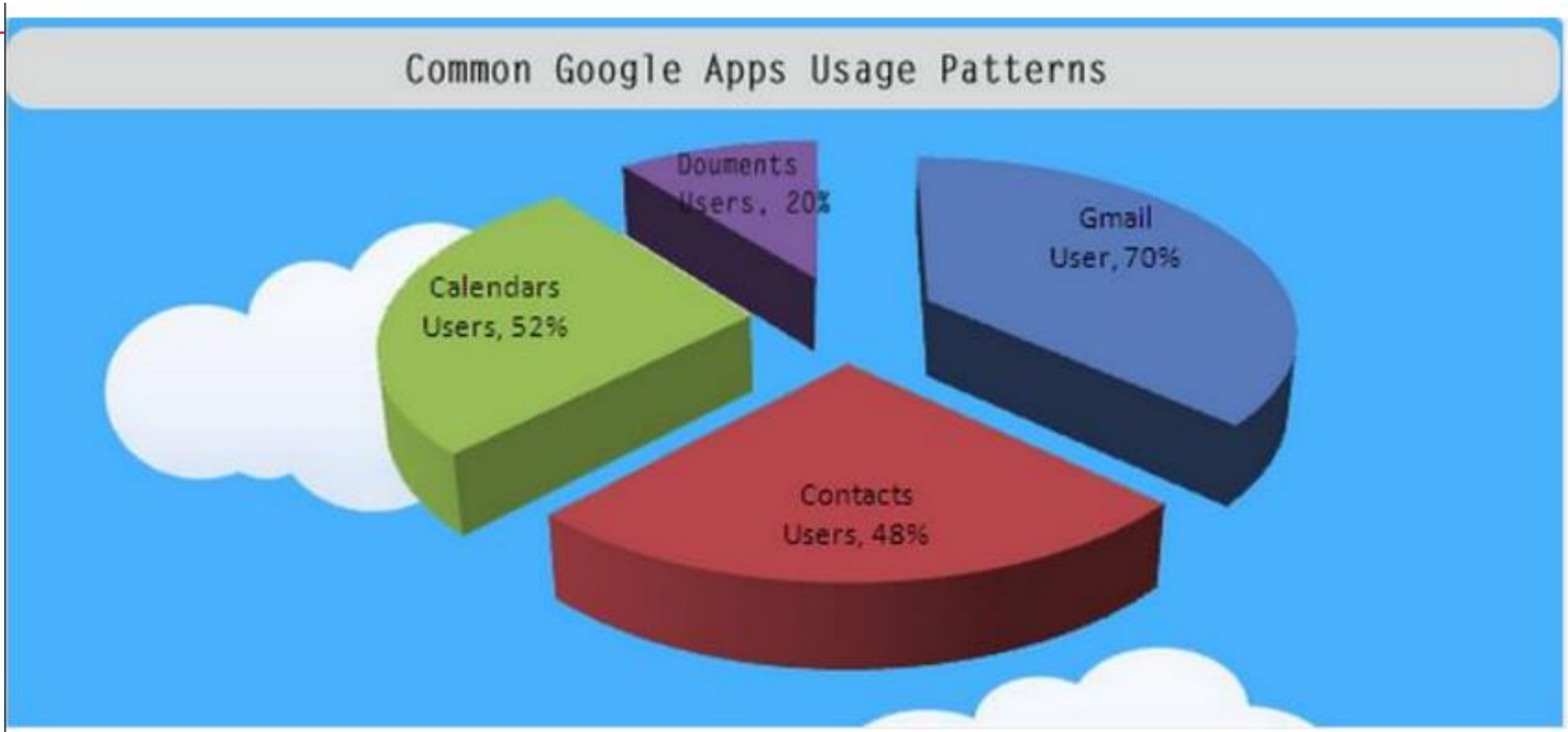
- It looks like features such as the blue background or 3D orientation are meant to convey anything.
- In fact, these additional unnecessary features distract from the main message and can be confusing to the audience.



- The message here is that people are most likely to choose bacon over other types of pig meat, so let's get rid of everything that can be distracting from this core message.
- It is simple, cleaner, less distracting, and much easier to read.

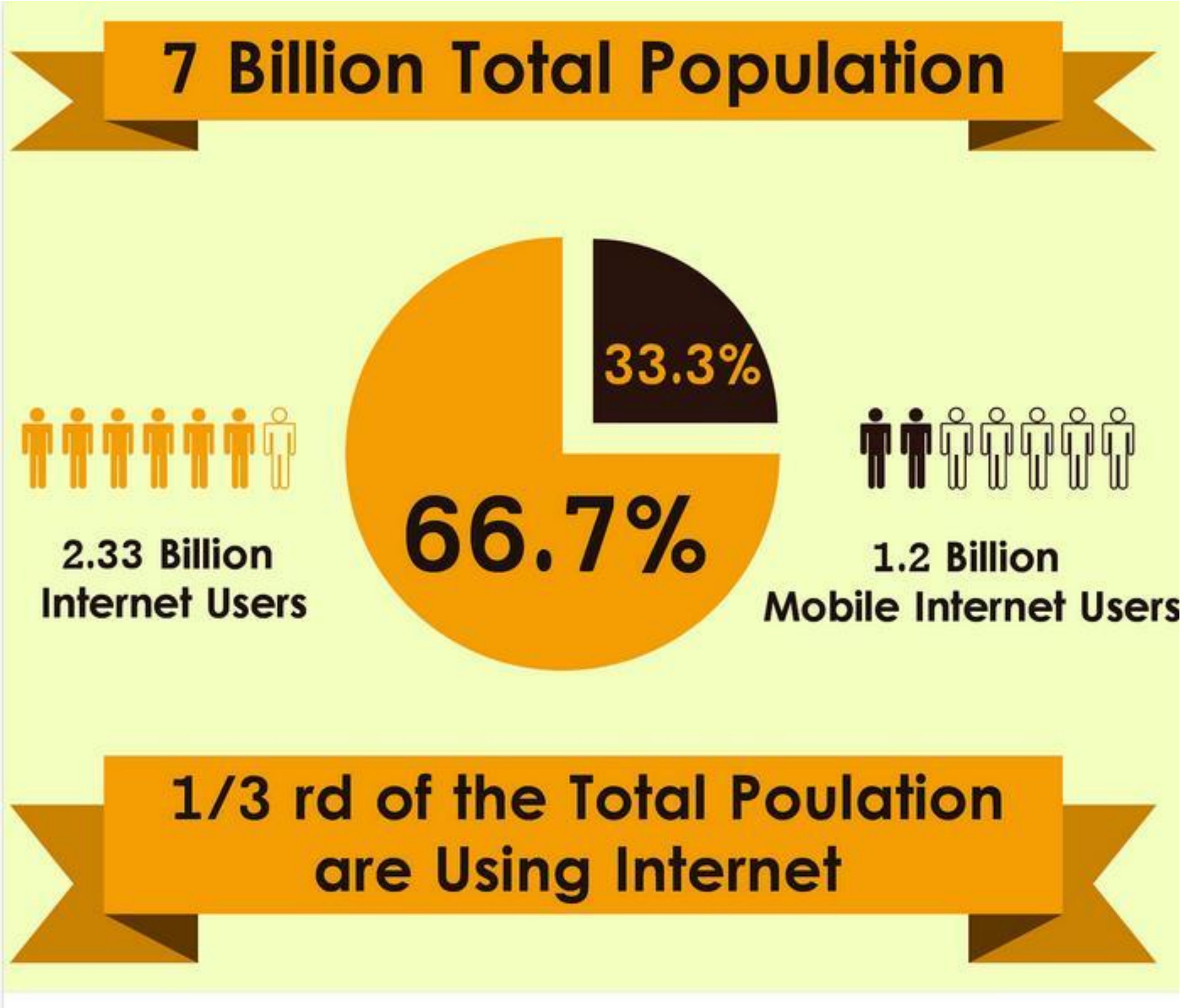


from wtfviz.net



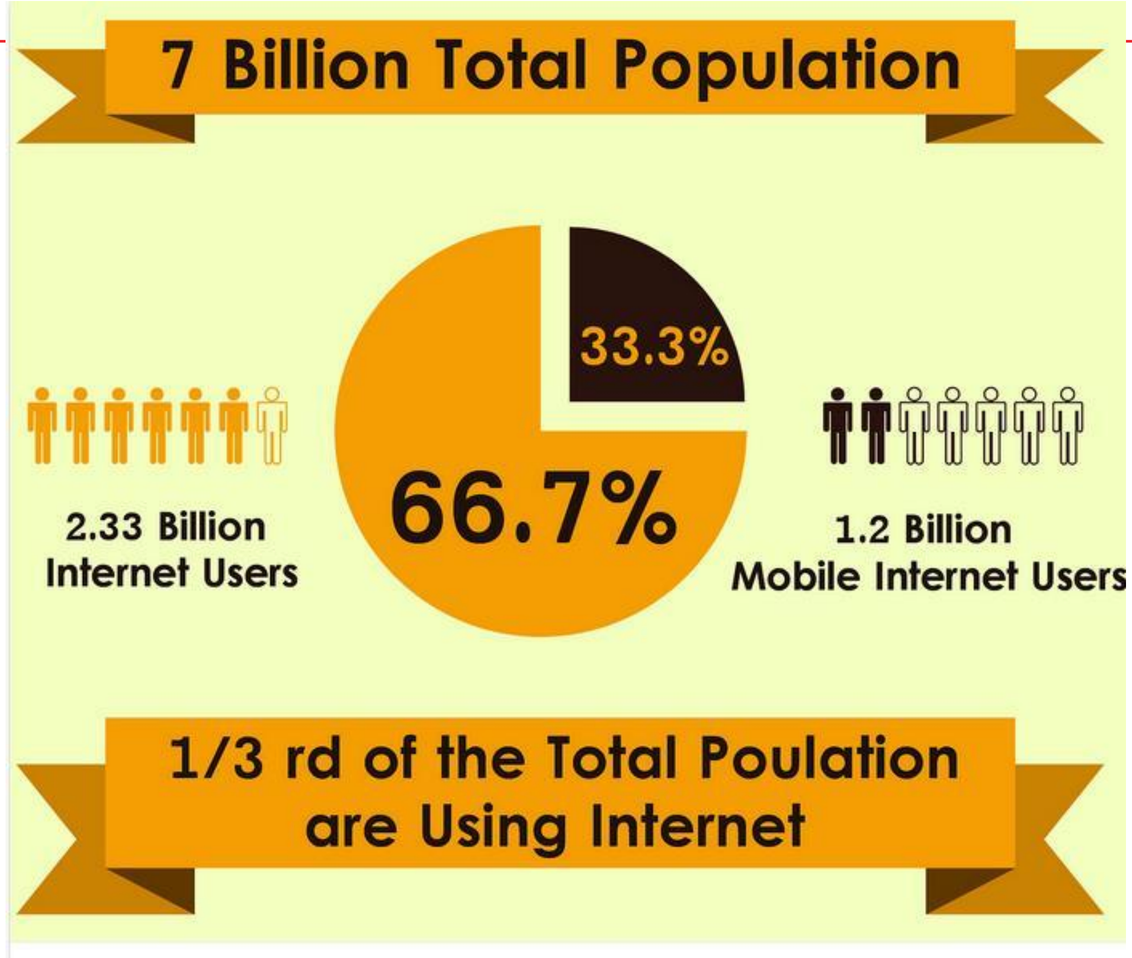
from wtfviz.net

- The proportion of each pie is wrong.
- Unnecessary sky background.

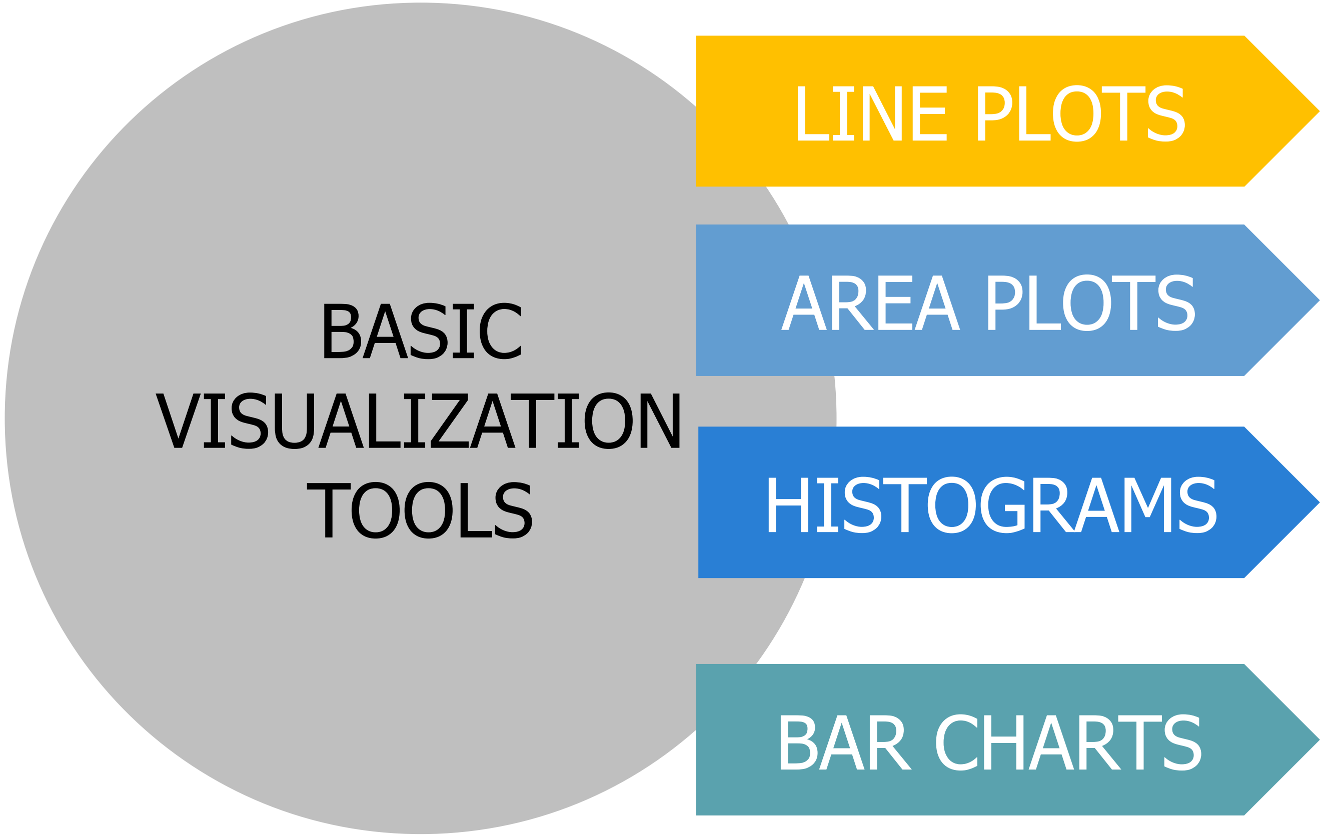


7 Billion Total Population

1/3 rd of the Total Poulation are Using Internet



- Are you sure the internet users are only 1/3rd of the total population?



