University of Science, VNU-HCM

Faculty of Information Technology

Introduction to Data Science

Data Visualization

=> Nằm từ bước hiểu Dl đến bước đã rút ra đc kết luận hữu ích

Presenter: Le Ngoc Thanh
lnthanh@fit.hcmus.edu.vn

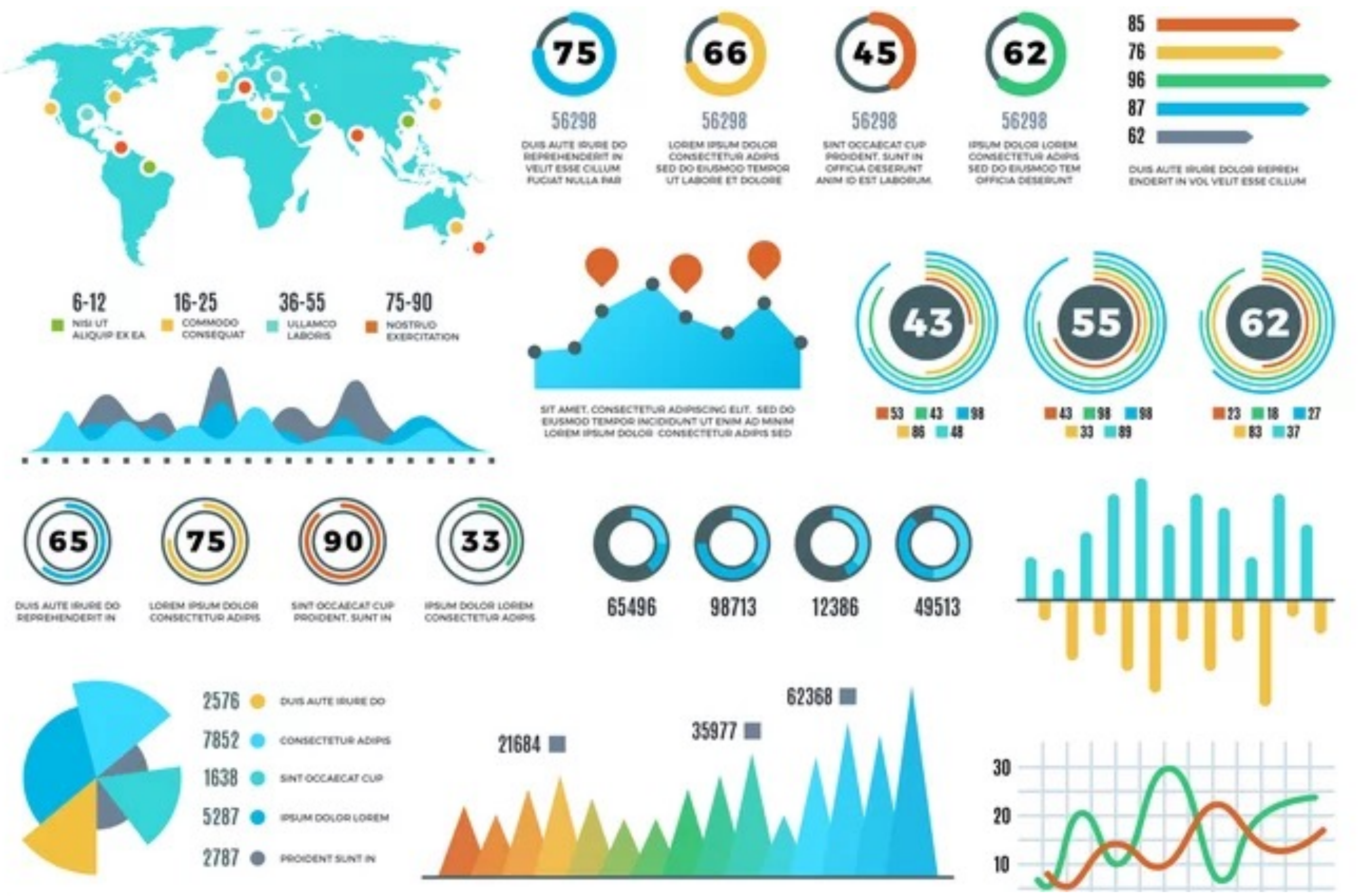
Ho Chi Minh City

Content

- Introduction
- Types of Visualization
  - Comparison Plots
  - Relation Plots
  - Composition Plots
  - Distribution Plots
  - Geo Plots

What is Data Visualization?

- Data visualization is the graphical representation of information and data.

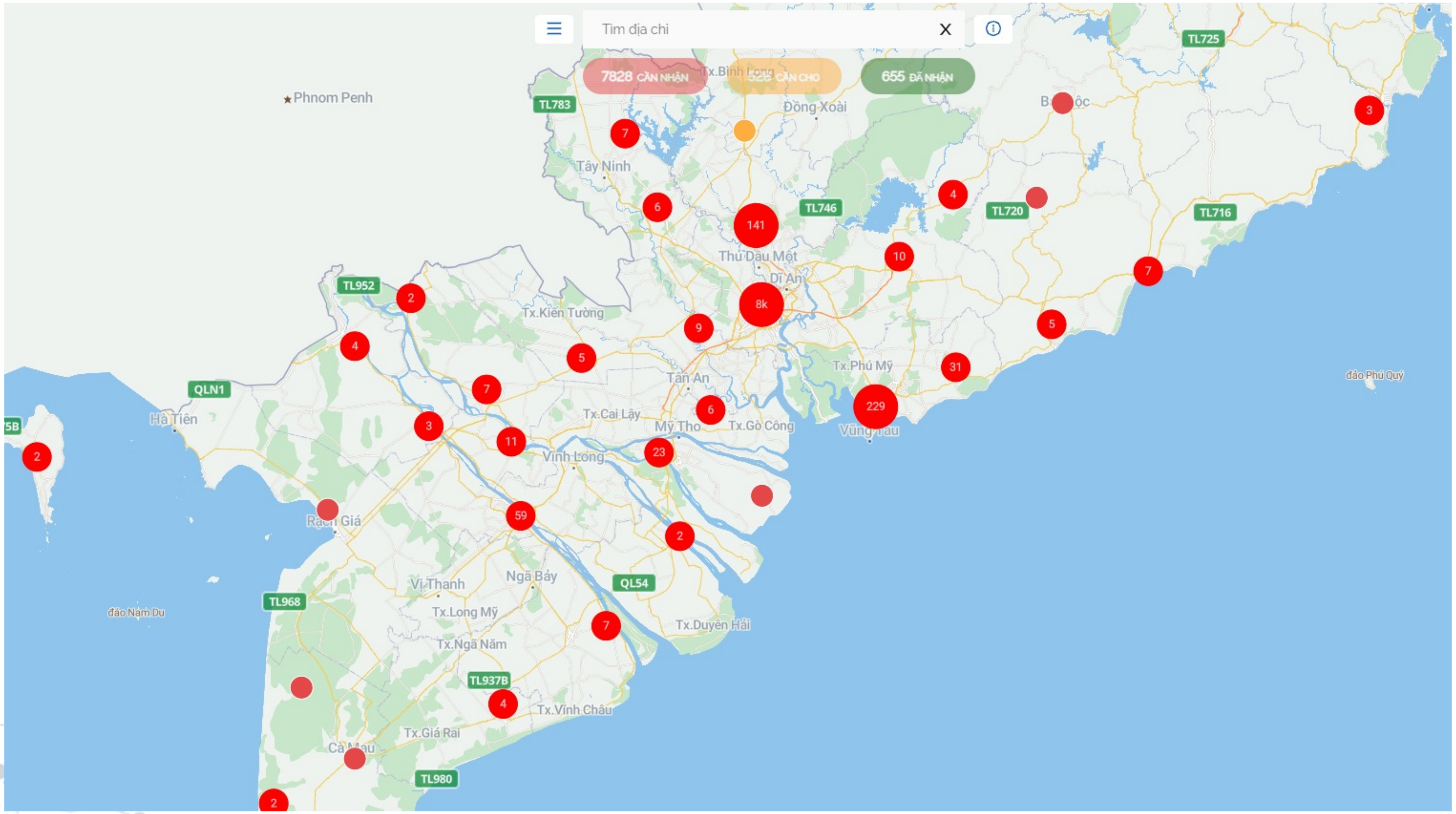


Why is Data Visualization important?

- Visual data is very easy to understand compared to data in any other form.
  - Our brains process images at a rapid pace, according to an MIT study.

Example of Data Visualization

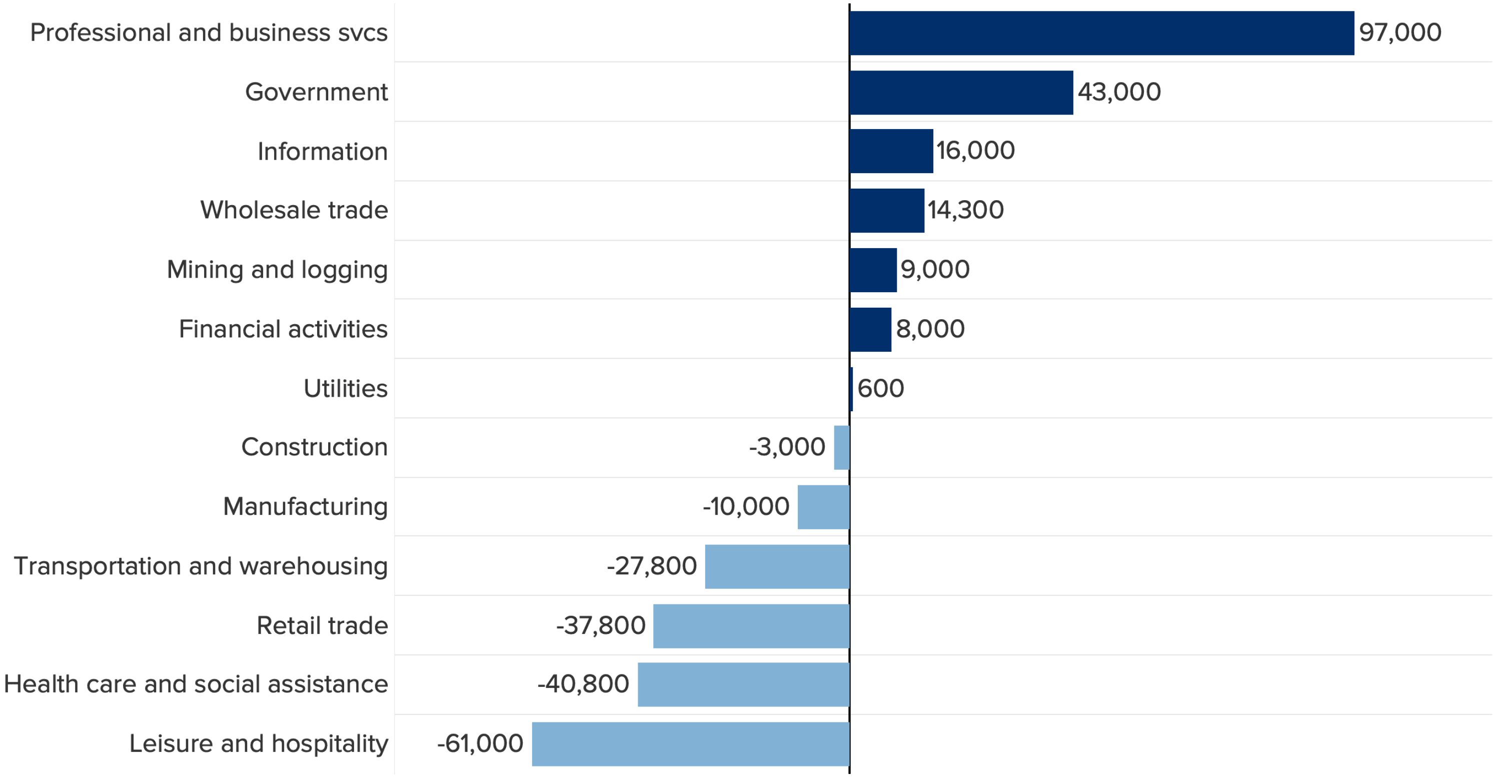
- Sosmap.net visualizes the covid infected cases with map



Example of Data Visualization

- The visual from CNBC uses a bar graph to visualize the industry-by-industry employment changes in the January 2021 jobs report.