BASIC VISUALIZATION TOOLS

LINE PLOTS

AREA PLOTS

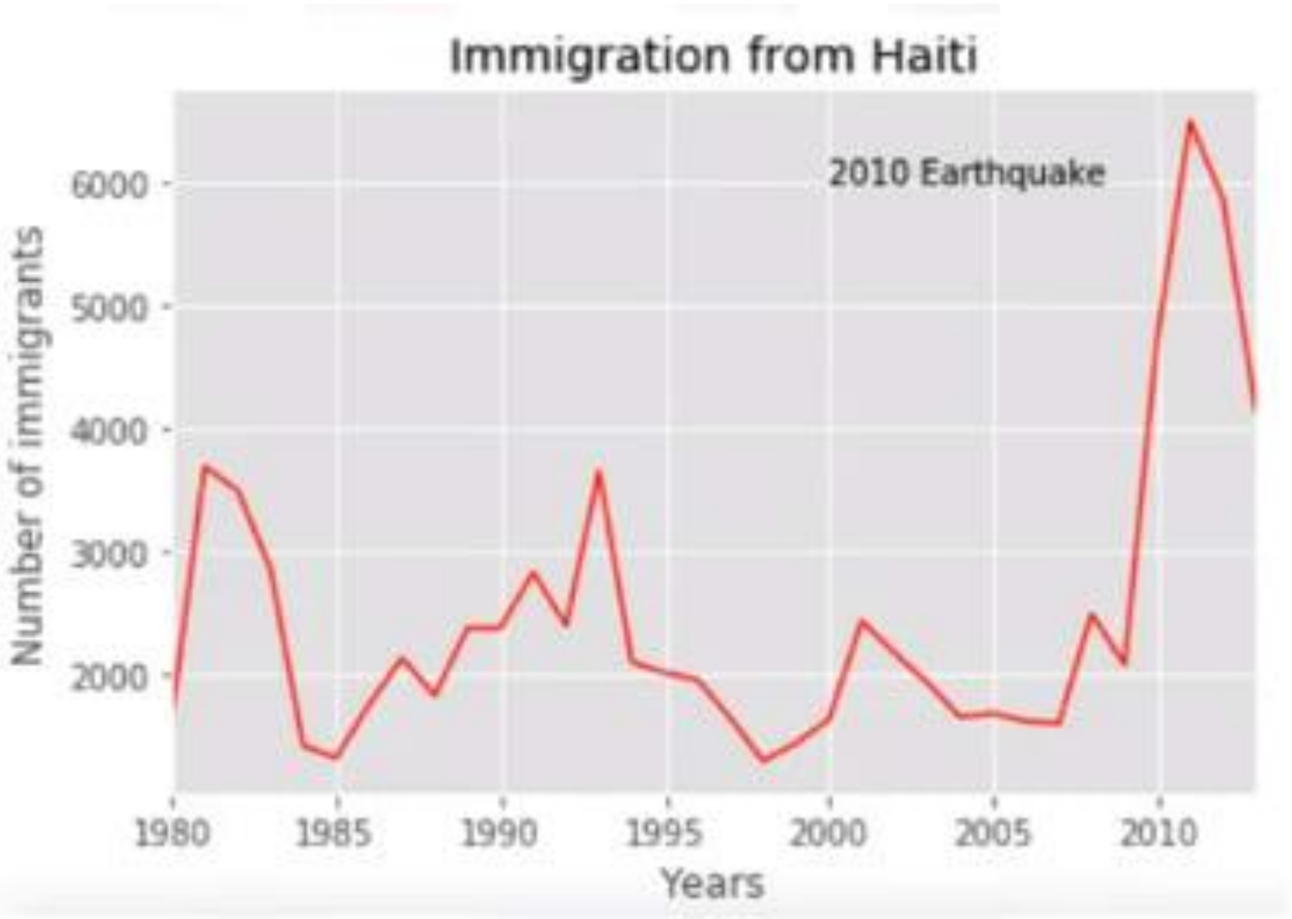
HISTOGRAMS

BAR CHARTS

LINE PLOTS

- Line plot is a plot in the form of a series of data points connected by straight line segments.
- The best use case for a line plot is when you have a continuous dataset and you're interested in visualizing the data over a period of time.

LINE PLOTS



For example, say we're interested in the trend of immigrants from Haiti to Canada.

We can generate a line plot and the resulting figure will depict the trend of Haitian immigrants to Canada from 1980 to 2013.

Based on the line plot, we can then research for justifications of obvious anomalies or changes

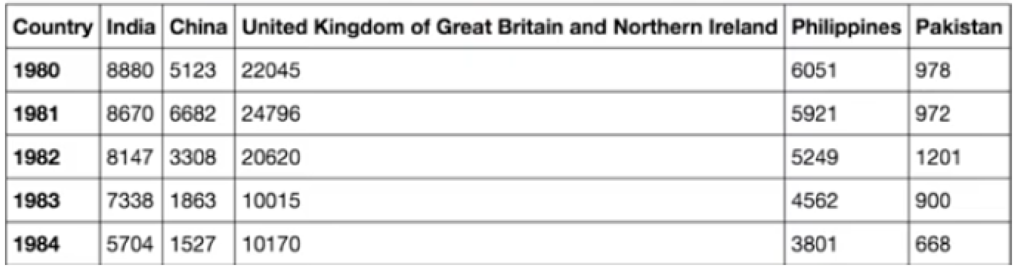
From previous plot, we see that there is a spike of immigration from Haiti to Canada in 2010.

A quick Google search for major events in Haiti in 2010 would return the tragic earthquake that took place in 2010, and therefore this influx of immigration to Canada was mainly due to that tragic earthquake.

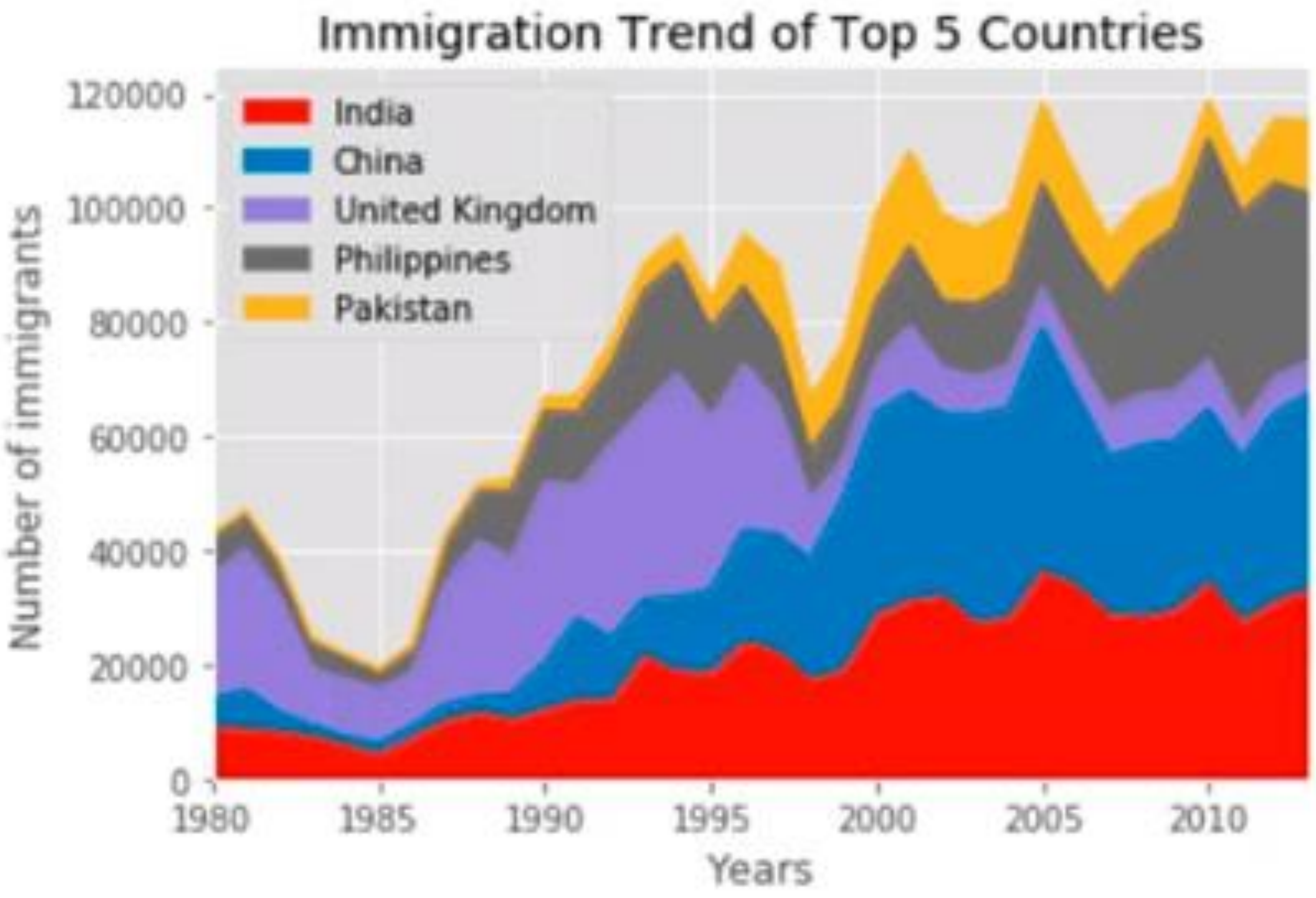
AREA PLOTS

- An area Plot (also known as an area chart or area graph) depicts accumulated totals using numbers or percentages over time.
- It is based on the line plot and is commonly used when trying to compare two or more quantities.

AREA PLOTS



Country	India	China	United Kingdom of Great Britain and Northern Ireland	Philippines	Pakistan
1980	8880	5123	22045	6051	978
1981	8670	6682	24796	5921	972
1982	8147	3308	20620	5249	1201
1983	7338	1863	10015	4562	900
1984	5704	1527	10170	3801	668



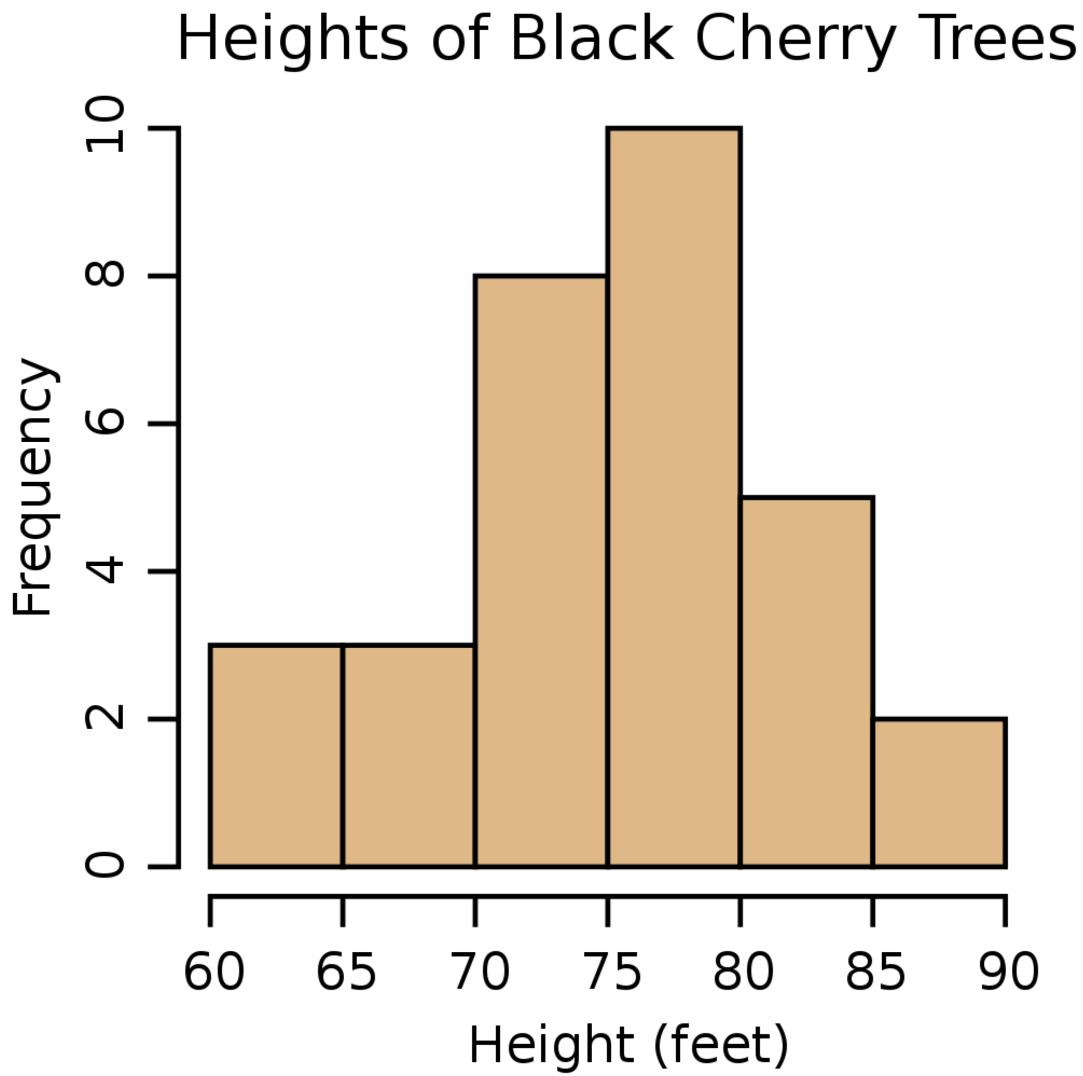
HISTOGRAMS

- **A histogram** is a way of representing the frequency **distribution of a numeric dataset**.
- It takes as input one numerical variable.