January jobs one-month net change

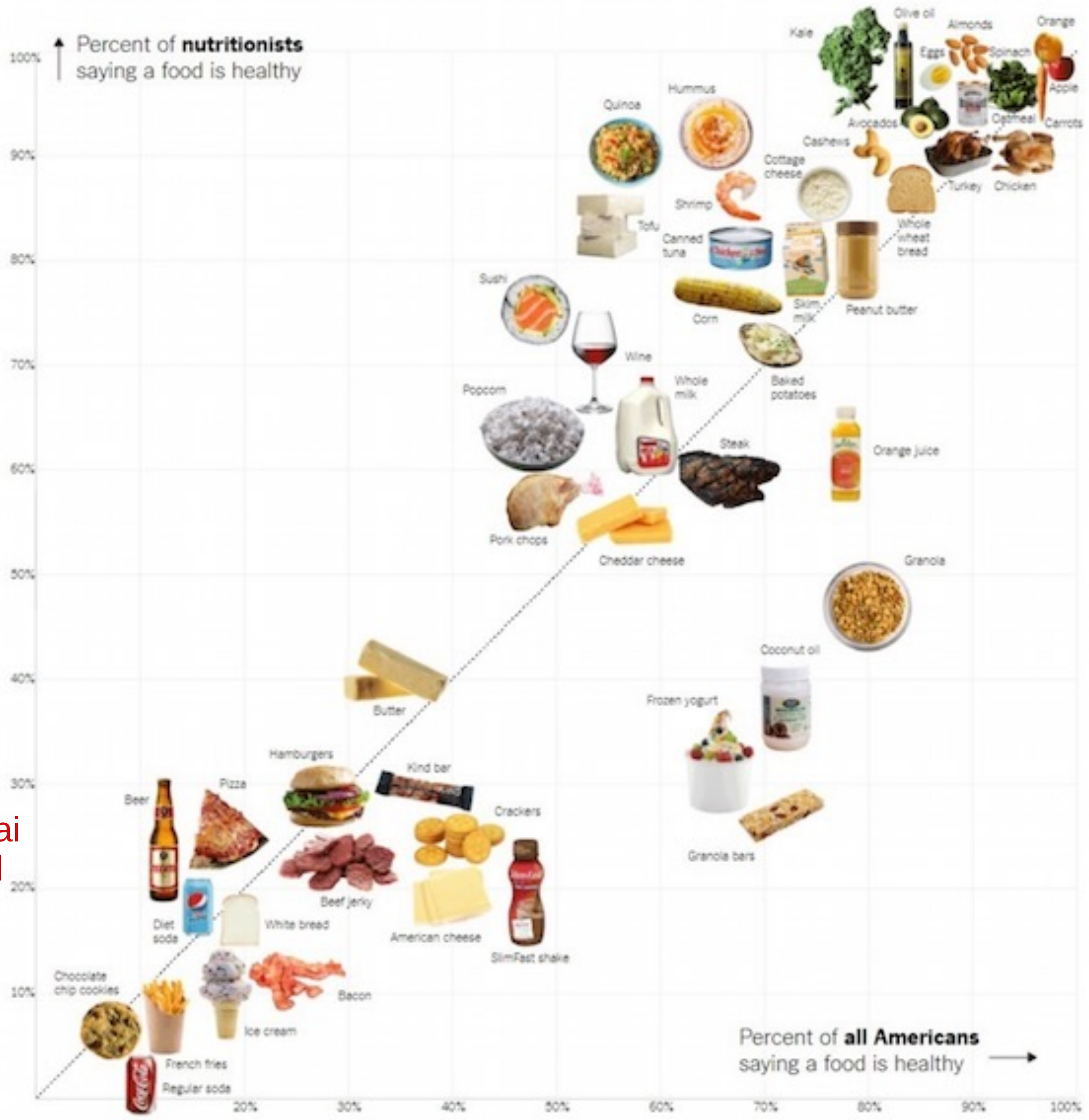


SOURCE: Bureau of Labor Statistics

CNBC

Example of Data Visualization

- The NY Times uses scatter plots to explain about healthy food



Pros:
+ dễ hiểu DL
+ dễ trao đổi và chia sẻ
+ dễ nhớ
+ Storytelling can be done using dashboard and animations
+ Có thể nhìn ra đc mối liên kết DL

Cons:
+ Thiếu sự chi tiết
+ Phải có kiến thức về biểu đồ mới hiểu đc
+ Chỉ đưa ra infor ước lượng chứ ko chính xác => có thể đưa ra dự đoán sai
+ Bias => có thể thiên vị vì góc nhìn or kinh nghiệm/quan điểm của ng visual
+ sử dụng sai biểu đồ
+ nhìn sai thông điệp trog data

Features of Visualizations

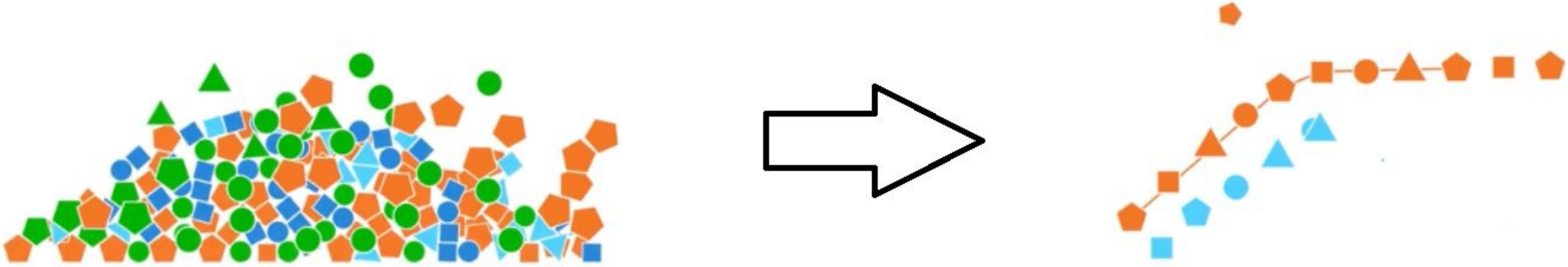
- All of data visualizations include the following features :
  - Indicators: They highlight the most important information.
  - Simplicity: The information is clear. The reader understands the information at hand immediately. => rõ ràng
  - Brevity: The message is short and clear, and no unnecessary information is visible. => ngắn gọn, súc tích, infor ko cần thì ko đưa vào
  - Originality: types of data are collected and displayed in a way that offers readers a new perspective on the subject. => cung cấp cho ng xem khía cạnh mới về DL
  - Colour: to draw the reader’s attention to the most important pieces of information, clear and easy-to-understand color palettes are used.

    => thu hút sự chú ý của ng đọc đến infor qtrong nhất bằng màu sắc

=> đưa DL từ ban đầu rồi ánh xạ or lm ccahs nào đó thành dạng dễ tiêu thụ

Data Wrangling

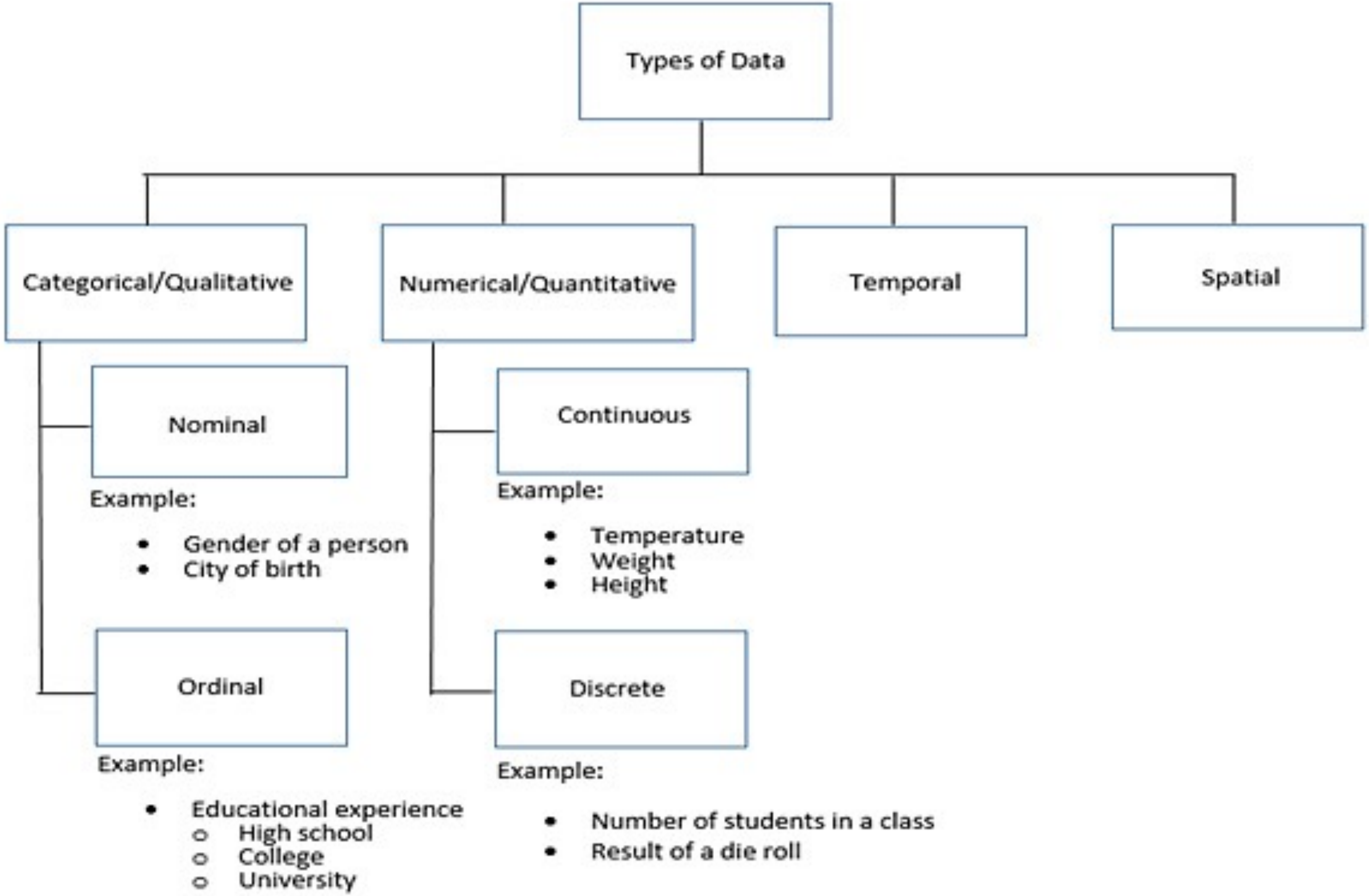
- **Data wrangling**—also called **data munging**—is the process of transforming and mapping data from one "raw" data form into the format that is convenient for the consumption of data.



Tools and Libraries for Visualization

- Non-coding tool:
  - Tableau
  - Power BI
- Coding (Python, Matlab, R):
  - Python libraries:
    - Matplotlib
    - Seaborn
    - Geospatial => hỗ trợ trực quan về mặt địa lý, bản đồ
    - Bokeh => tạo ra sự tương tác : zoom out, zoom in, kéo thả

Types of Data



Types of Data

- Categorical/Qualitative
  - Nominal

    Example:

    - Gender of a person
    - City of birth
  - Ordinal

    Example:

    - Educational experience
      - High school
      - College
      - University
- Numerical/Quantitative
  - Continuous

    Example:

    - Temperature
    - Weight
    - Height
  - Discrete

    Example:

    - Number of students in a class
    - Result of a die roll
- Temporal
- Spatial

Content

- Introduction
- **Types of Visualization**
  - Comparison Plots => nhóm lược đồ so sánh
  - Relation Plots
  - Composition Plots
  - Distribution Plots
  - Geo Plots

Comparison Plots

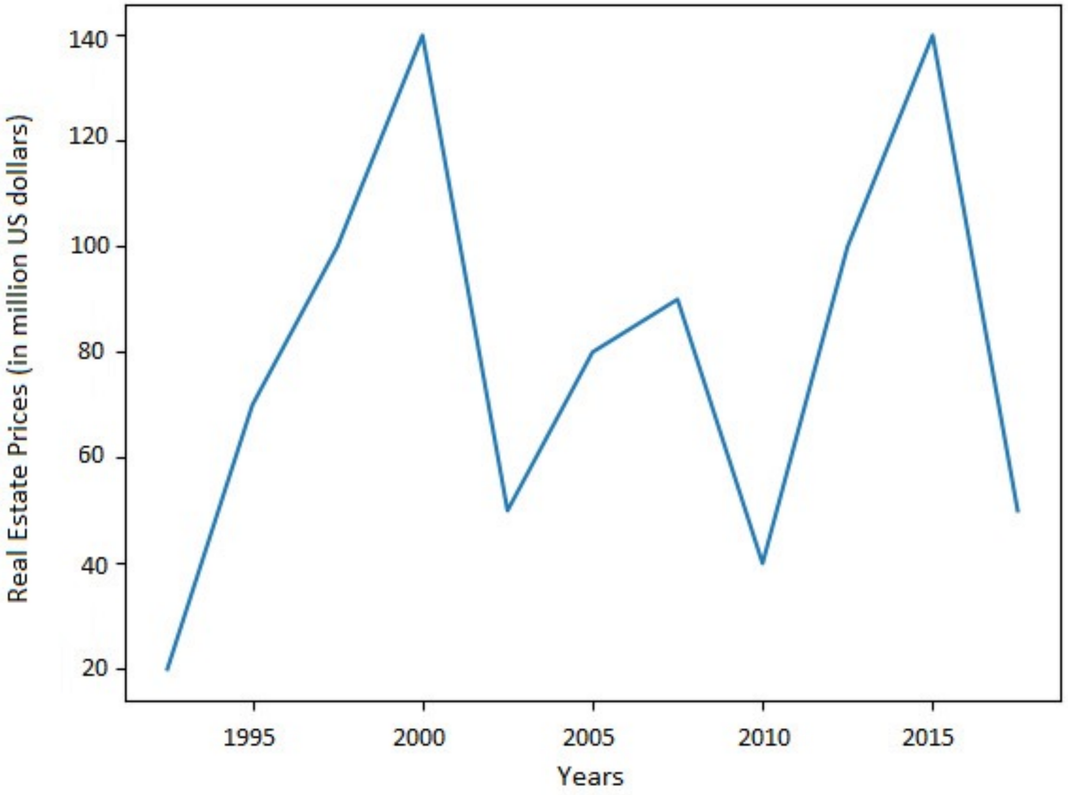
- Comparison plots include charts that are well-suited for comparing multiple variables or variables over time.
  - Line chart
  - Bar chart
  - Radar chart

Line Chart

=> hướng đến trend, tìm ra quy luật/ xu hướng
=> Mục tiêu: so sánh sự thay đổi của biến theo thời gian

=> đối vs những khoảng tg ngắn or chu kì ngắn thì ko dùng line chart mà dùng bar chart

- Line charts are used to display quantitative values over a continuous time period and show information as a series.

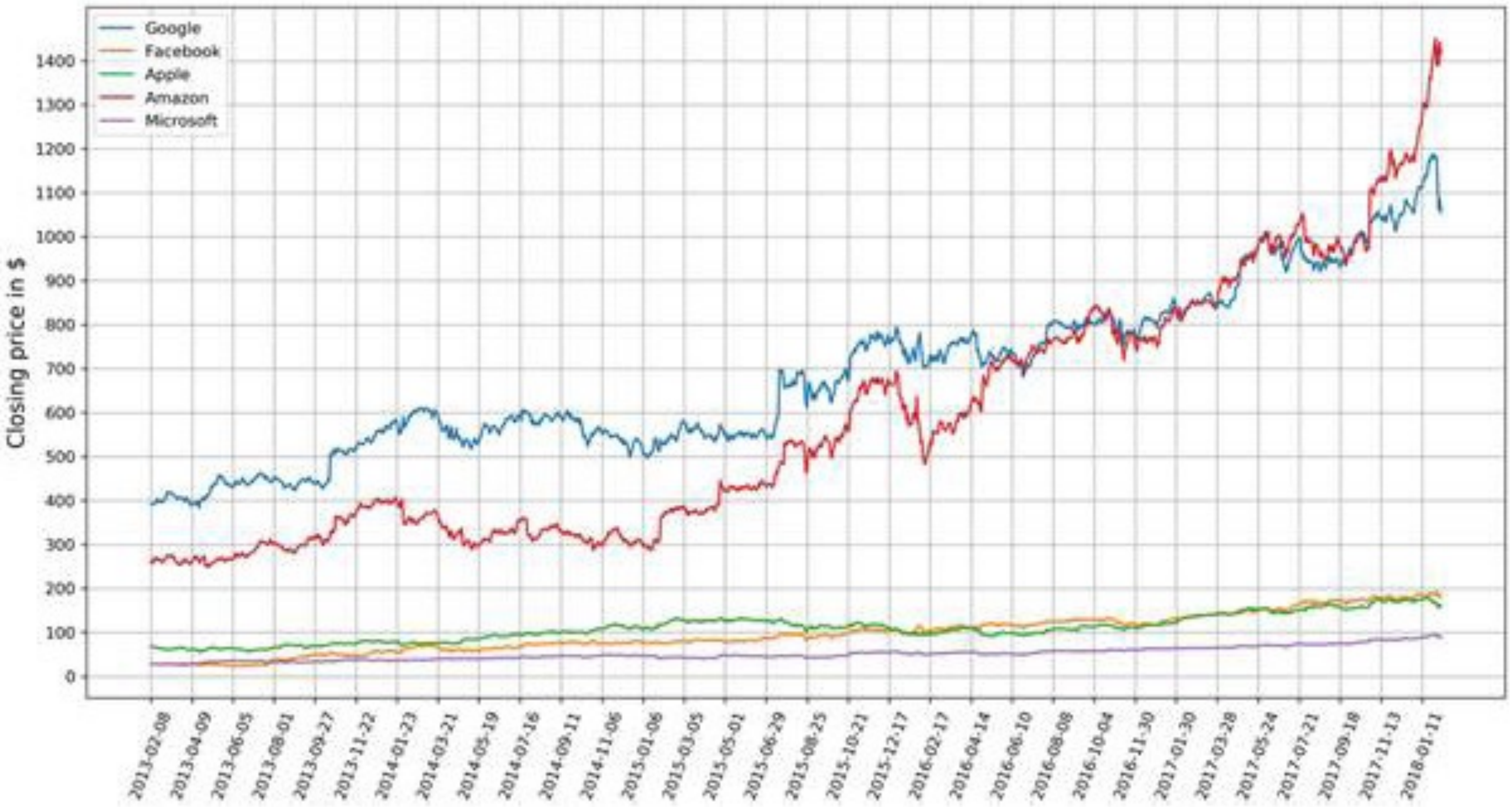


Line Chart

- Uses:
  - Line charts are great for comparing multiple variables and visualizing trends for both single as well as multiple variables, especially if your dataset has many time periods (roughly more than ten).
  - For smaller time periods, vertical bar charts might be the better choice.

Line Chart Example

- Compares the stock-closing prices for Google, Facebook, Apple, Amazon, and Microsoft.



Line Chart Pratices

=> mất tính đơn giản

=> Câu hỏi: cải thiện: chia nhóm, giới hạn đg trog nhóm. có thể dùng TB or max, min

- Design practices:
  - Avoid too many lines per chart
  - Adjust the scale so that the trend is clearly visible

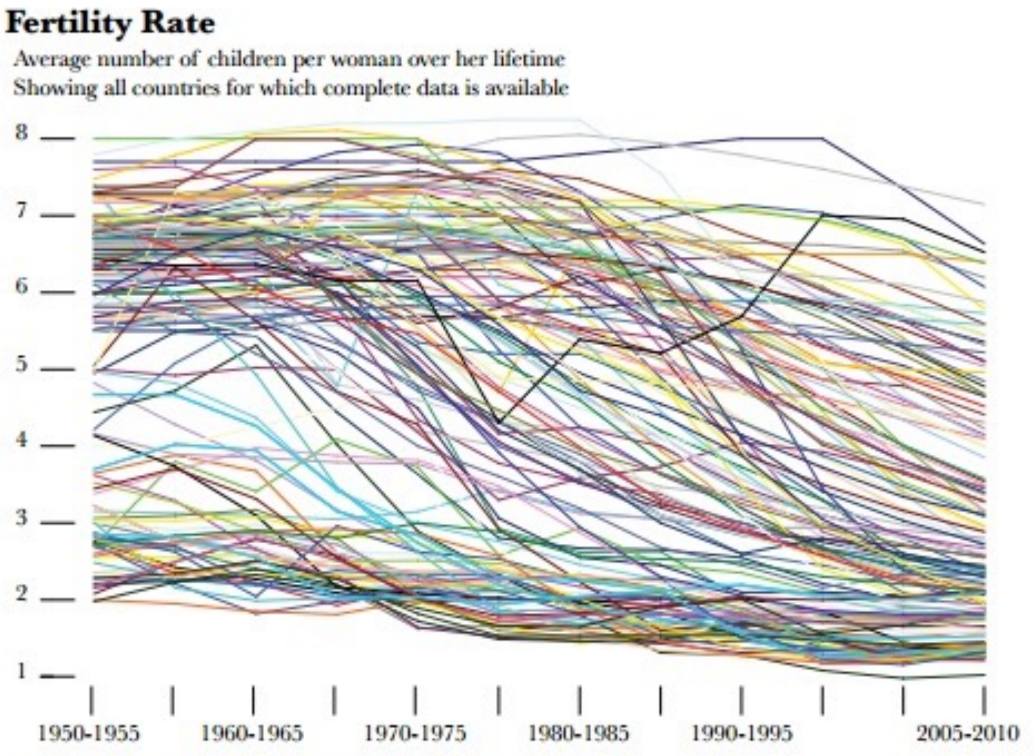


Figure 1.5 Too many lines obscure the message.

Bar Chart

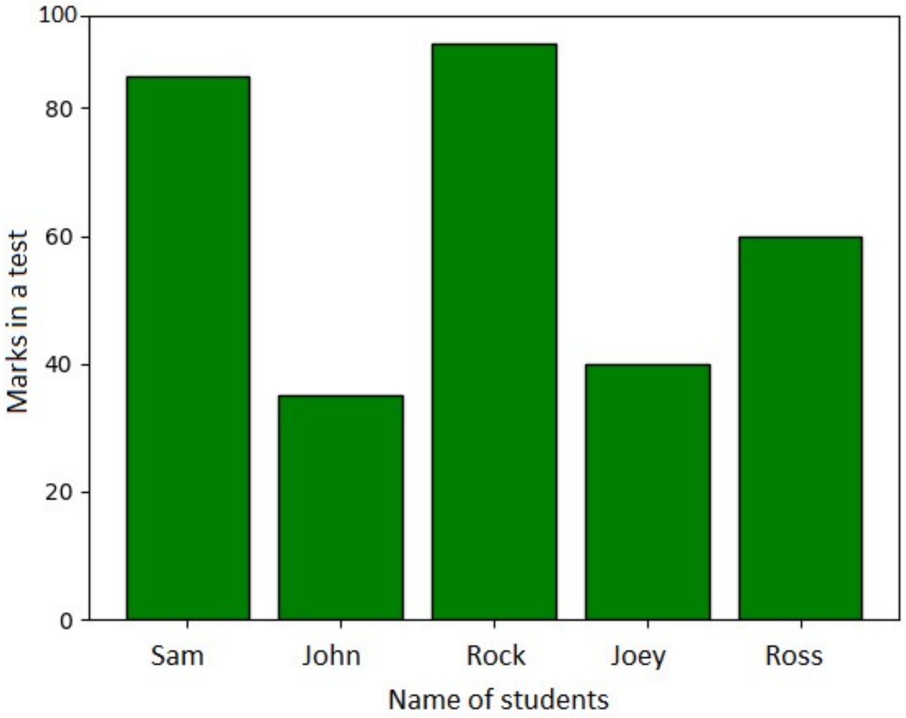
- The bar length encodes the value. thanh đứng
- There are two variants of bar charts: vertical bar charts and horizontal bar charts. thanh ngang
  - While they are both used to compare numerical values across categories, vertical bar charts are sometimes used to show a single variable over time.
  - Don't confuse vertical bar charts with histograms. Bar charts compare different variables or categories, while histograms show the distribution for a single variable.