- The variable is **cut into several bins**, and **the number of observations per bin** is represented by **the height of the bar**.

- To construct a histogram, the first step is to “bin” the range of values—that is, divide the entire range of values into a series of intervals—and then count how many values fall into each interval. The bins are usually specified as consecutive, non-overlapping intervals of a variable.

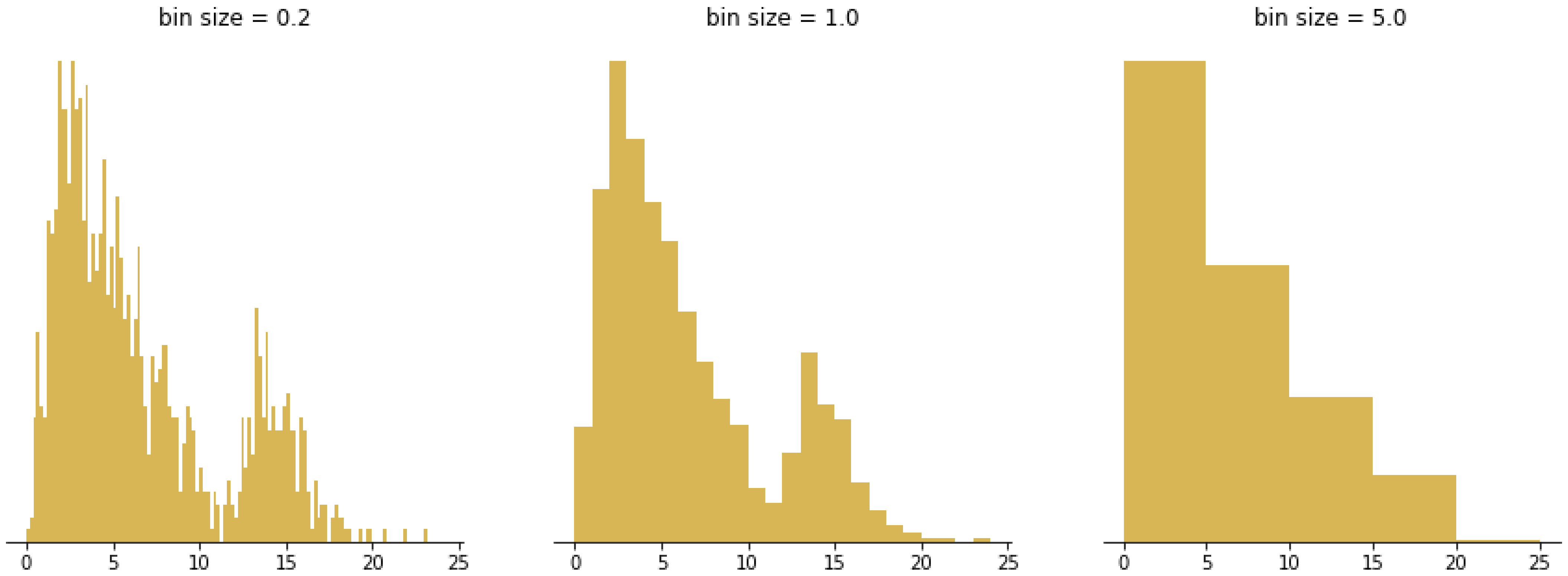
HISTOGRAMS



https://id.wikipedia.org/wiki/Histogram#/media/Berkas:Black_cherry_tree_histogram.svg

HISTOGRAMS

Histogram with different bin size



https://chartio.com/learn/charts/histogram-complete-guide/

HISTOGRAMS

The number of bins needs to be:

- large enough to reveal interesting features;
- small enough not to be too noisy.

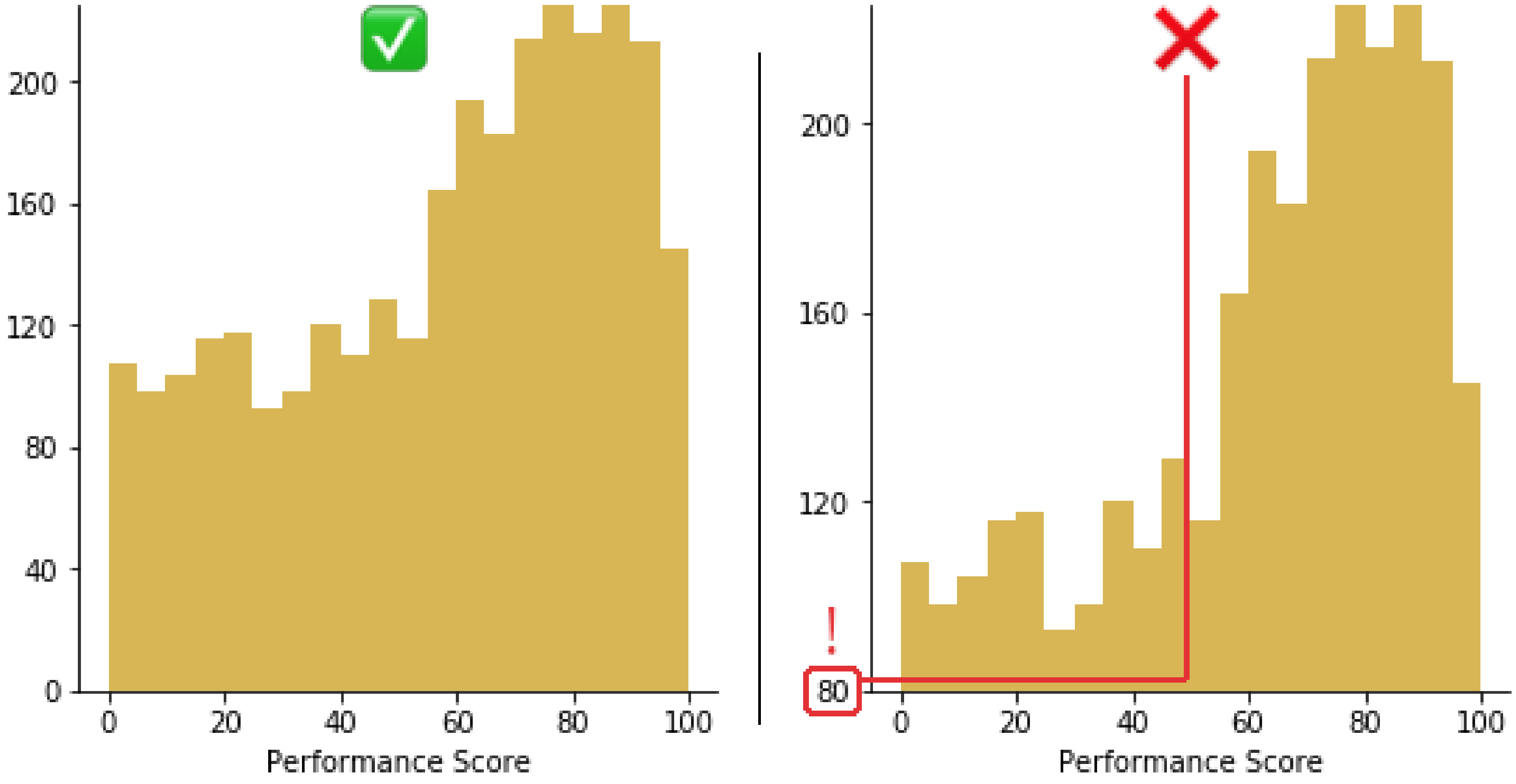
Choice of bin size has an inverse relationship with the number of bins.

  - The larger the bin sizes, the fewer bins there will be to cover the whole range of data.
  - With a smaller bin size, the more bins there will need to be.
  - It is worth taking some time to test out different bin sizes to see how the distribution looks in each one, then choose the plot that represents the data best.

https://chartio.com/learn/charts/histogram-complete-guide/

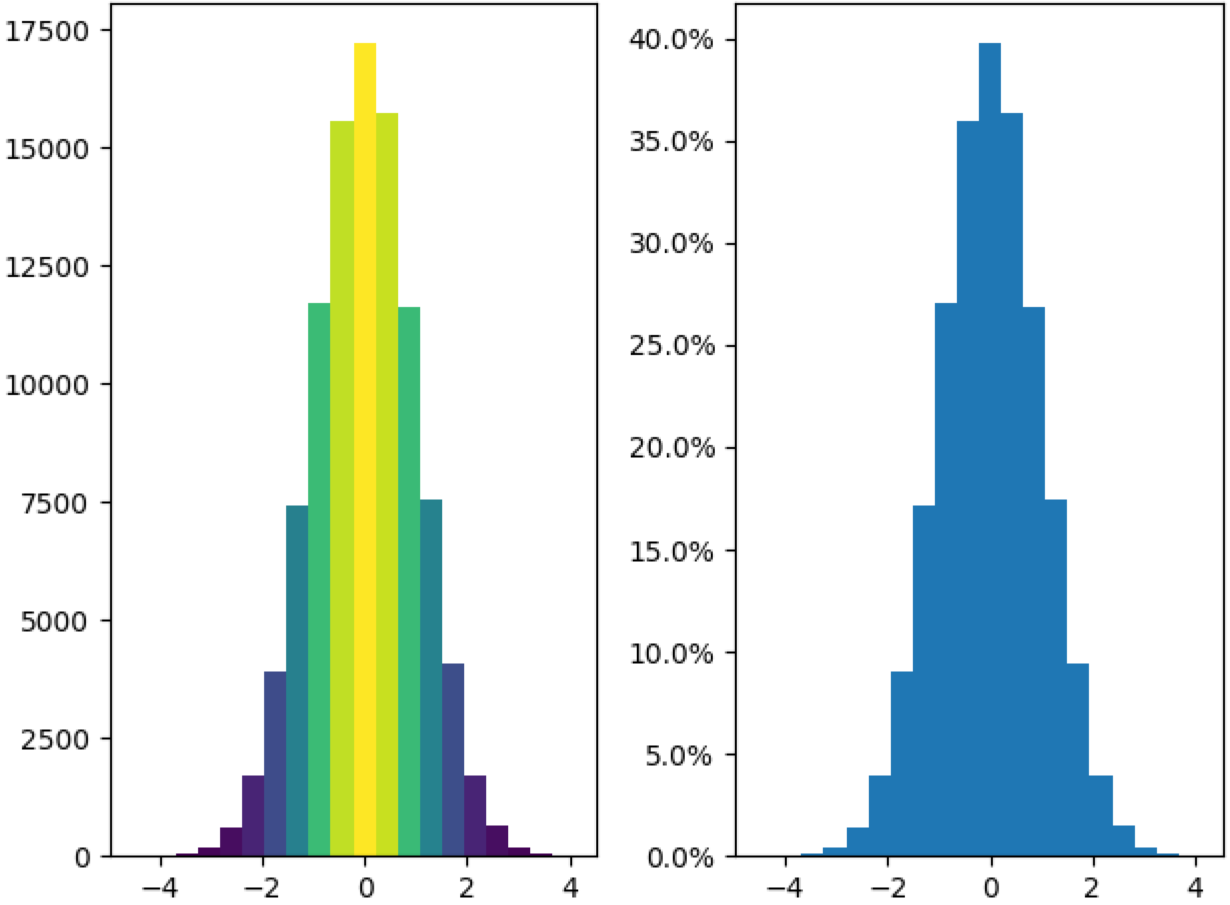
HISTOGRAMS

Use a zero-valued base line



https://chartio.com/learn/charts/histogram-complete-guide/

HISTOGRAMS



Updating Histogram with Colors

https://matplotlib.org/3.3.4/gallery/statistics/hist.html

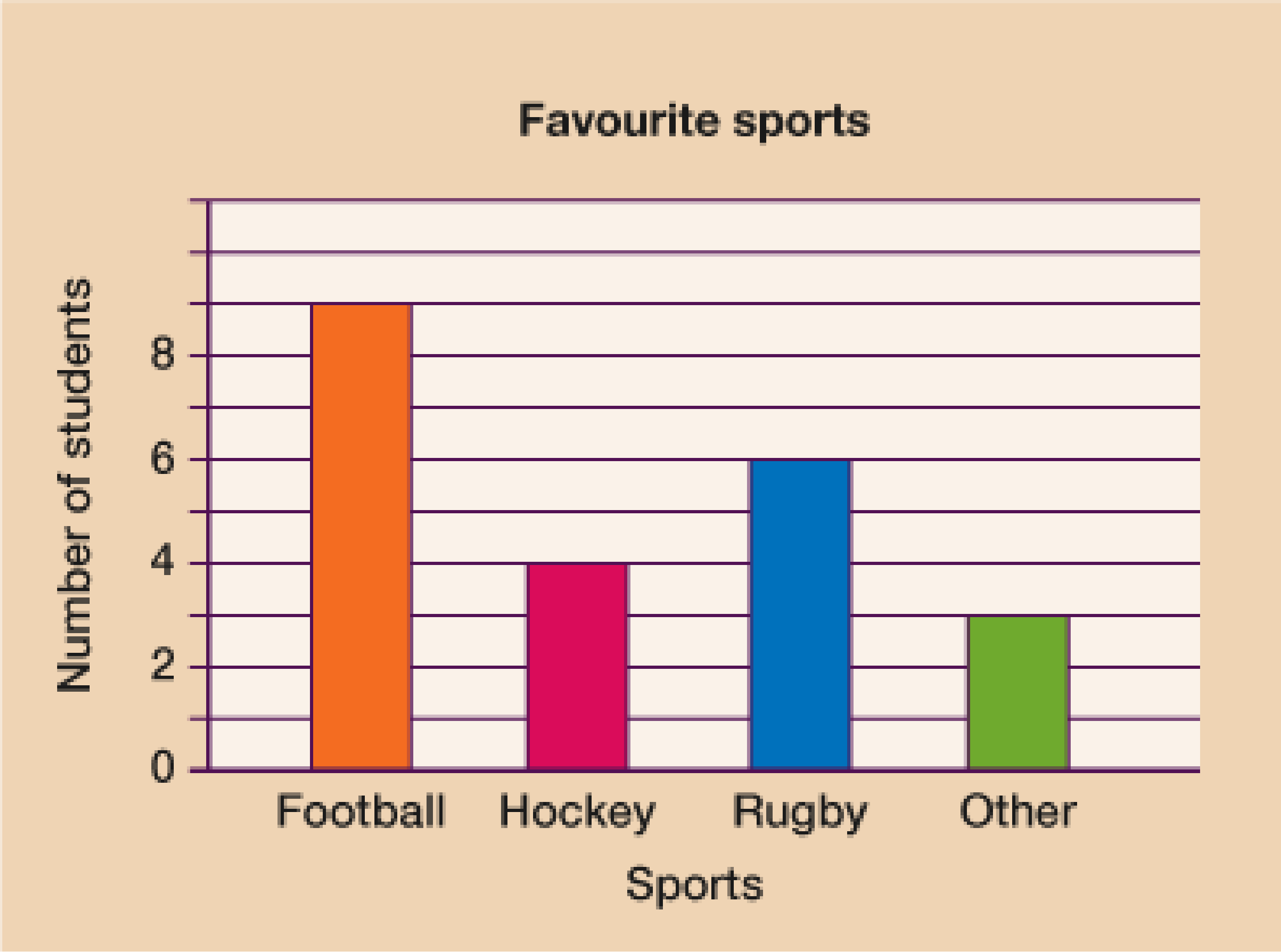
BAR CHARTS

- Unlike a histogram, **a bar chart** also known as a **bar graph is a type of plot where the length of each bar is proportional to** the value of the item that it represents.
- It is commonly used to compare the values of a variable at a given point in time.