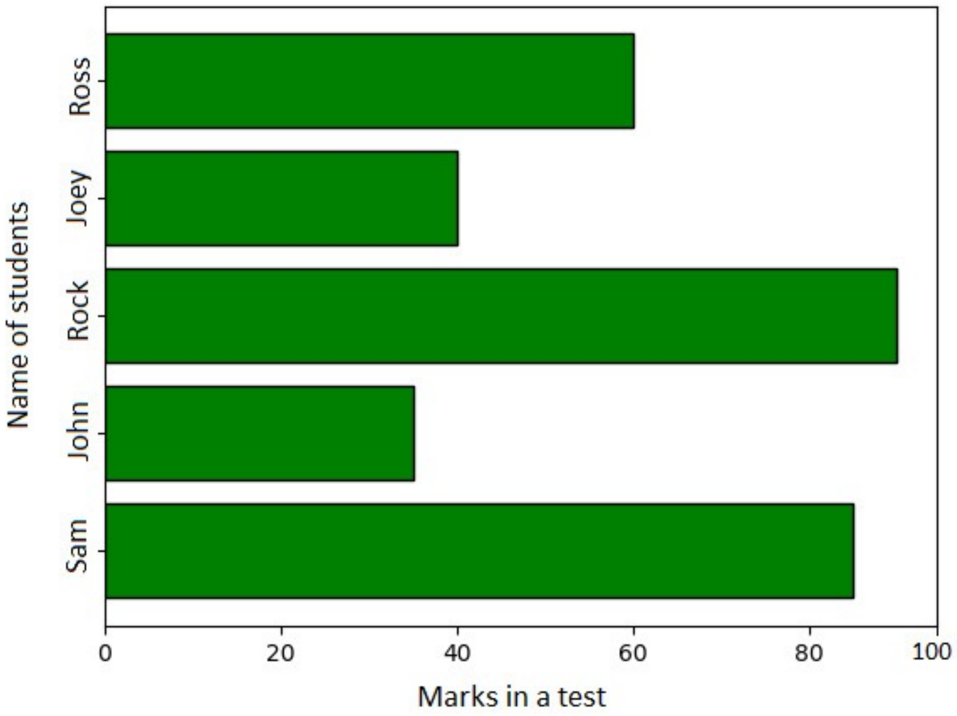
=> Mục tiêu của histogeam là thể hiện sự phân bố chứ ko phải so sánh như bar chart

Bar Chart Example

=> ko nên dùng line chart vì ở đây có thể swap lại và tạo ra nhìu đg khác nhau và dùng bar dễ so sánh hơn

- The marks out of 100 that five students obtained in a test





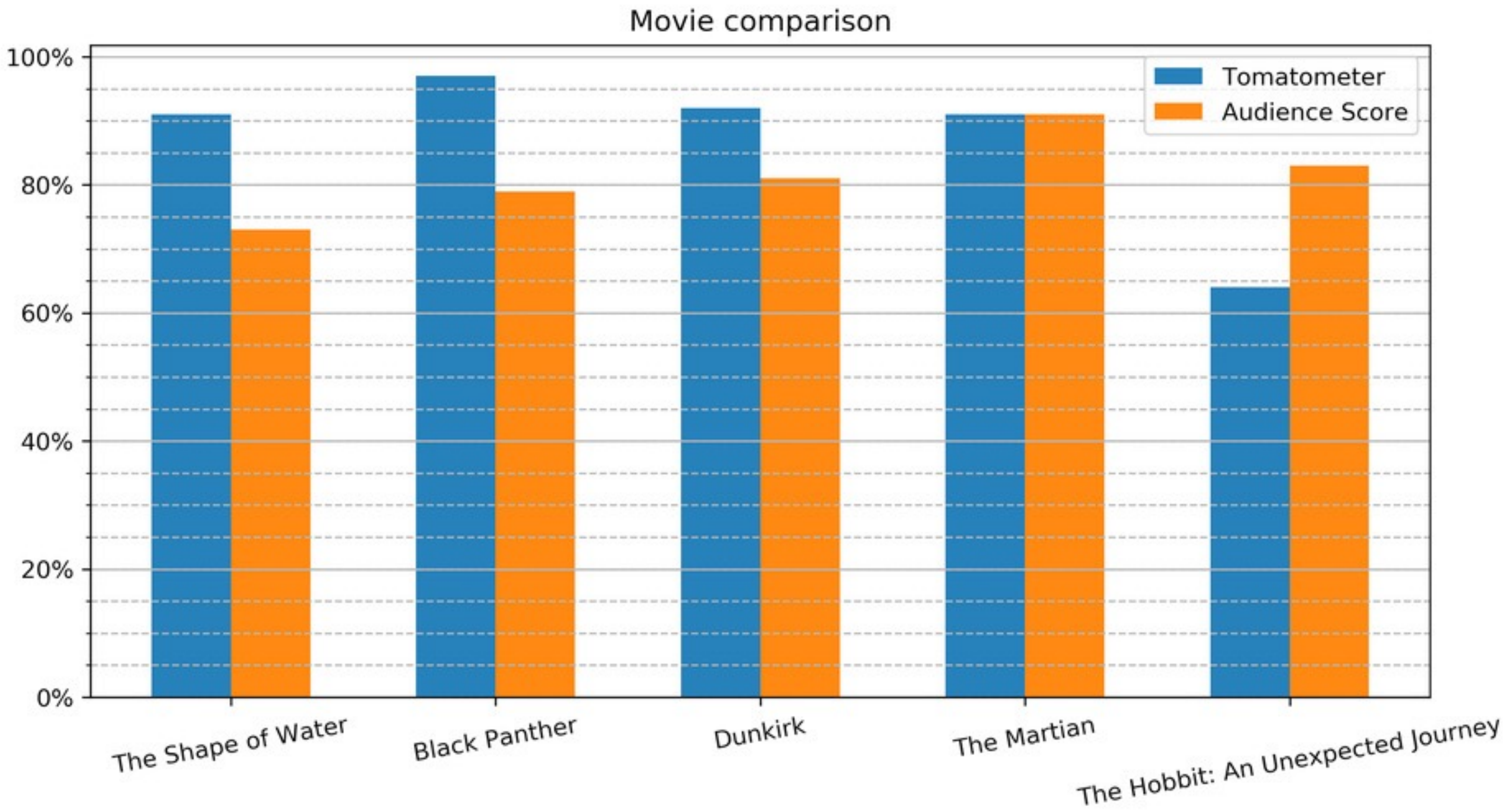
Dùng bd đứng: có ít giá trị

Dùng bd ngang: thể hiện hiều giá trị, nên sắp xếp tăng or giảm dần

Bar Chart Example

Exercise:
+ Điểm ko hợp lí: mục tiêu của họ ko có xu hướng , thiếu ghi chú
+ Nên dùng bar chart
+ Ta hoán vị thì tạo ra đg khác => ko dùng line chart

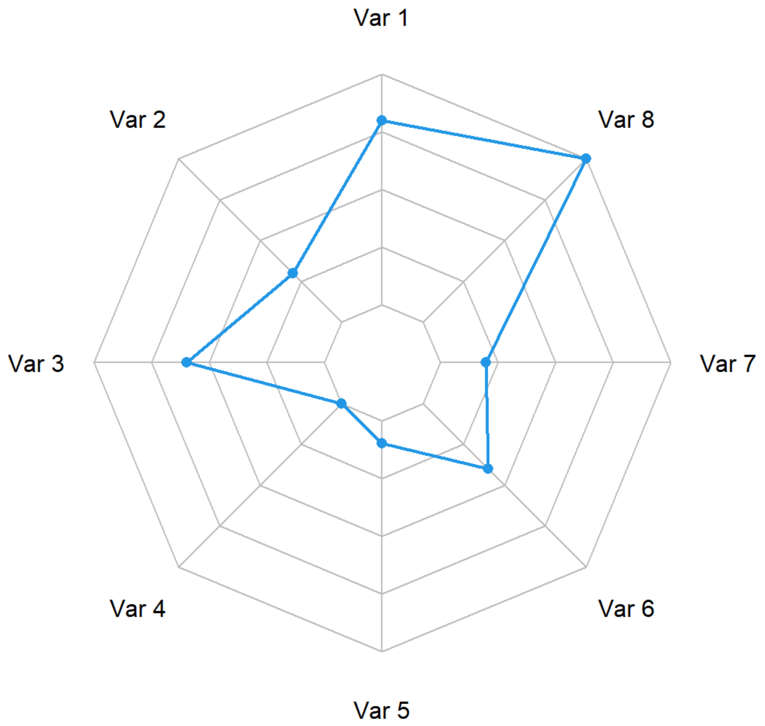
- The following diagram compares movie ratings, giving two different scores:



=> visual nhiều biến vs mỗi biến đc vẽ trên trục riêng của nó
=> mỗi trụ thể hiện 1 biến và vòng tròn thể hiện mức độ

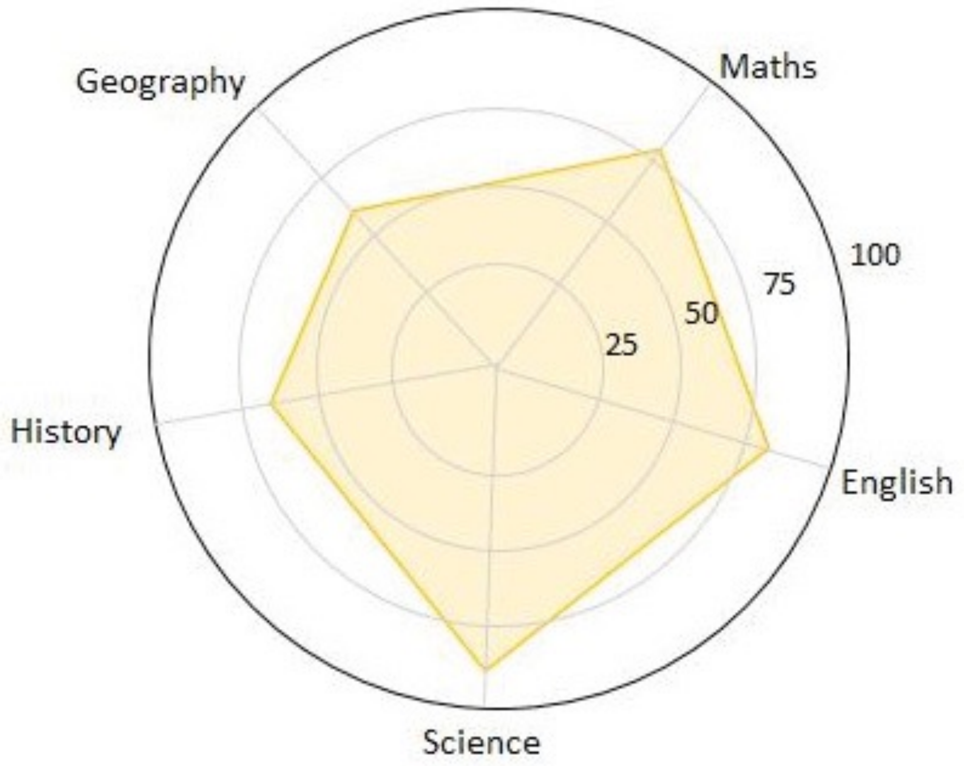
Radar Chart

- Radar charts, also known as spider or web charts, visualize multiple variables with each variable plotted on its own axis, resulting in a polygon.
  - All axes are arranged radially, starting at the center with equal distances between one another and have the same scale.



Radar Chart Example

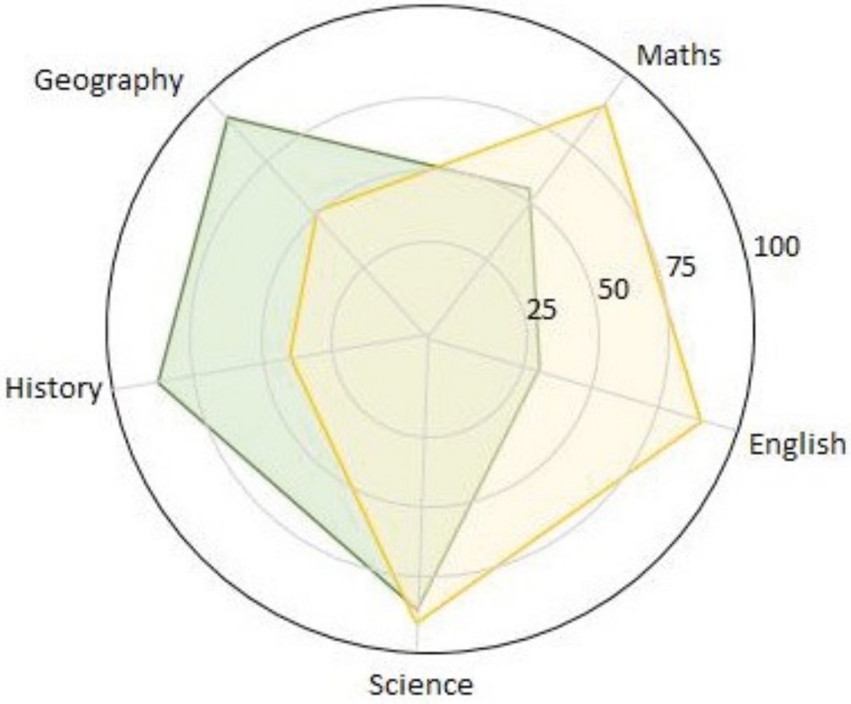
- The following radar chart displays data about a student (a single variable) scoring marks in different subjects.



=> dùng radar chart vì
+ vùng phủ: thiên hướng của sv là về khtn
+ có thể so sánh trên toàn bộ không gian

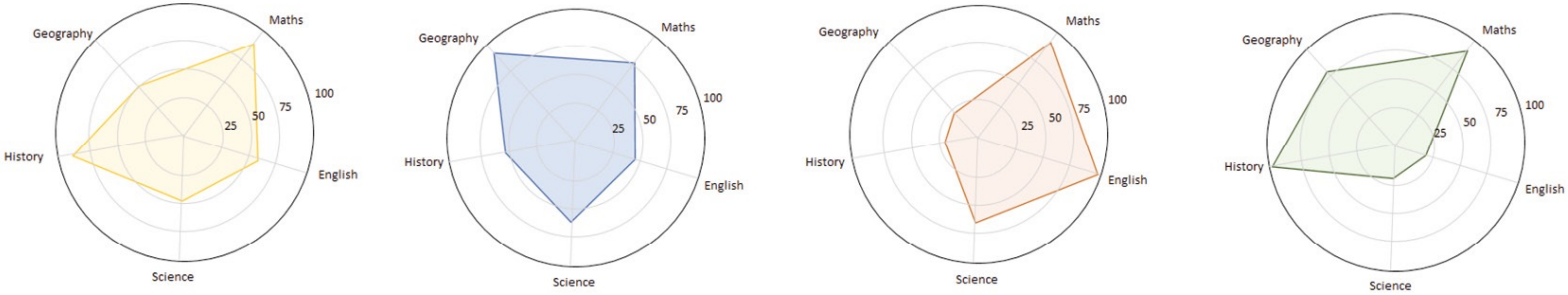
Radar Chart Example

- The following diagram shows a radar chart for two variables/groups:



Radar Chart Example

- Radar chart with faceting for multiple variables (multiple subjects)



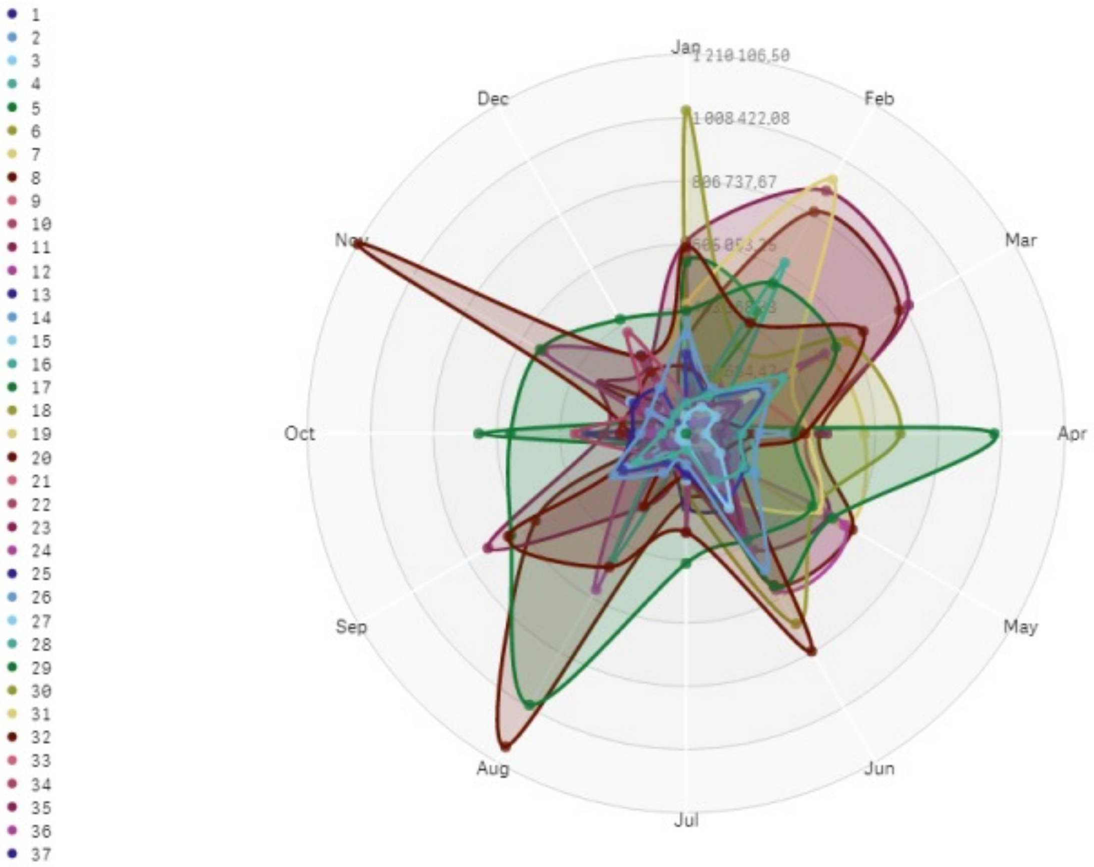
Radar Chart Pratices

Exercise 2:
+ Vẽ bar chart: quá nhìu cột, ko nhìn đc thế mạnh của từng bạn (thường 2-3 cột)

+ Vẽ radar chart: dễ so sánh và nhìn tổng thể => đánh giá tổng thể tốt

Improve: bổ sung thêm màu, xếp chồng lên nhau, thêm region

- Design practices:
  - Display ten factors or fewer on one radar chart to make it easier to read.



Content

- Introduction
- **Types of Visualization**
  - Comparison Plots
  - Relation Plots => nhóm lược đồ thể hiện sư tương quan
  - Composition Plots
  - Distribution Plots
  - Geo Plots

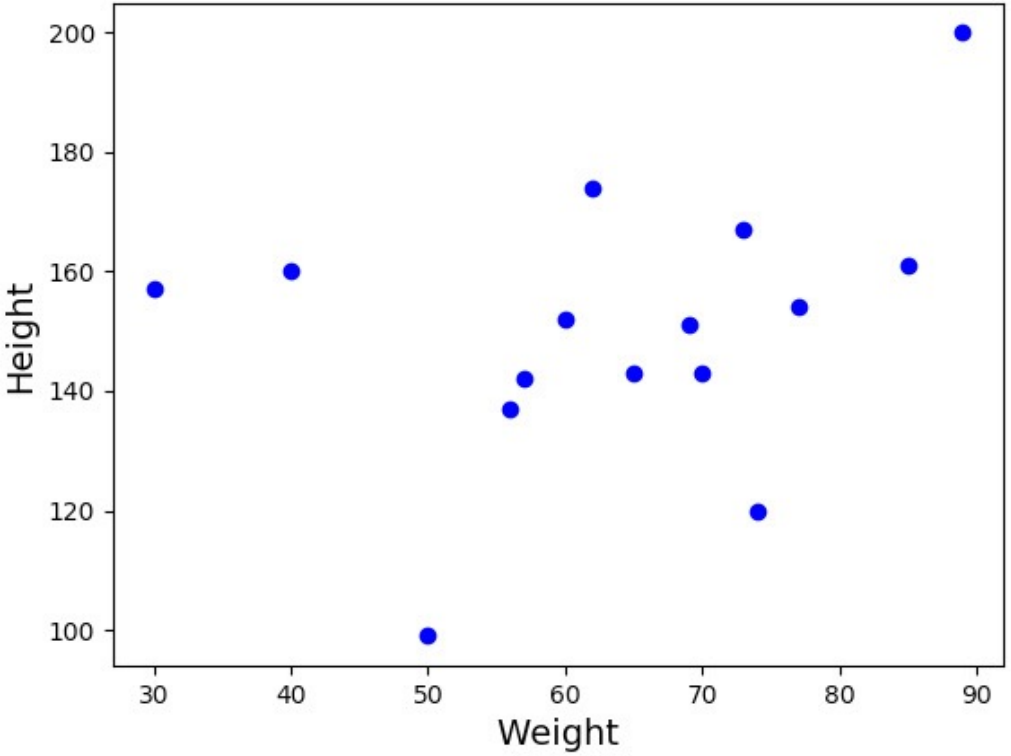
Relation Plots

- Relation plots are perfectly suited to show relationships among variables.
  - Scatter plot => dùng nhiều
  - Bubble plot
  - Correlogram
  - Heatmap

=> Mục tiêu: tìm mối tương quan của 2 biến liên tục, mối liên hệ giữa các biến

Scatter Plot

- Scatter plots show data points for two numerical variables, displaying a variable on both axes.



=> nếu các điểm vex lên thành 1 đg thagwr thì cod thể đi 1 trường vì từ trường kia có thể nội suy ra

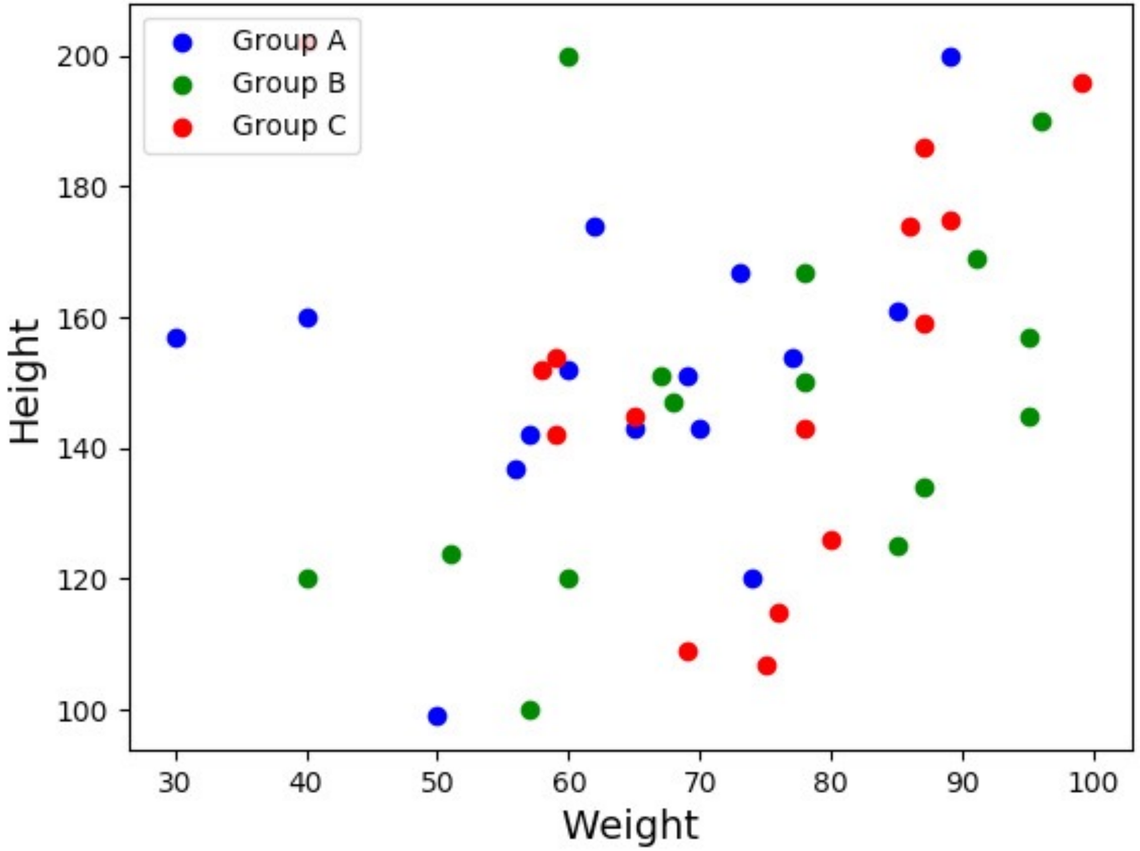
A scatter plot of height and weight of persons belonging to a single group

Scatter Plot

- With scatter plot:
  - Can detect whether a correlation (relationship) exists between two variables.
  - Can plot the relationship for multiple groups or categories using different colors.
  - A bubble plot, which is a variation of the scatter plot, is an excellent tool for visualizing the correlation of a third variable.