BAR CHARTS

- Unlike a histogram, **a bar chart** also known as a **bar graph is a type of plot where the length of each bar is proportional to** the value of the item that it represents.
- It is commonly used to compare the values of a variable at a given point in time.

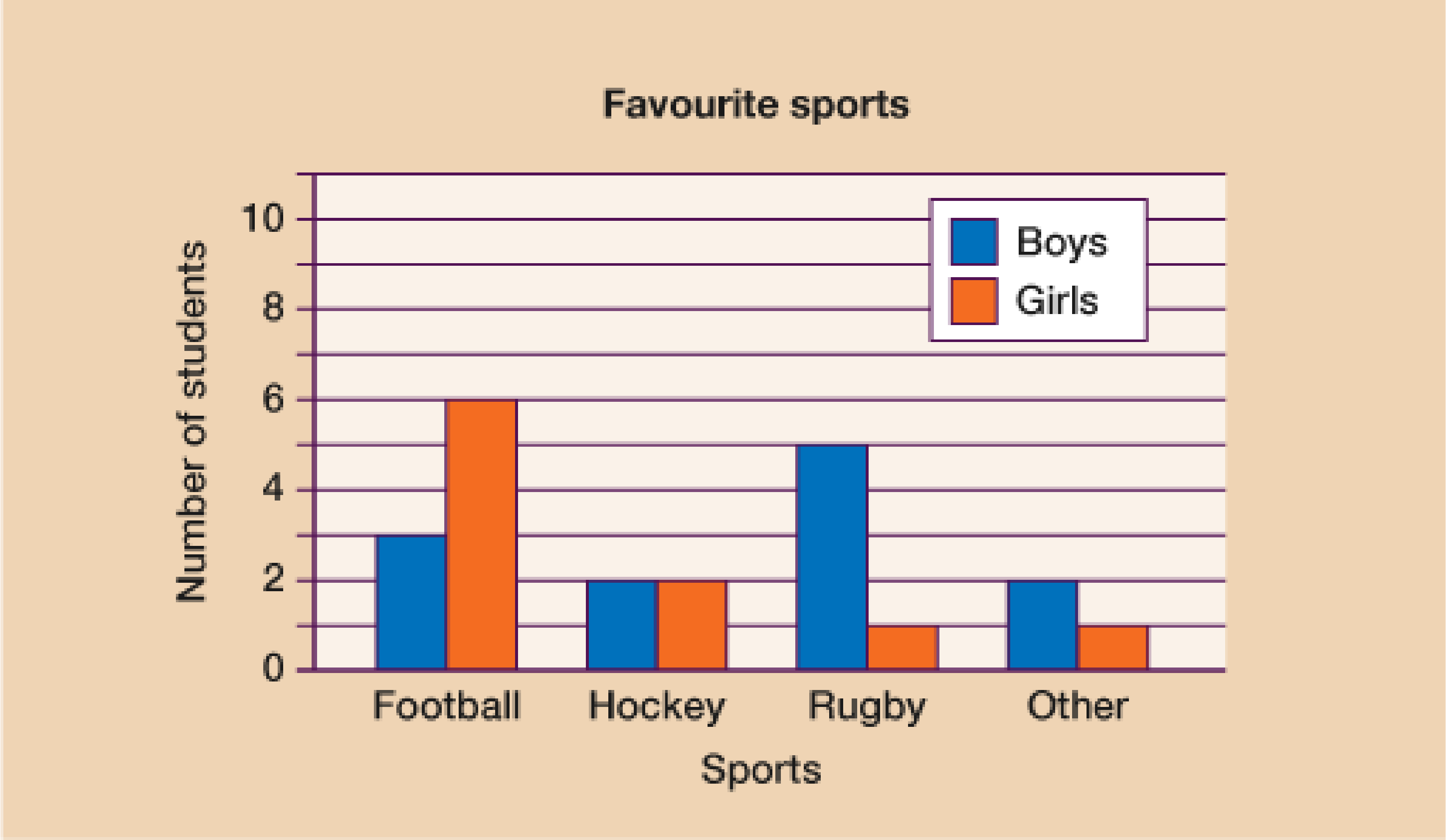
BAR CHARTS



Single Bar Chart

https://www.open.edu/openlearn/mod/oucontent/view.php?id=90853&extra=thumbnailfigure_idm333

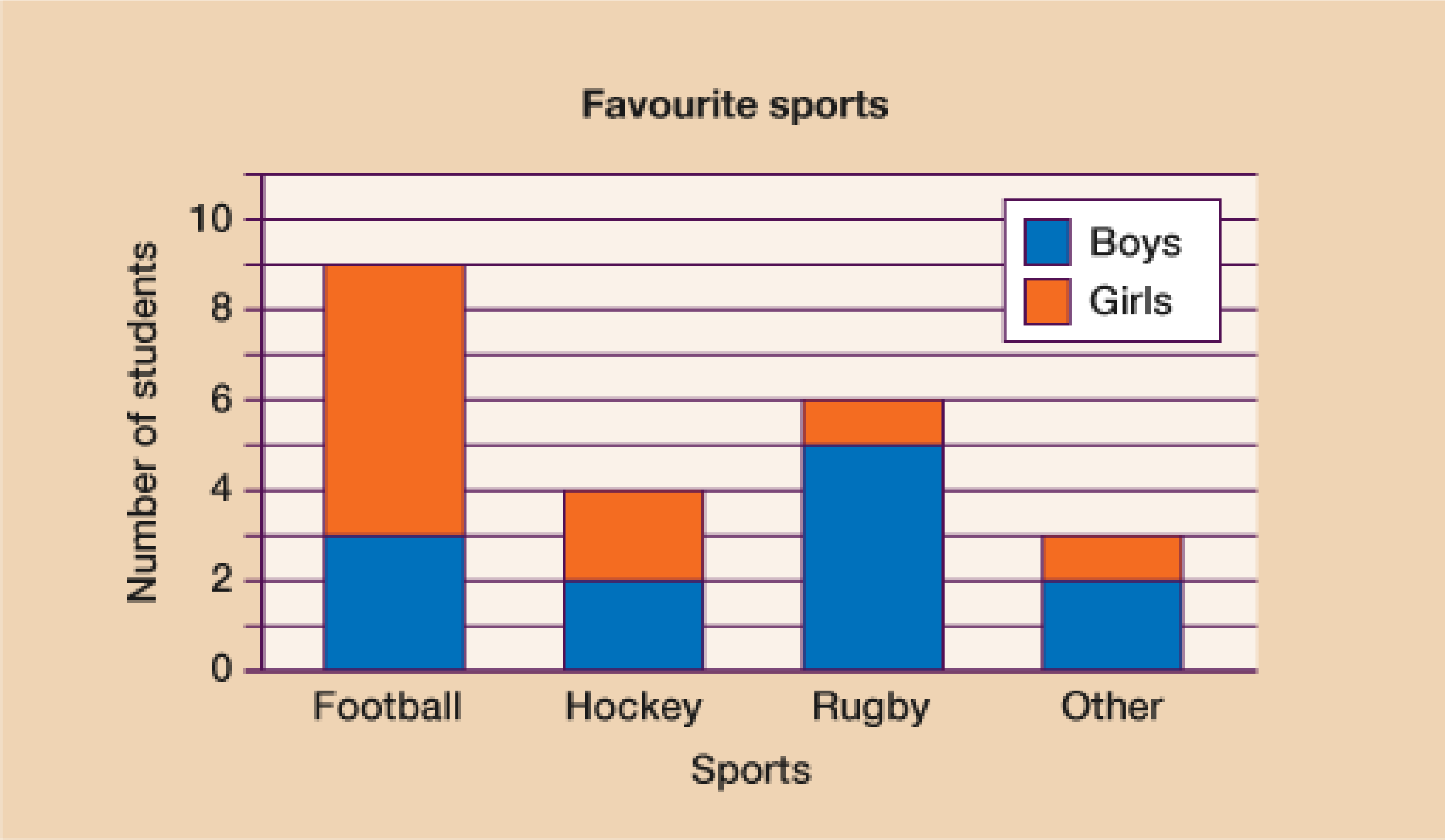
BAR CHARTS



Dual Bar Chart

https://www.open.edu/openlearn/mod/oucontent/view.php?id=90853&extra=thumbnailfigure_idm333

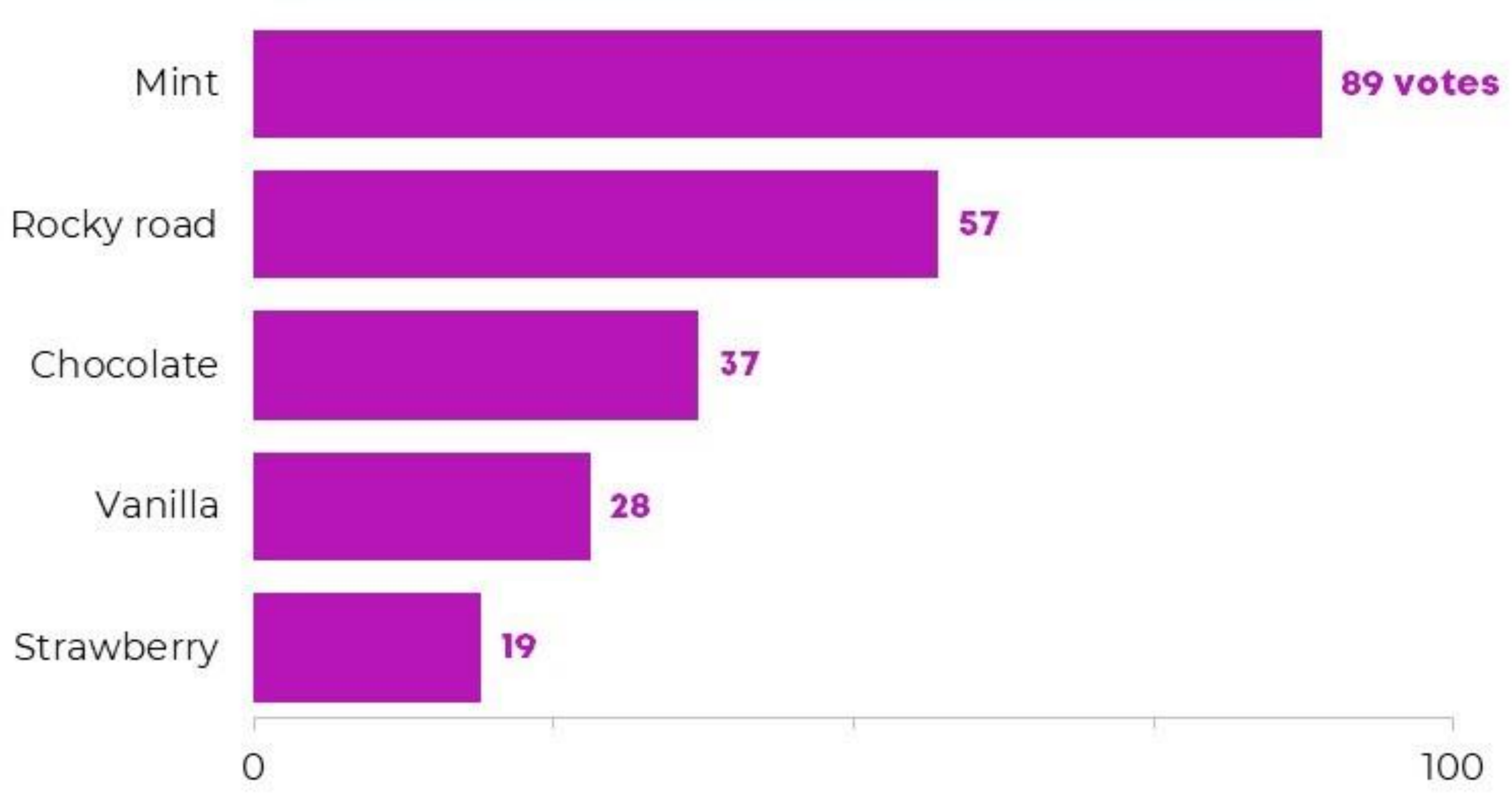
BAR CHARTS



Stacked Bar Chart

https://www.open.edu/openlearn/mod/oucontent/view.php?id=90853&extra=thumbnailfigure_idm333

BAR CHARTS

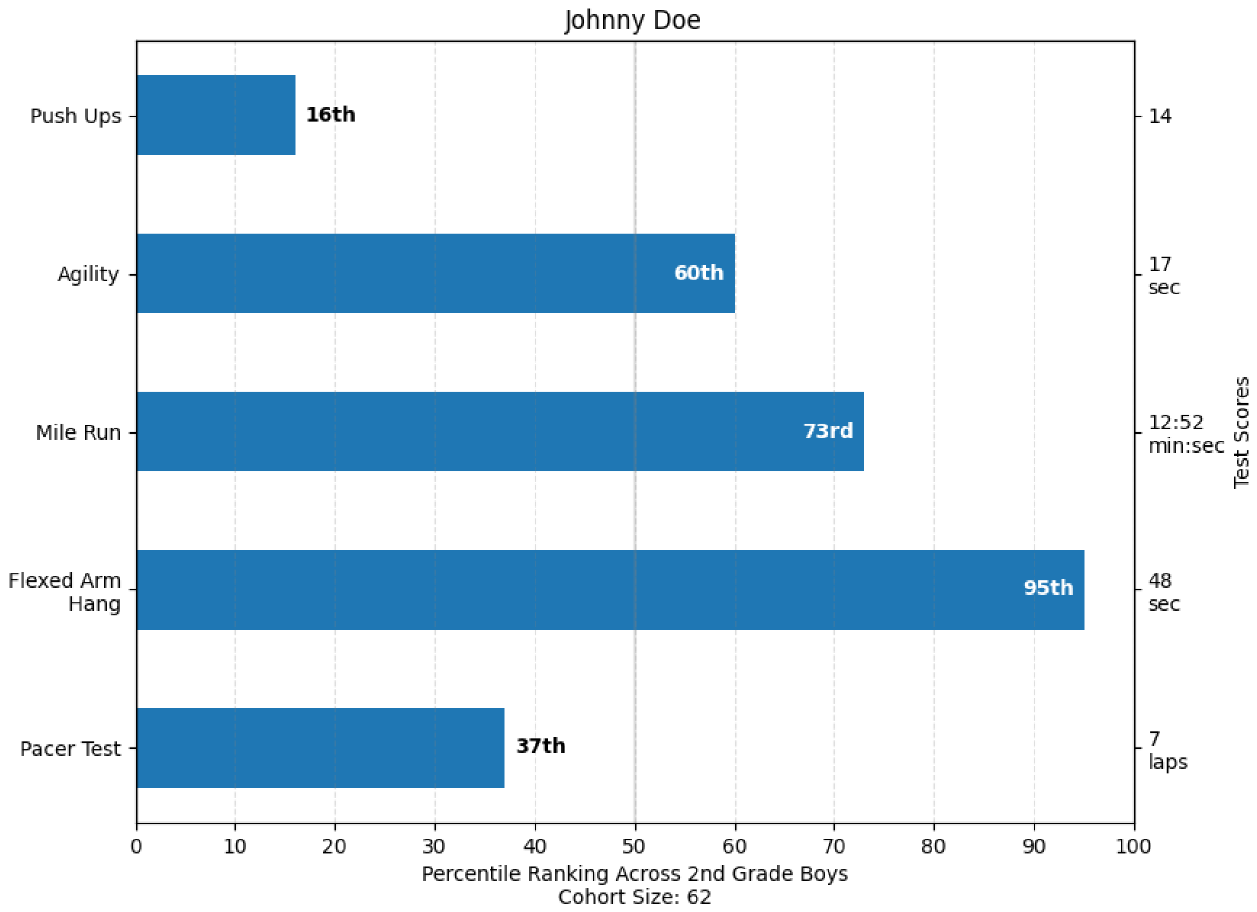


Horizontal Bar Chart

Apa keunggulannya?

https://matplotlib.org/3.4.0/gallery/statistics/barchart_demo.html#sphx-glr-gallery-statistics-barchart-demo-py

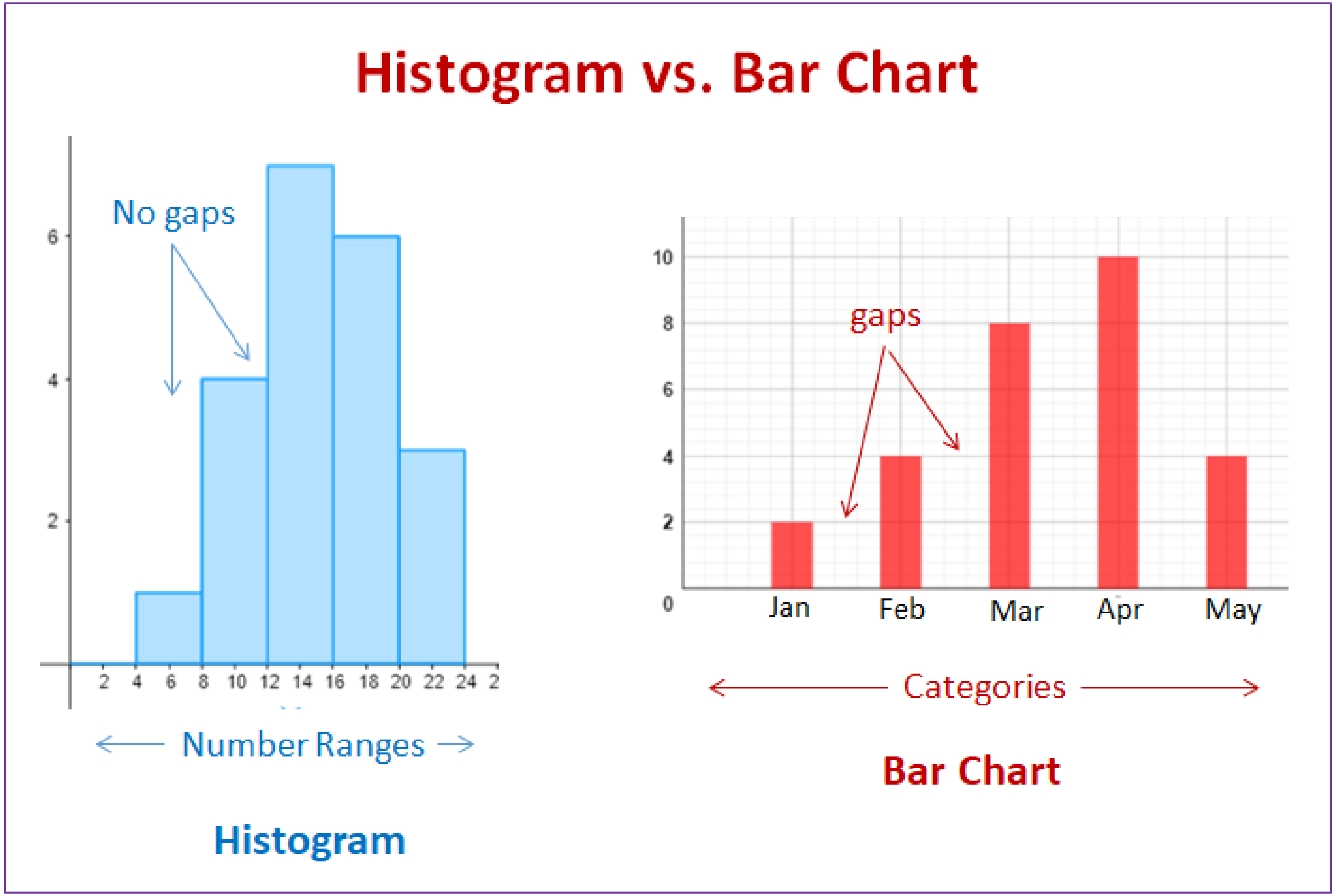
BAR CHARTS



Percentiles as Horizontal Bar Chart

https://matplotlib.org/3.4.0/gallery/statistics/barchart_demo.html#sphx-glr-gallery-statistics-barchart-demo-py