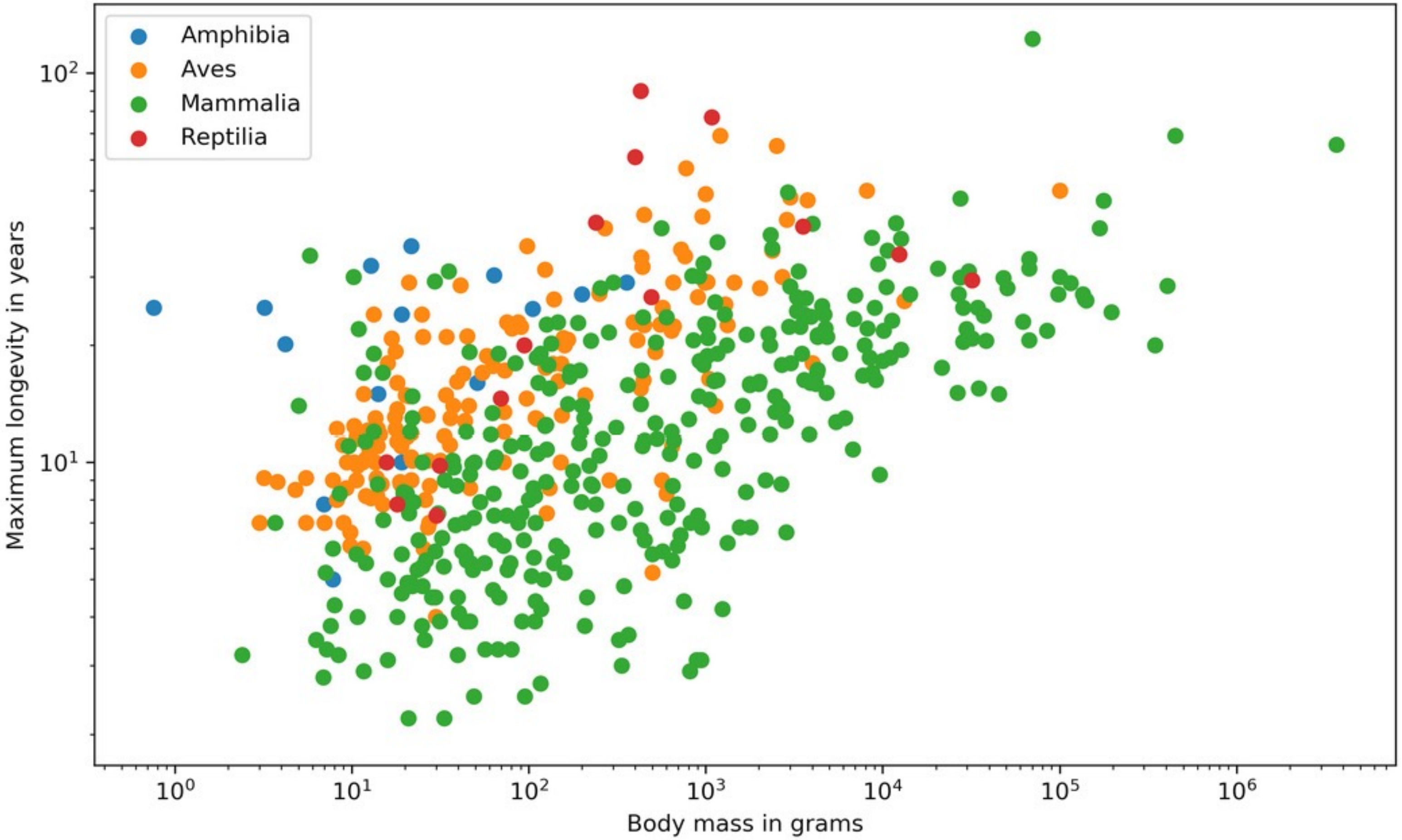
Scatter Plot Example

- Scatter plot with multiple variables (three groups)



Scatter Plot Example

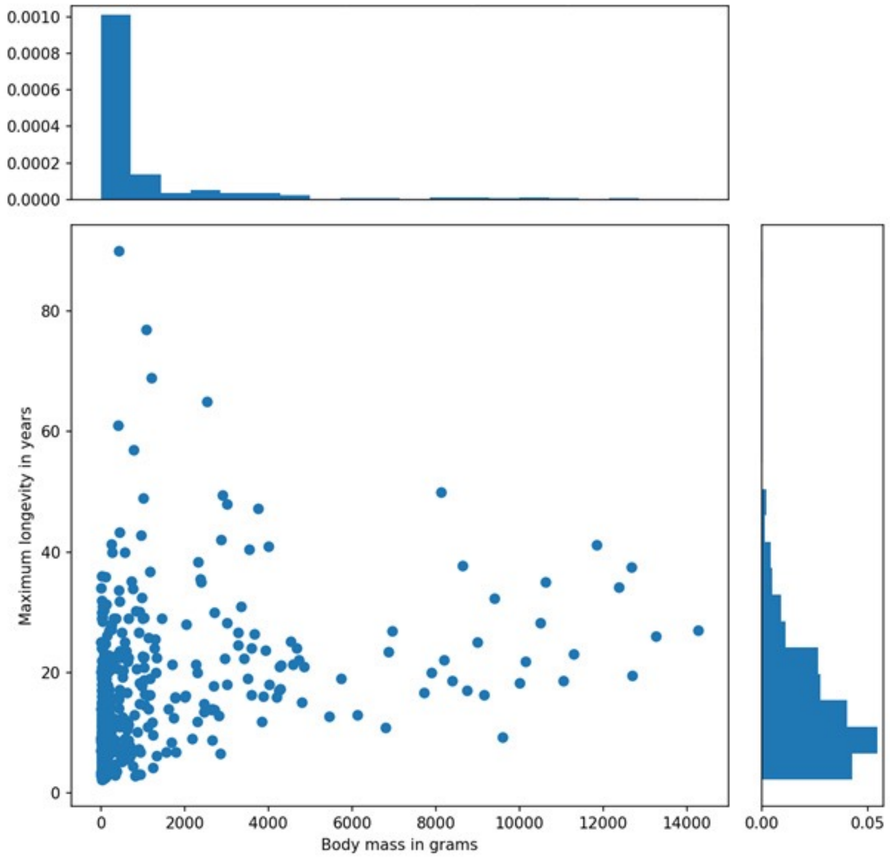
- The following diagram shows the correlation between the body mass and the maximum longevity for various animals grouped by their classes.
  - There is a positive correlation between the body mass and the maximum longevity



Variants: scatter plots with marginal histograms

- In addition to the scatter plot, you can plot the marginal distribution for each variable in the form of histograms to give better insight into how each variable is distributed.

=> bổ sung thêm sự phân bố

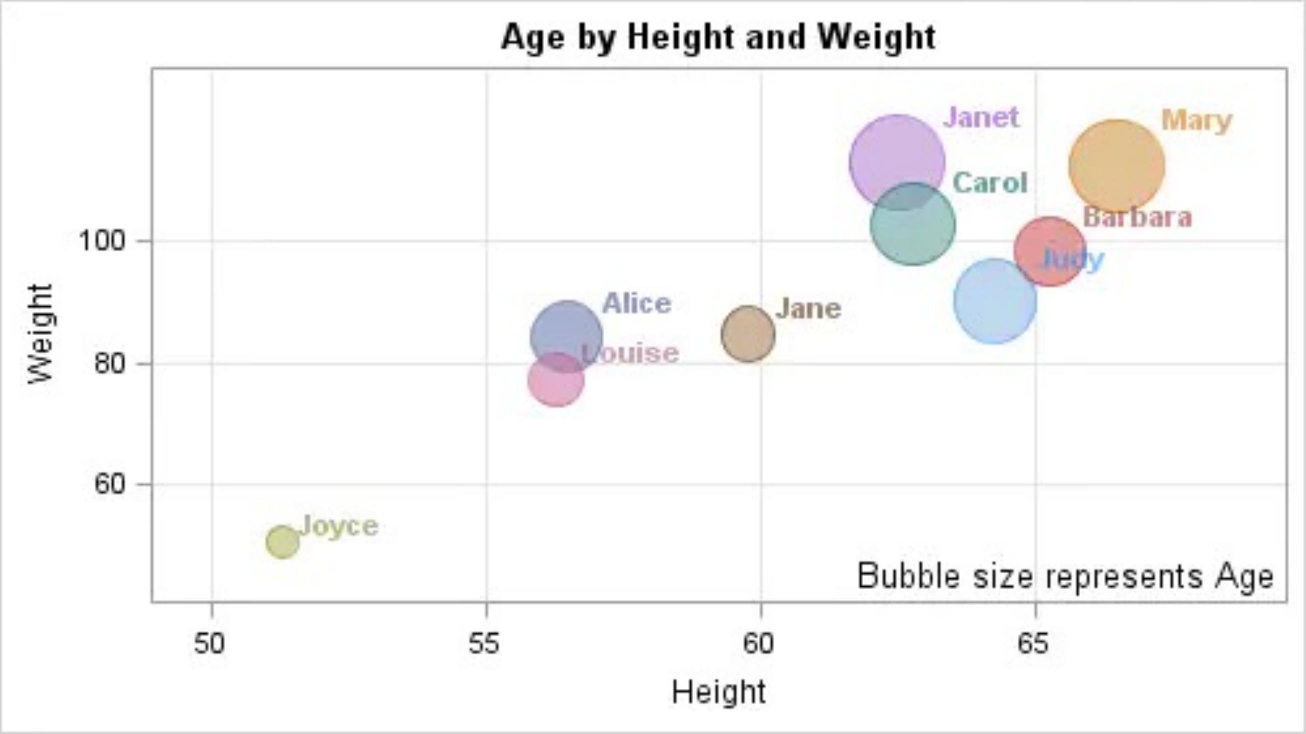


Correlation between body mass and maximum longevity of the Aves class with marginal histograms

Bubble Plot

=> bổ sung thêm tuổi

- A bubble plot extends a scatter plot by introducing a third numerical variable.
  - To show a correlation between three variables.
  - The value of the variable is represented by the size of the dots. The area of the dots is proportional to the value.
  - A legend is used to link the size of the dot to an actual numerical value.



Correlogram

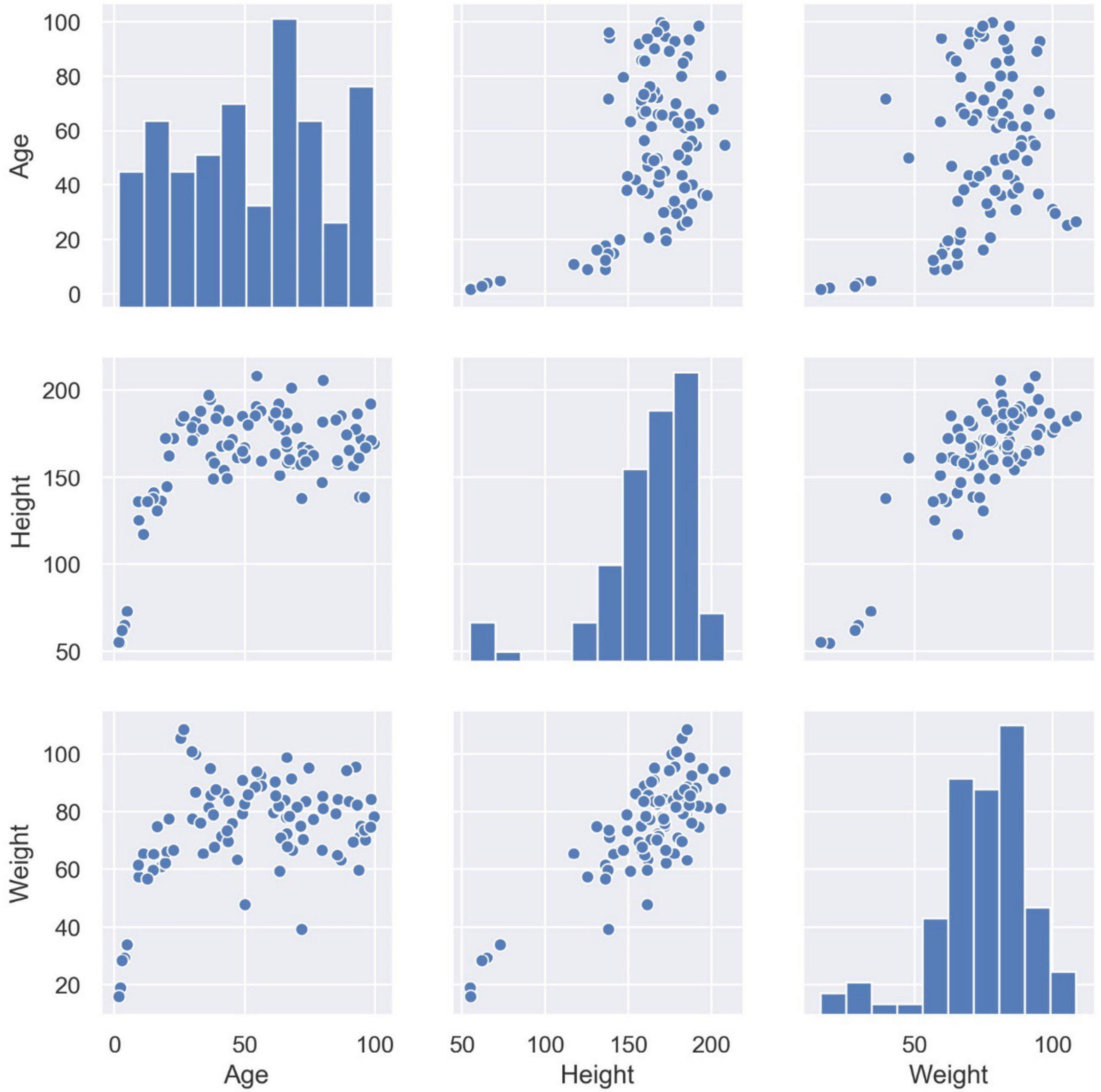
- A correlogram is a combination of scatter plots and histograms.
  - A correlogram or correlation matrix visualizes the relationship between each pair of numerical variables using a scatter plot.
  - The diagonals of the correlation matrix represent the distribution of each variable in the form of a histogram.

Correlogram Example

=> so sánh từng cặp

=> trên đg chéo là lược đồ phân phối của giá trị đó

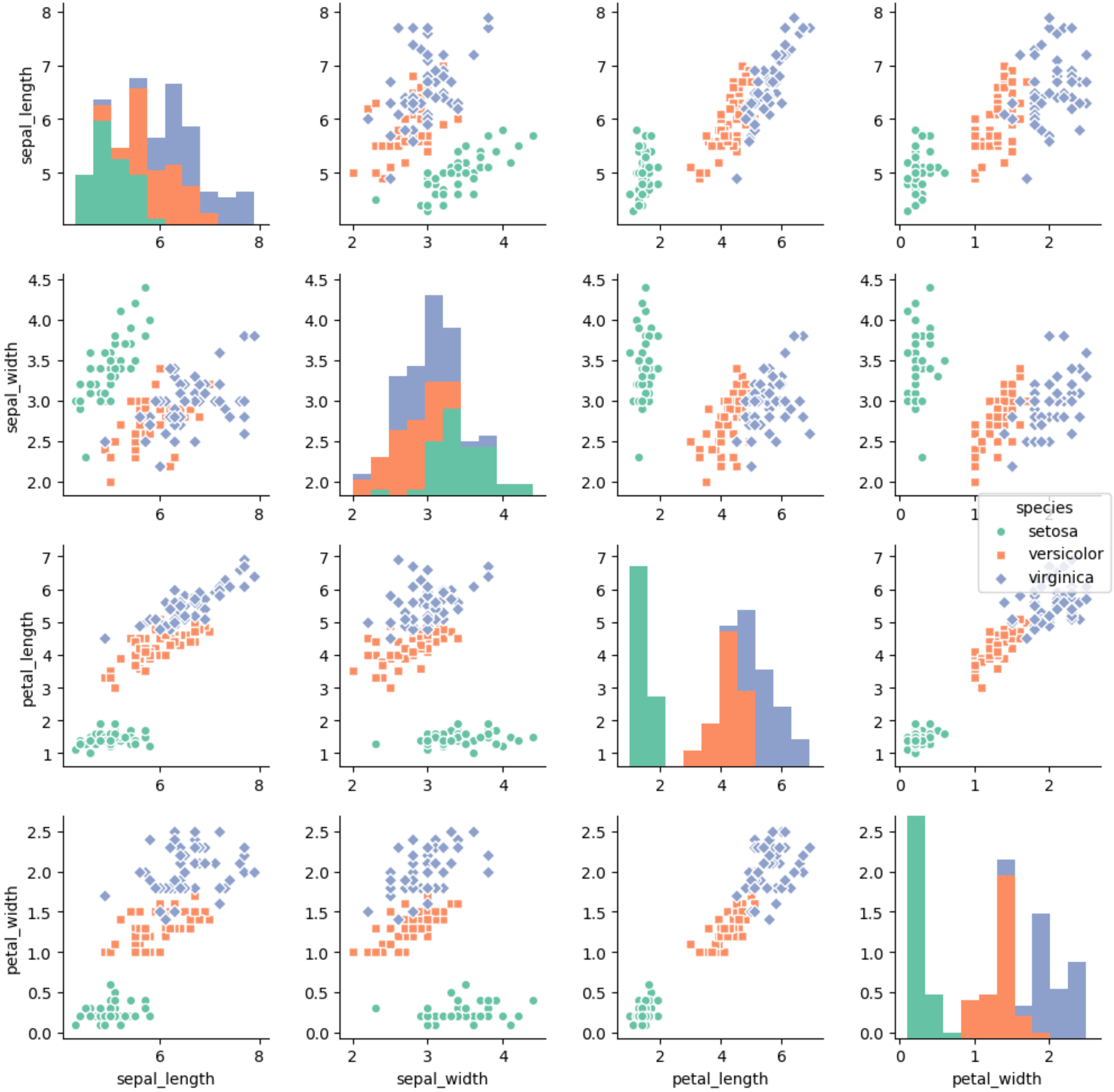
- The following diagram shows a correlogram for height, weight, and age of humans.



Correlogram Example

- Correlogram with multiple categories

=> bổ sung thêm group, thêm màu
=> cái nào phân bố thấp hơn thì nằm trước

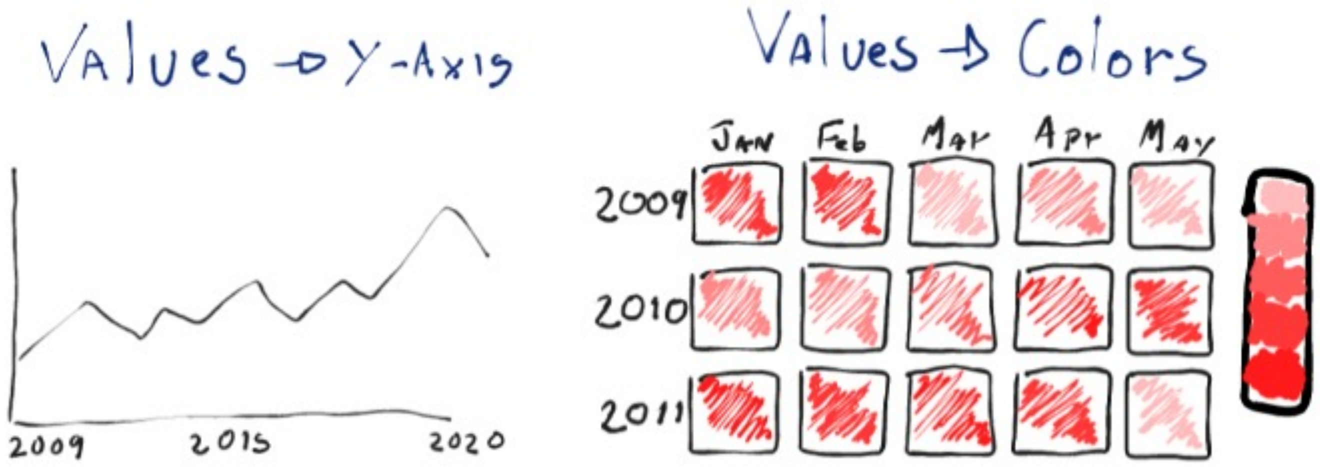


Heatmap

=> Bản đồ nhiệt: hình thành ma trận trong đó so sánh thành phần giá trị. Giá trị cột, dòng là rời rạc. Còn bên trong là các giá trị tương quan. VD: bài toán hoa lá sen.
=> So sánh biến cate dựa trên giá trị liên tục

- A heatmap is a visualization where values contained in a matrix are represented as colors or color saturation.
- Heatmaps are great for visualizing multivariate data
  - categorical variables are placed in the rows and columns
  - numerical or categorical variable is represented as colors or color saturation.

EX: các sàn thương mại và sản phẩm là biến rời rạc.+ biểu thị sự tương quan của sp vs các àn TMDT đã bán nó + Bfont-size:14pt; font-style:normal;font-weight:normal; color:#C80F12;font-family:Helvetica" >đc nhiều+ AA bán đc nhiều nhất trog các sp



Content

- Introduction
- **Types of Visualization**
  - Comparison Plots
  - Relation Plots
  - Composition Plots => mô tả 1 phần so với tổng thể
  - Distribution Plots
  - Geo Plots

Composition Plots

- Composition plots are ideal if you think about something as a part of a whole.
  - Pie chart
  - Stacked bar chart
  - Venn diagram

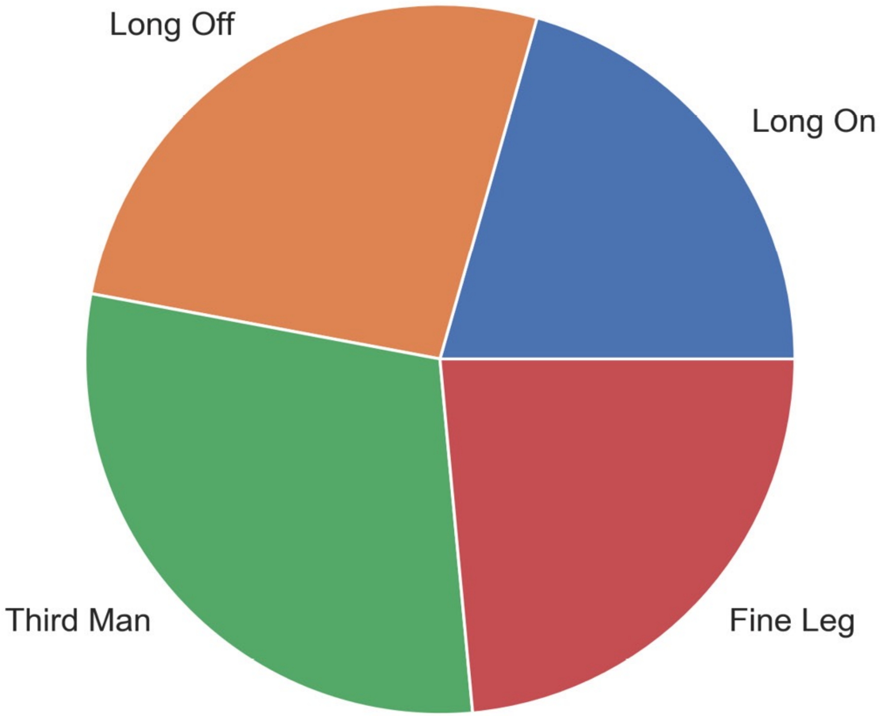
Pie Chart

=> so sánh giữa 2 cá thể độc lập thì nên dùng bar chart (chart đứng)

- Pie charts illustrate numerical proportion by dividing a circle into slices.
    - Each arc length represents a proportion of a category.
    - The full circle equals to 100%.
    - For humans, it is easier to compare bars than arc lengths; therefore, it is recommended to use bar charts or stacked bar charts most of the time.