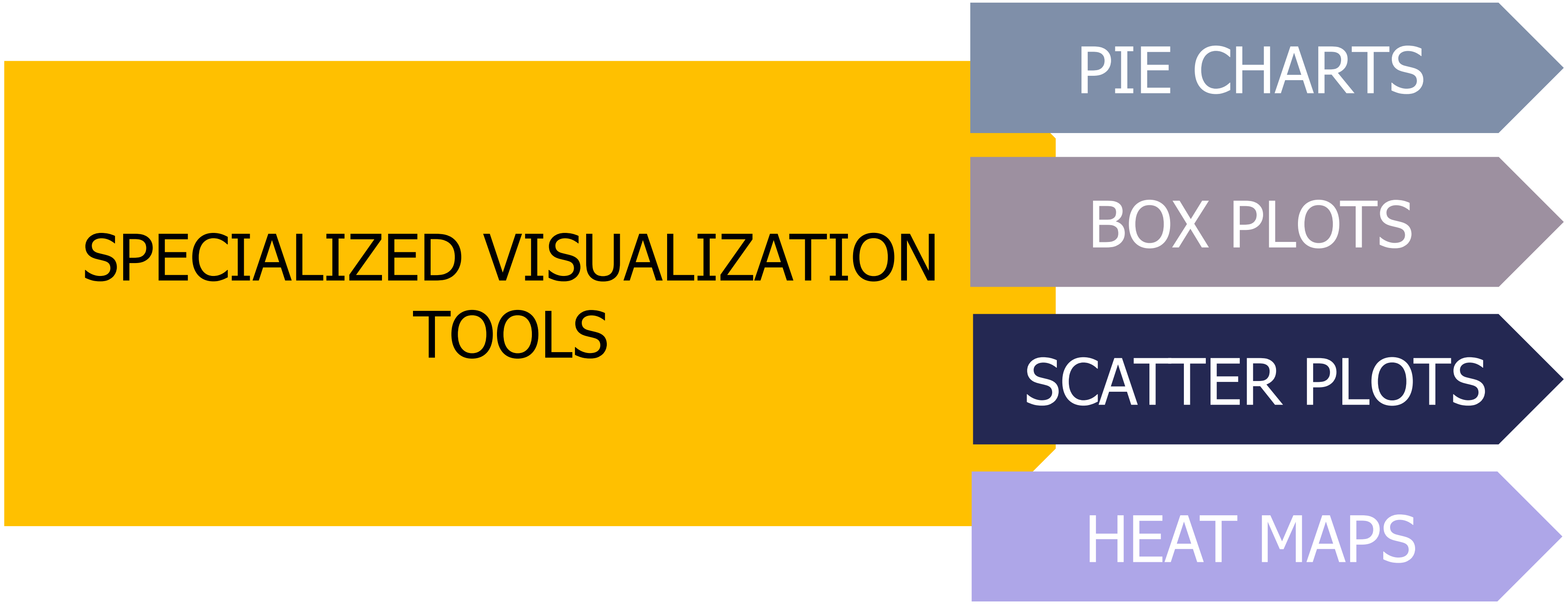
Histogram vs Bar Chart



Histogram vs. Bar Chart

Histogram

Bar Chart



SPECIALIZED VISUALIZATION TOOLS

PIE CHARTS

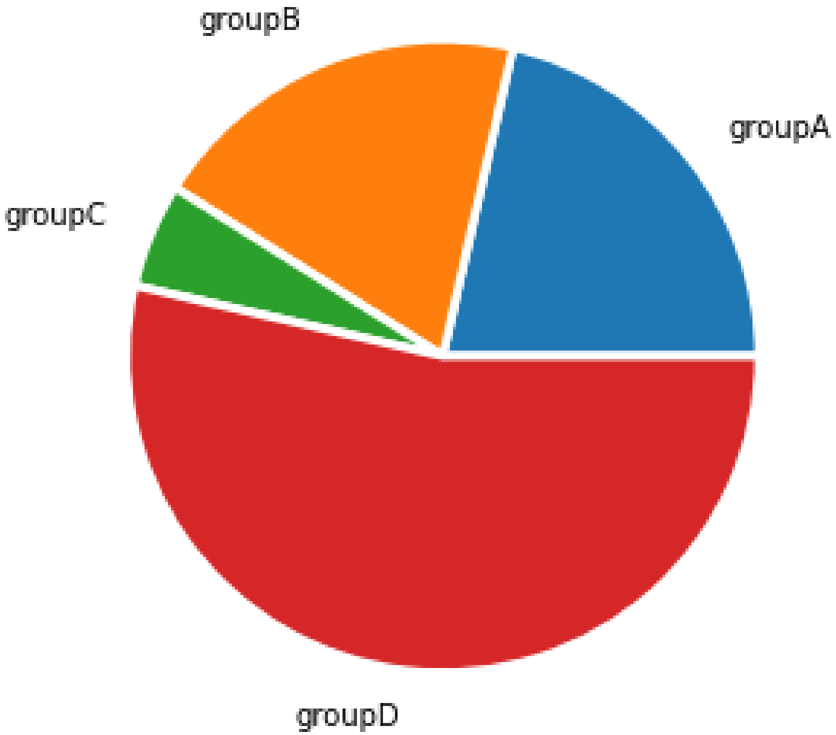
BOX PLOTS

SCATTER PLOTS

HEAT MAPS

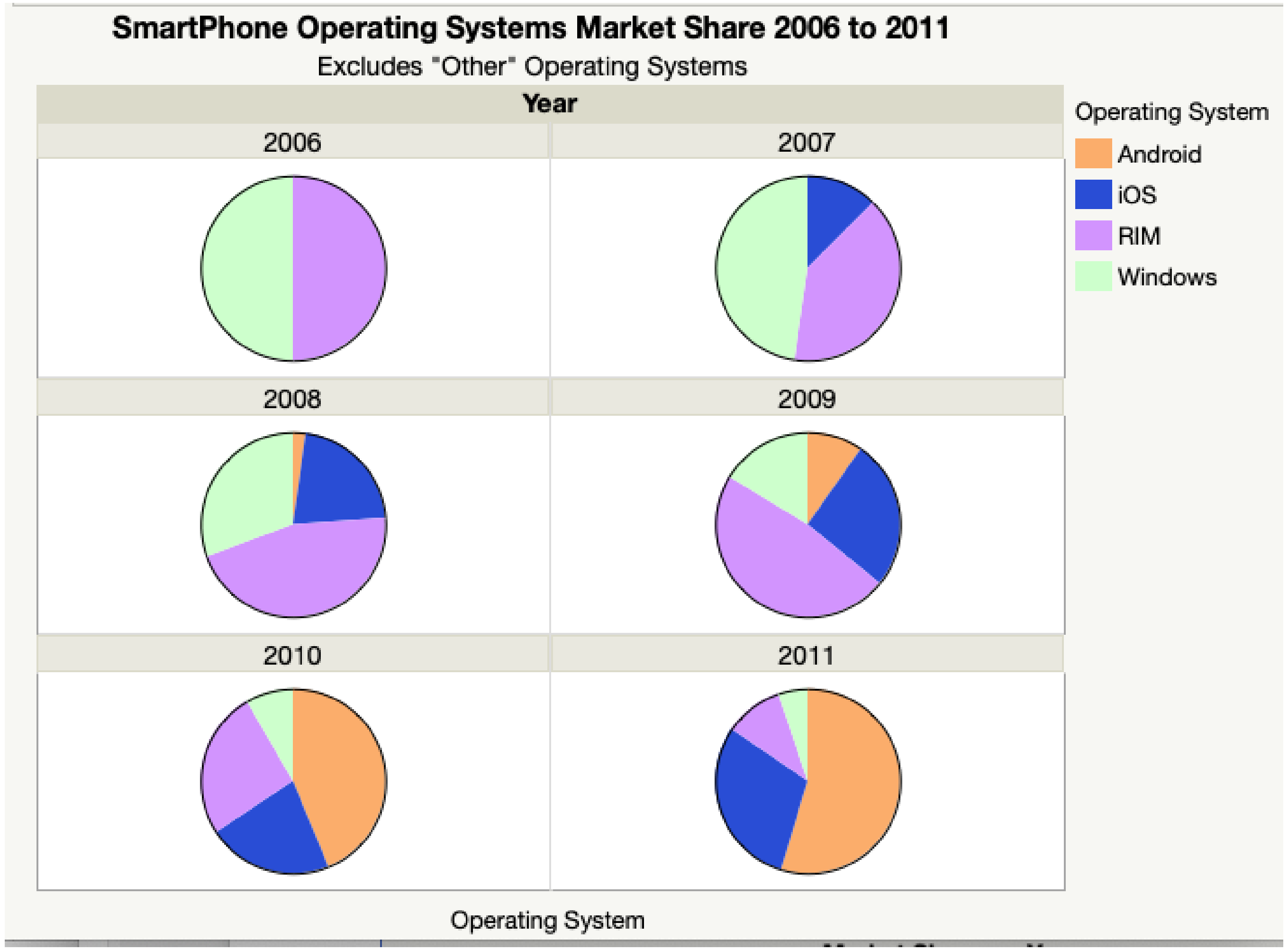
PIE CHARTS

- **A pie chart** is a circular statistical graphic divided into slices to illustrate numerical proportion.
- The **input data** you must provide is an array of **numbers**, where **each** numbers will be **mapped to one of the pie item**.



Source: https://www.python-graph-gallery.com/pie-plot-matplotlib-basic

PIE CHARTS



Multiple pie charts to show changes in parts-to-whole relationship

Source: https://www.jmp.com/en_us/statistics-knowledge-portal/exploratory-data-analysis/pie-chart.html

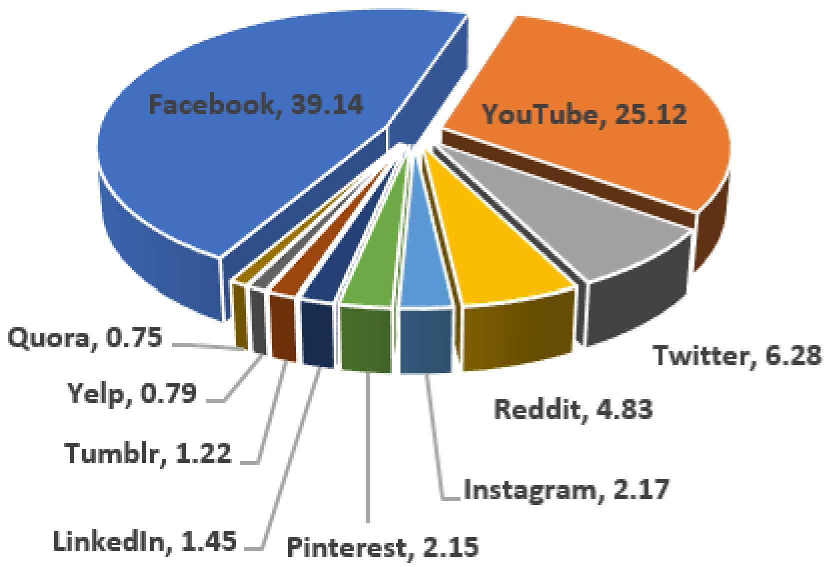
PIE CHARTS

- Some people suggest no to use Pie Charts
- Graphs of data should tell us about the quantities involved and help us to make accurate comparisons between these quantities. The quantities in each category should be easy to estimate and the category labels should be clear.
- *Pies and doughnuts fail because:*
  - Quantity is represented by slices; humans aren’t particularly good at estimating quantity from angles, which is the skill needed.
  - Matching the labels and the slices can be hard work.
  - Small percentages (which might be important) are tricky to show.

Source: https://www.jmp.com/en_us/statistics-knowledge-portal/exploratory-data-analysis/pie-chart.html

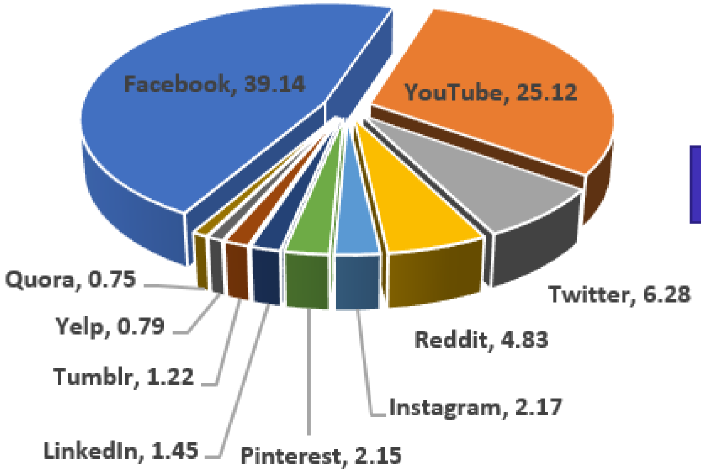
PIE CHARTS

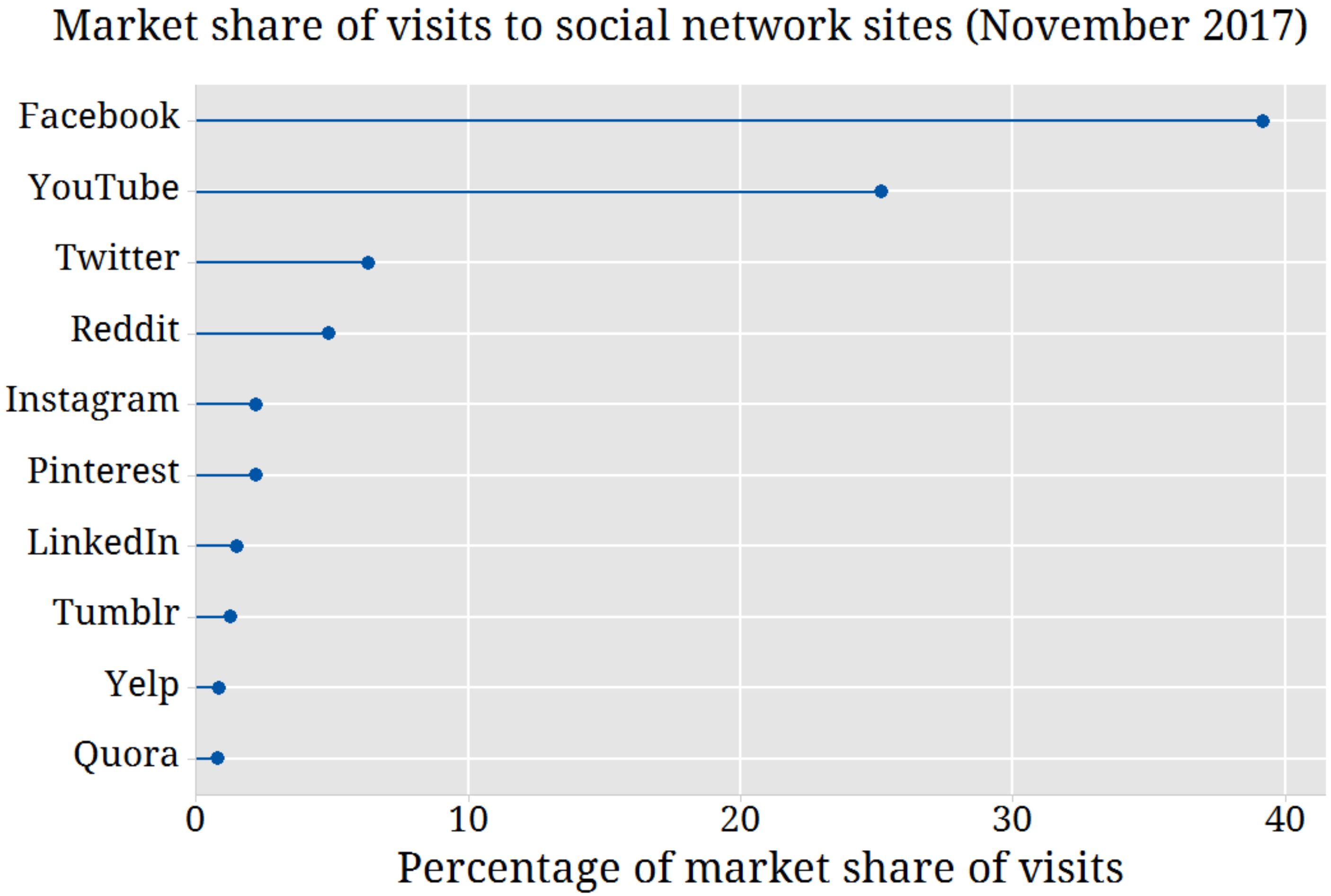
- You need to add the percentage to every slice.
- You need to directly label every slice.
- You have run out of colors for the slices.
- You decide to explode the chart to solve your first three problems.



https://scc.ms.unimelb.edu.au/resources/data-visualisation-and-exploration/no_pie-charts

PIE CHARTS



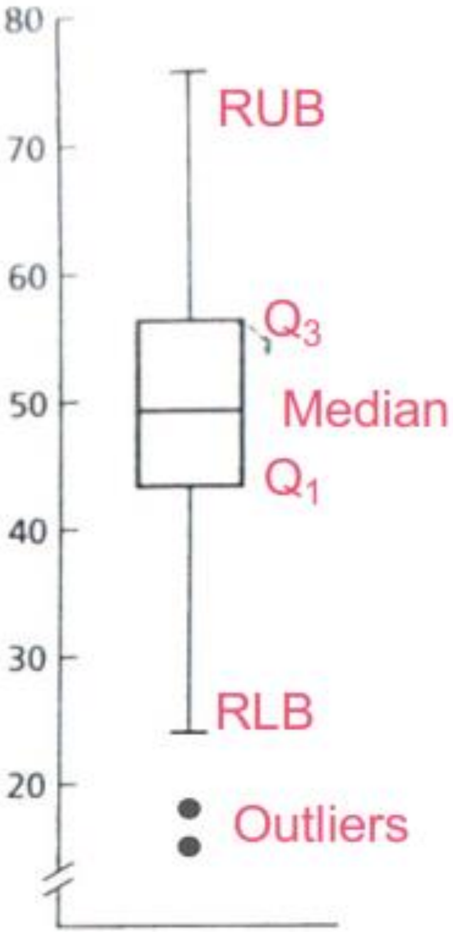


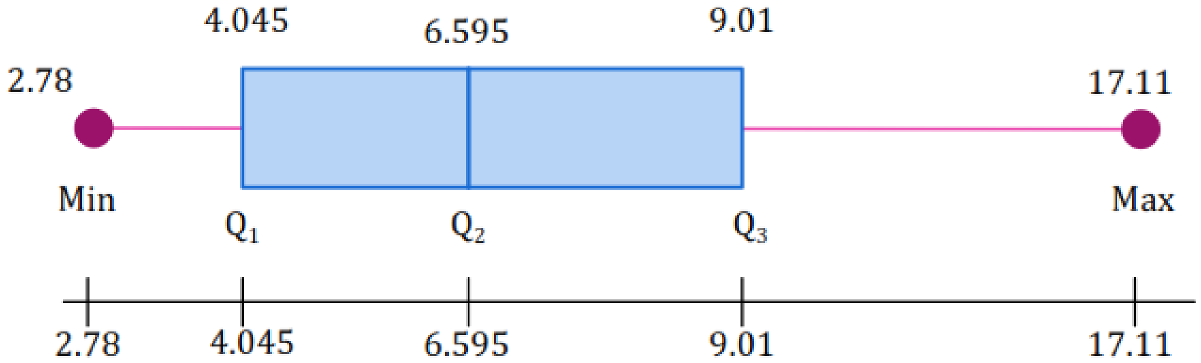
https://scc.ms.unimelb.edu.au/resources/data-visualisation-and-exploration/no_pie-charts

BOX PLOTS

- **A box plot is a way of statistically representing the distribution of given data through five main dimensions**:
  - The first dimension is minimum of the data.
  - The second dimension is first quartile.
  - The third dimension is median.
  - The fourth dimension is third quartile.
  - And the final dimension is maximum of the data.

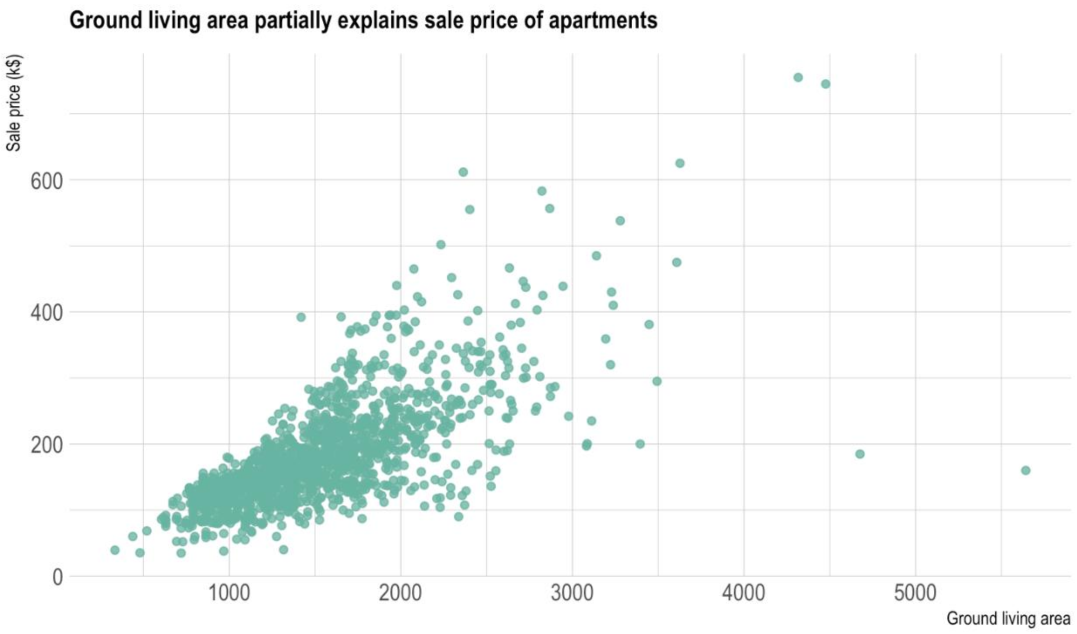
BOX PLOTS


80
70
60
50
40
30
20
RUB
Q3
Median
Q1
RLB
Outliers


4.045
6.595
9.01
2.78
17.11
Min
Q1
Q2
Q3
Max
2.78
4.045
6.595
9.01
17.11

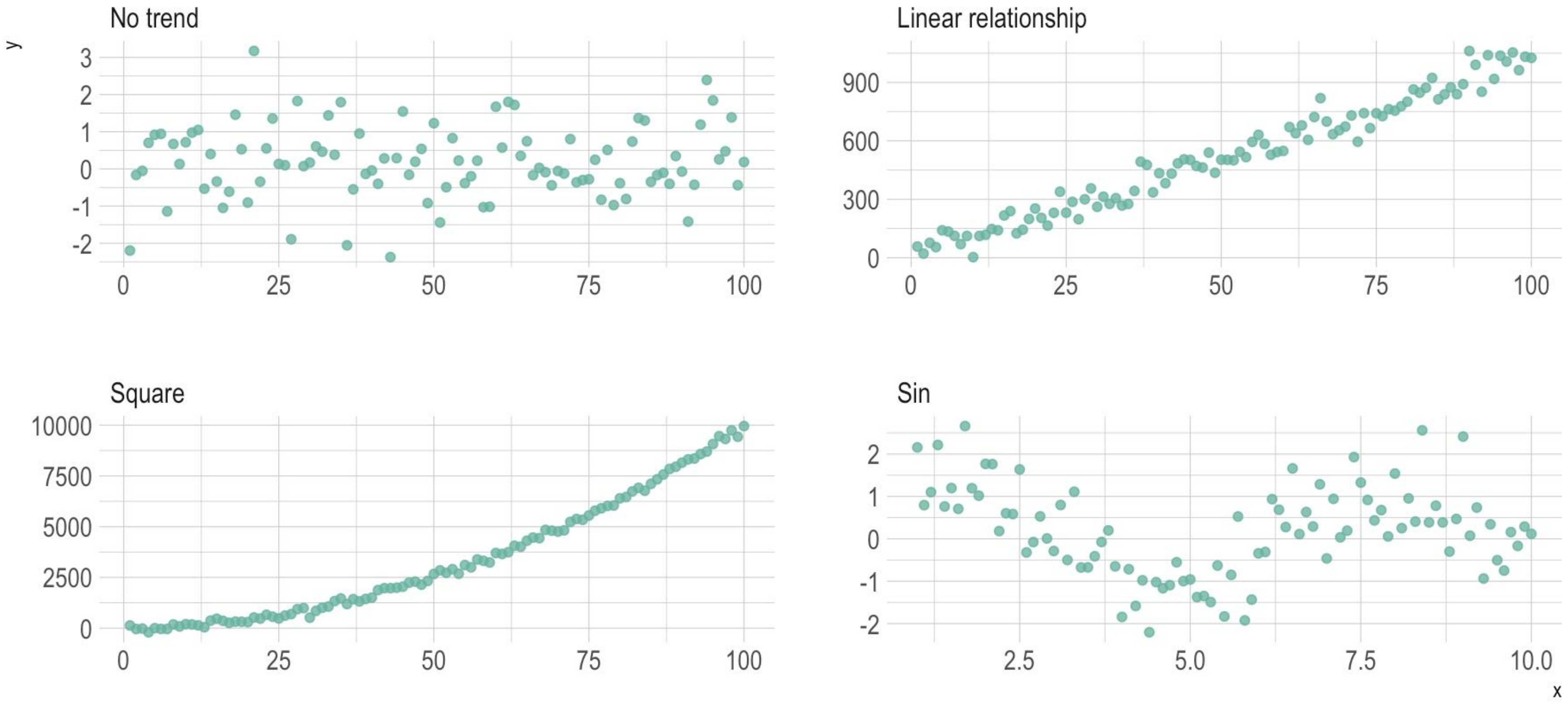
SCATTER PLOTS

- A scatter plot is a type of plot that displays values pertaining to typically two variables against each other.
- Usually it is a dependent variable to be plotted against an independent variable in order to determine if any **correlation** between the two variables exists.



https://www.data-to-viz.com/graph/scatter.html

SCATTER PLOTS

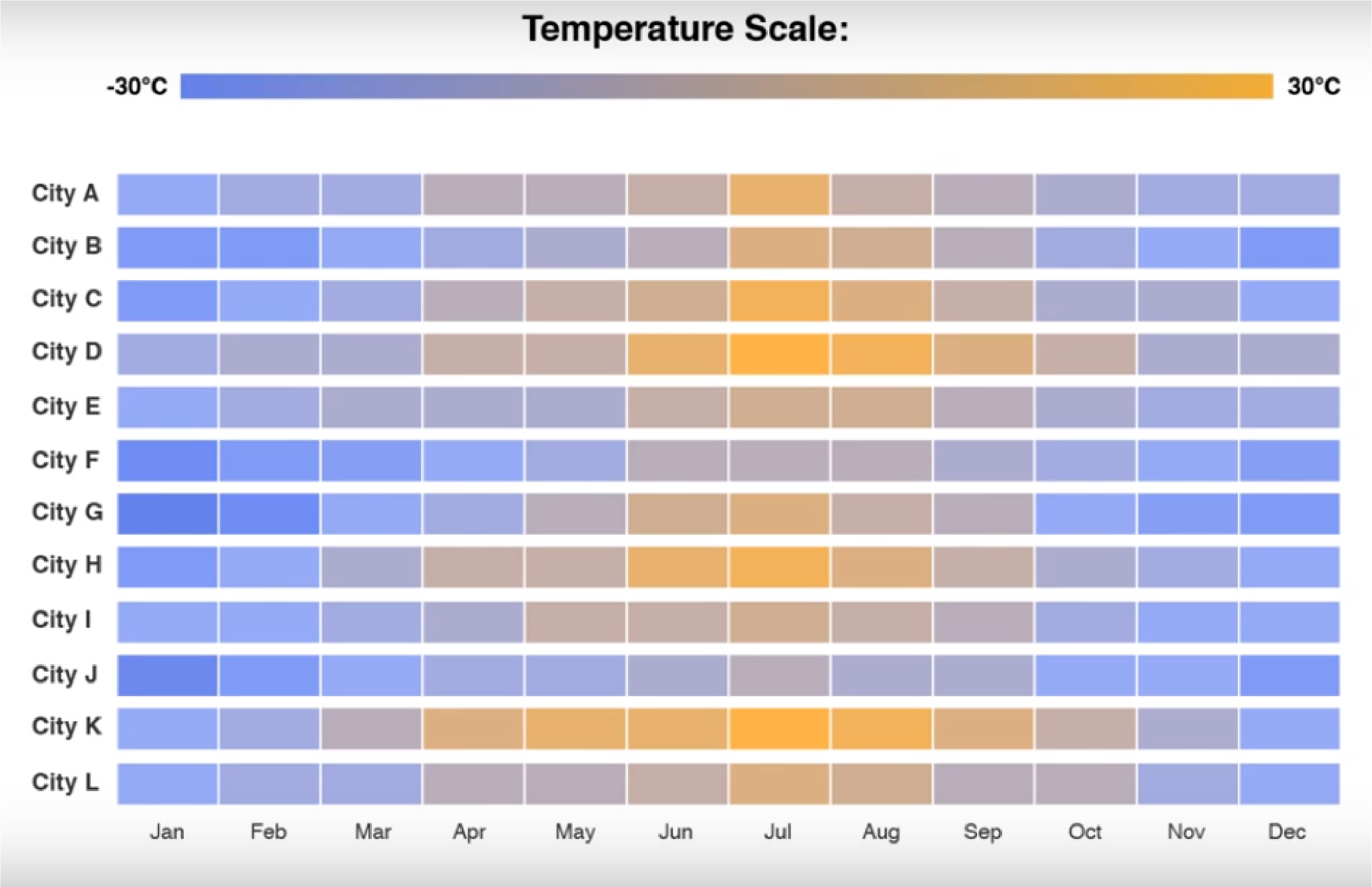


https://www.data-to-viz.com/graph/scatter.html

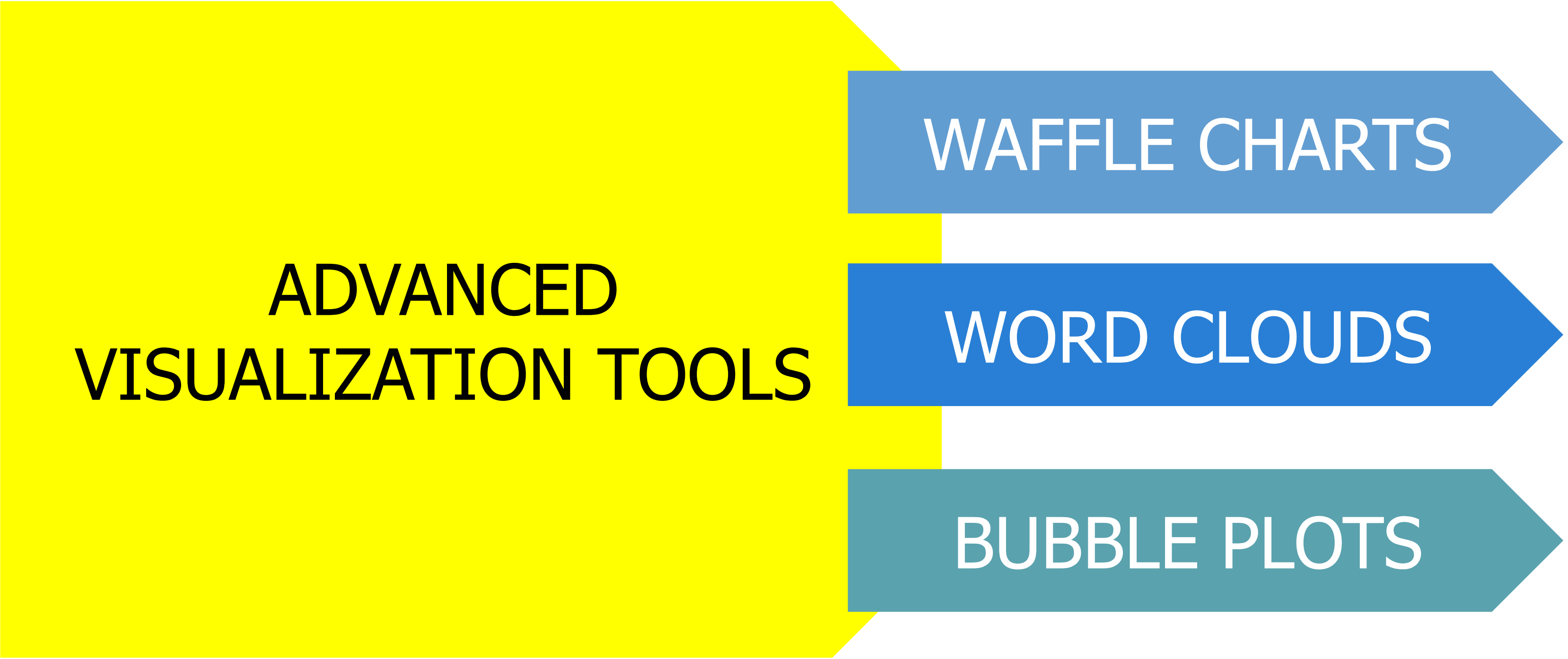
HEAT MAPS

- Heatmaps visualise data through variations in colouring.
- When applied to a tabular format, Heatmaps are useful for cross-examining multivariate data, through placing variables in the rows and columns and colouring the cells within the table.
- Heatmaps are good for showing variance across multiple variables, revealing any patterns, displaying whether any variables are similar to each other, and for detecting if any correlations exist in-between them.

HEAT MAPS



https://datavizcatalogue.com/methods/heatmap.html



ADVANCED VISUALIZATION TOOLS

WAFFLE CHARTS

WORD CLOUDS

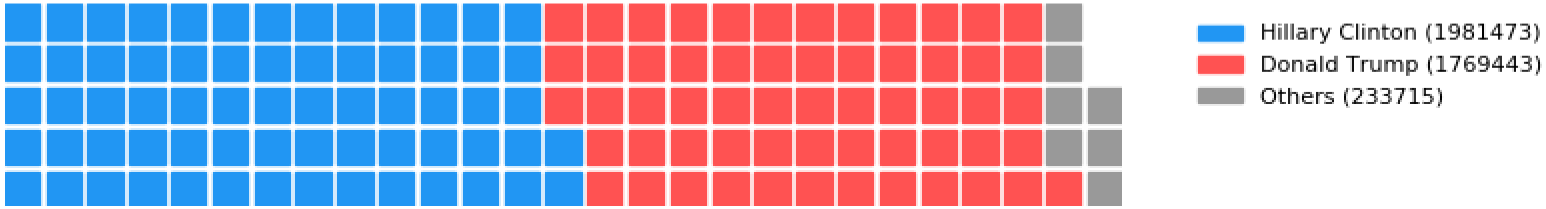
BUBBLE PLOTS

WAFFLE CHARTS

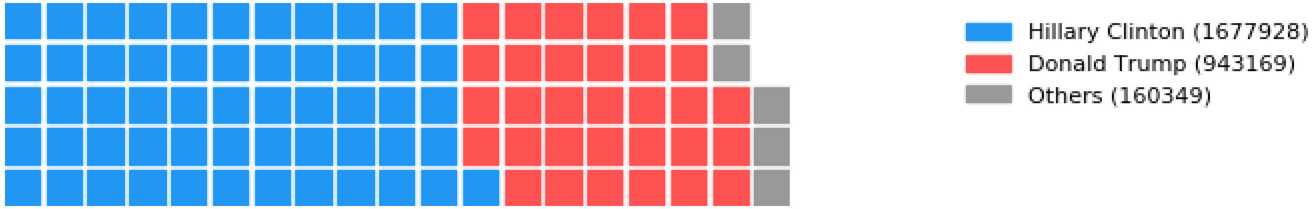
- **A Waffle Charts** is an interesting visualization that is normally created to display progress towards goals.
- As its name, it usually consists some **small squares** arranged in a M-by-N layout.
- The **squares are colored according** to the **proportions** you are aiming to visualize, **similarly to** how you would color different slices of a **pie chart**.

WAFFLE CHARTS

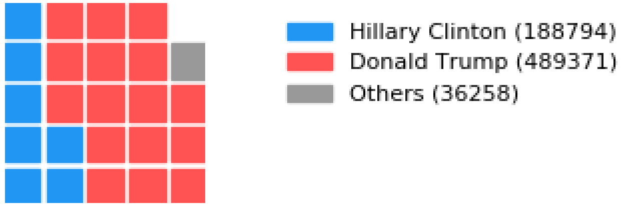
2016 Virginia Presidential Election Results



2016 Maryland Presidential Election Results



2016 West Virginia Presidential Election Results

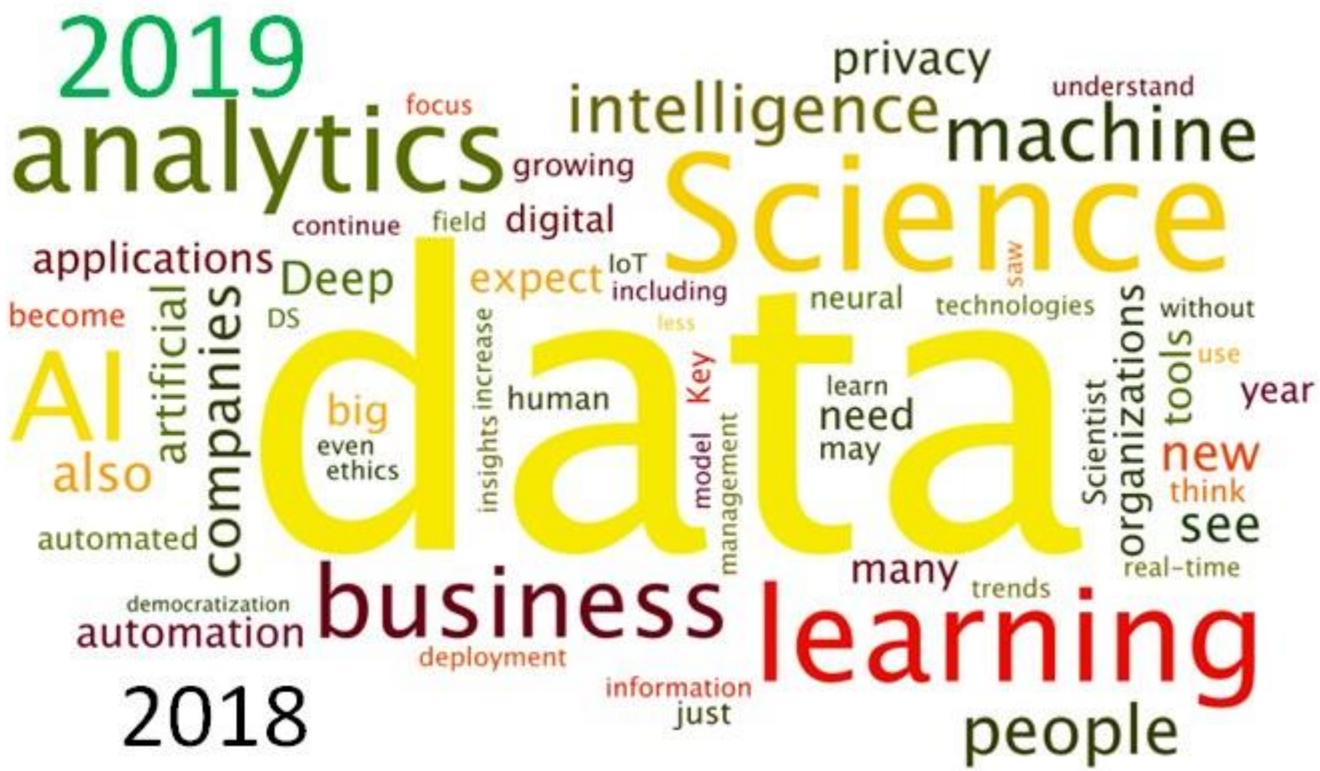


https://datascience.stackexchange.com/questions/57603/how-this-visualisation-was-made

WORD CLOUDS

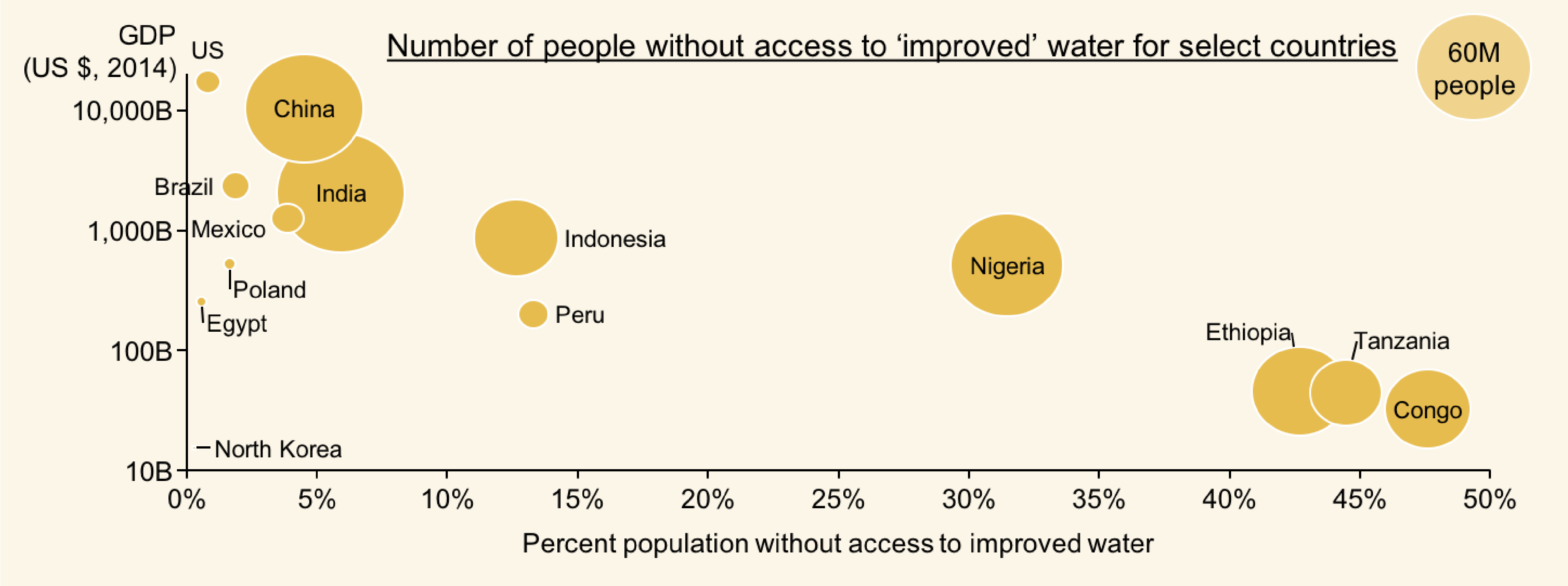
- **A word cloud** is simply a **depiction of the importance of different words in the body of text.**
- A word cloud works in a simple way; the **more a specific word appears** in a source of textual data **the bigger and bolder it appears** in the world cloud.
- Assuming that we didn't know anything about the content of these documents, a word cloud can be very **useful to assign a topic to some unknown textual data**.

WORD CLOUDS


2019
analytics
focus
intelligence
privacy
understand
machine
growing
Science
continue
field
digital
applications
Deep
expect
IoT
including
saw
become
DS
less
neural
technologies
without
AI
artificial
companies
data
big
even
ethics
insights
increase
human
Key
model
management
learn
need
may
Scientist
organizations
tools
use
year
new
think
see
also
automated
many
trends
real-time
democratization
automation
business
deployment
learning
information
just
2018
people

BUBBLE PLOTS

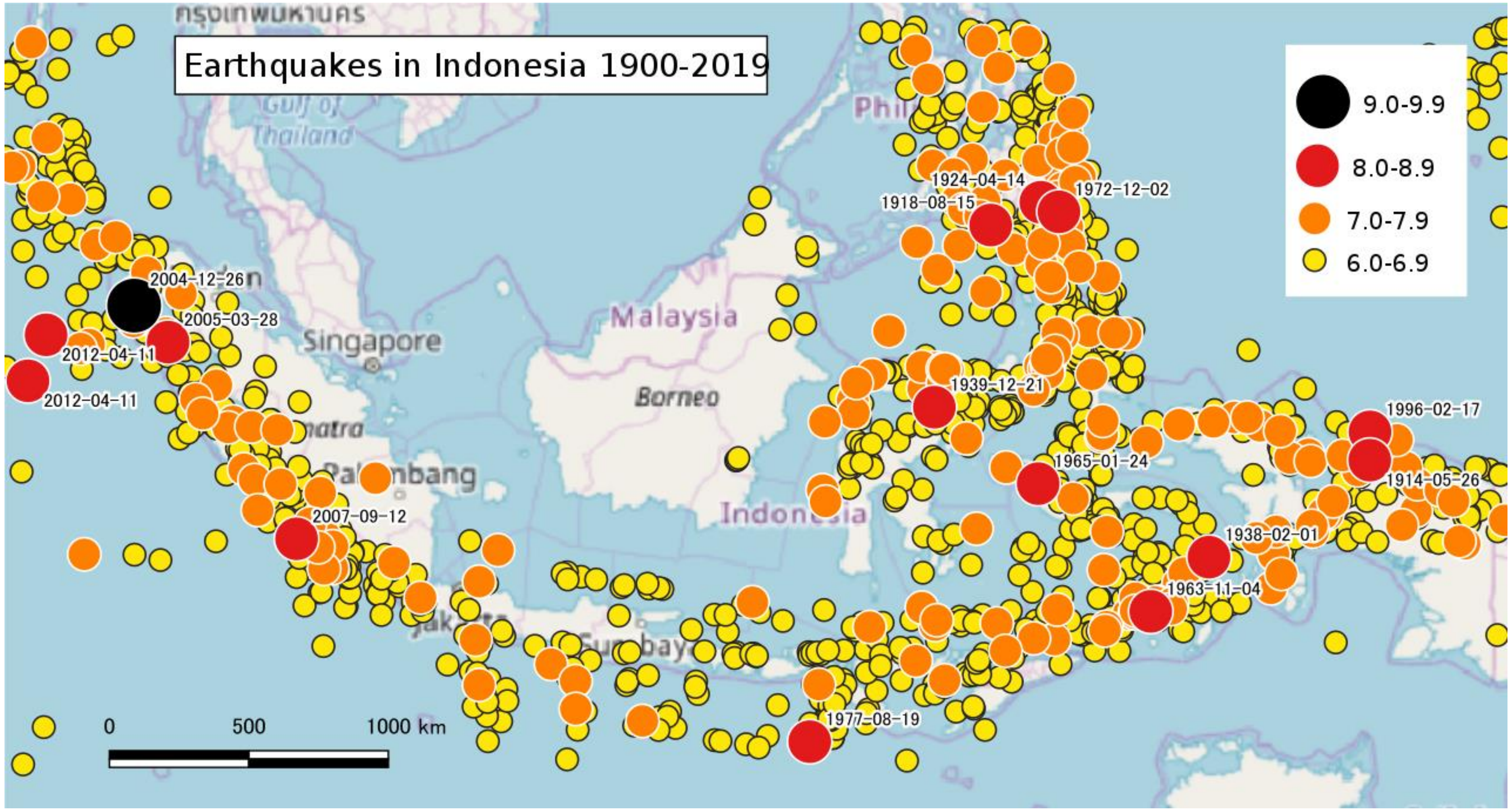
A bubble plot is a scatterplot where a third dimension is added: the value of an additional variable is represented through the size of the dots. You need 3 numerical variables as input: one is represented by the X axis, one by the Y axis, and one by the size.



Source: National Geographic (2015 data)

Note: 'Improved' water is water from pipes or wells that are protected from contamination

BUBBLE PLOTS



Bubble plots over Maps

https://upload.wikimedia.org/wikipedia/commons/thumb/8/8e/Map_of_earthquakes_in_Indonesia_1900-2019.svg/1280px-Map_of_earthquakes_in_Indonesia_1900-2019.svg.png

References & Credits

- Chirag Shah, Hands on Introduction to Data Science, Cambridge University Press, 2020
- Data Visualization from IBM Data Science Training Materials and cognitiveclass.ai
- Siti Aminah & Dhimas Arief Darmawan, Data Visualization, Salindia Mata Kuliah Data Sains Semester Genap 2020/2021, Fakultas Ilmu Komputer, Universitas Indonesia
- Fariz Darari, EDA & Visualization, Salindia Mata Kuliah Data Sains Semester Gasal 2020/2021, Fakultas Ilmu Komputer, Universitas Indonesia

- Gambar dan tangkapan layar hanya untuk kebutuhan penjelasan
  - Hak cipta tetap ada pada pemilik aslinya.

Wish You Success

☺