Pie Chart Example

- The following diagram shows a pie chart that shows different fielding positions of the cricket ground, such as long on, long off, third man, and fine leg.

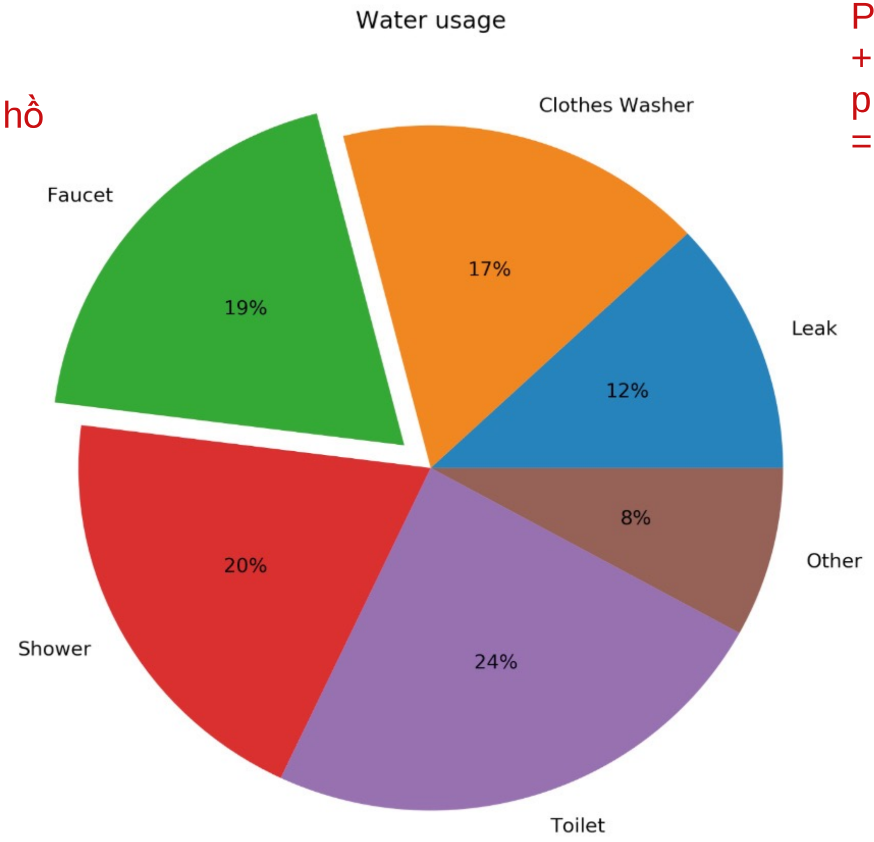


Pie Chart Example

=> Có miếng bánh đc tách ra để nhấn mạnh, tách tối đa 2 phần
+ Nếu muốn tách 2 cái thì gom nó lạo gần nhau, độ rời giữa 2 phần ko giống nhau

- The following diagram shows water usage around the world:

Design practise:
+ Nên xếp tăng dần or giảm dần theo chiều đồng hồ

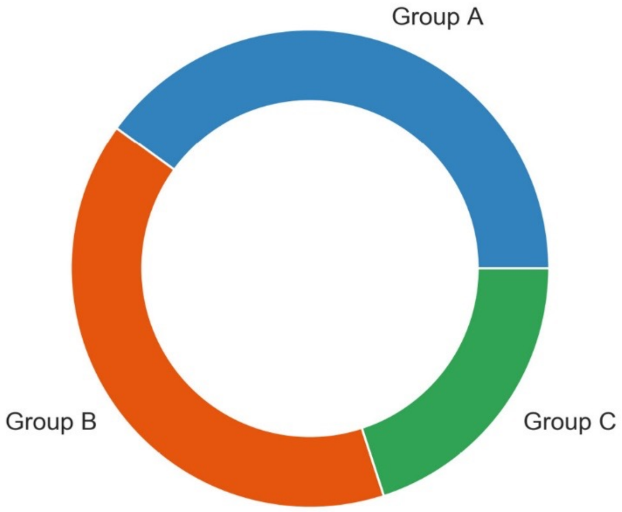


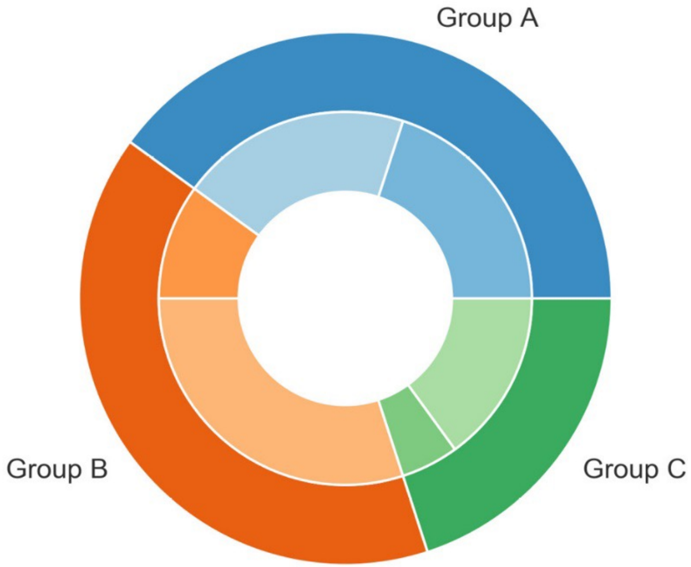
Pie Exercise (Youtube):
+ Cons: có quá nhiều giá trị. Ko thể dùng pie vì ko phù hợp, nó ko so sánh vs tổng thể
=> Dùng bar chart vì ta đag so sánh giữa các ytbs

Pie Chart Variants: Donut Chart

=> đỡ tốn màu, ko ai dùng như hình 1
=> chỉ dùng như hình 2 khi ta mún lồng những phần trăm vs nhau => có thể chia đc các cấp độ thấp hơn

- Donut charts are more space-efficient because the center is cut out, so it can be used to display information or further divide groups into sub-groups.

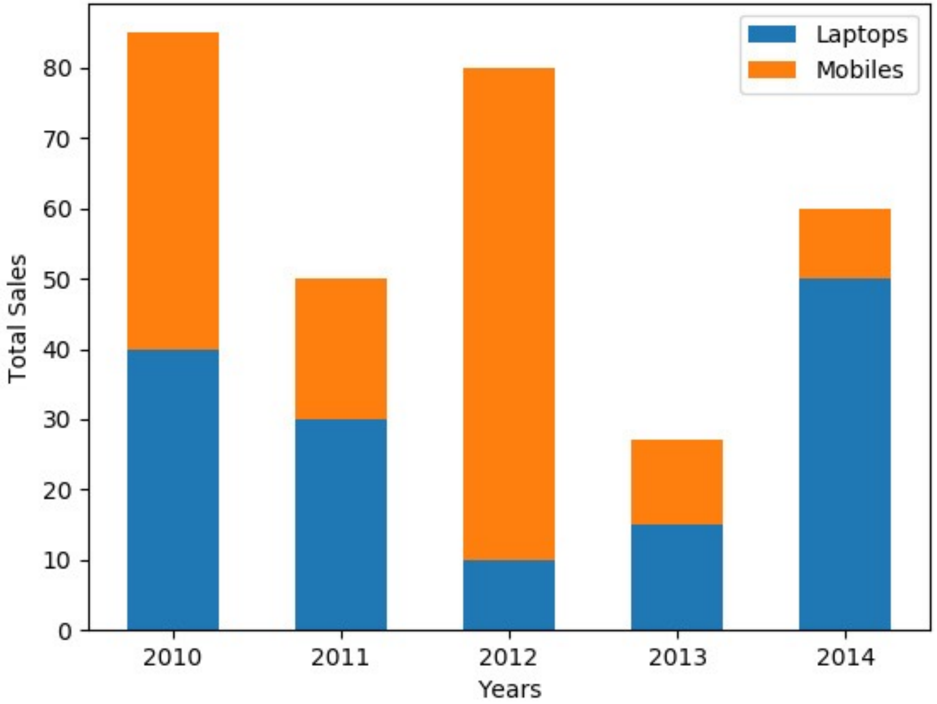




Stacked Bar Chart

=> chia 1 loại nào đó thành các loại con và xem % của nó
=> ko dùng bar chart vì ta muốn so sánh (so sánh tỉ lệ thành phần) tỷ trọng thay đổi qua từng năm và so sánh các năm

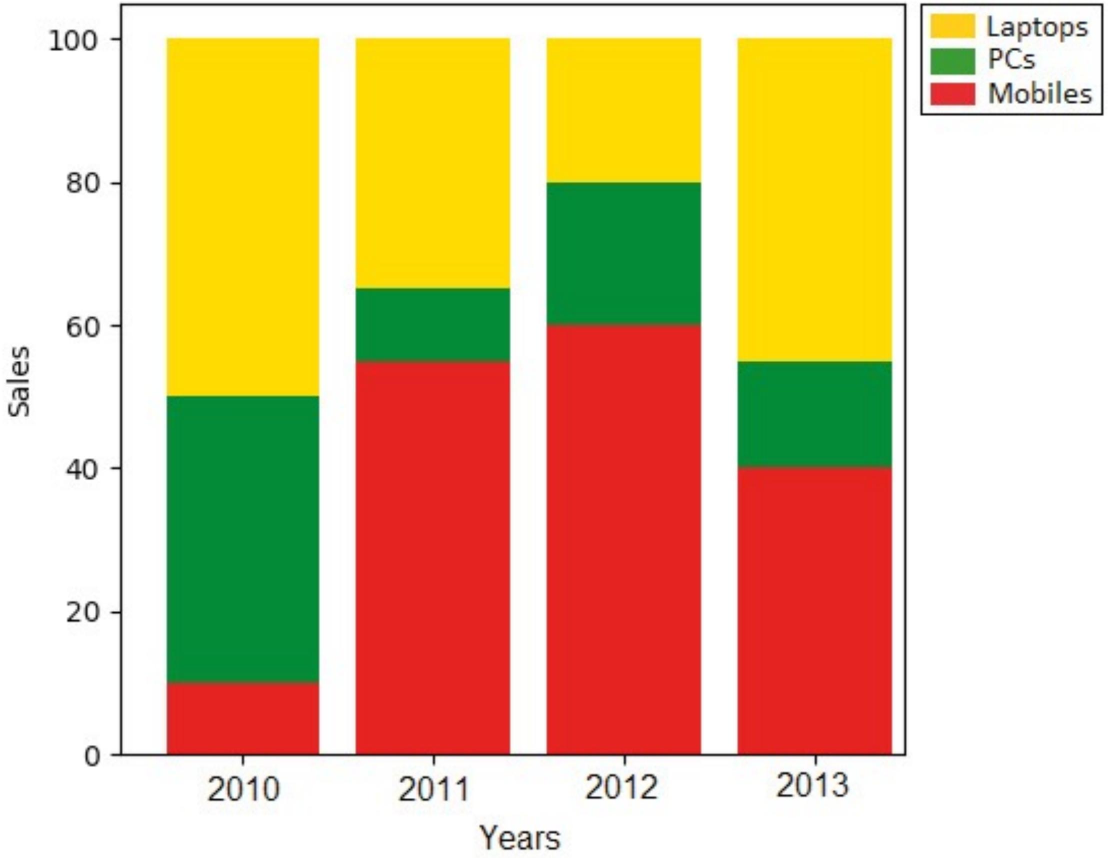
- Stacked bar charts are used to show how a category is divided into sub-categories and the proportion of the sub-category, in comparison to the overall category.
  - You can either compare total amounts across each bar or show a percentage of each group.



Stacked Bar Chart

Mục tiêu: ko phải là so sánh doanh thu qua từng năm mà muốn so sánh tỷ lệ thành phần

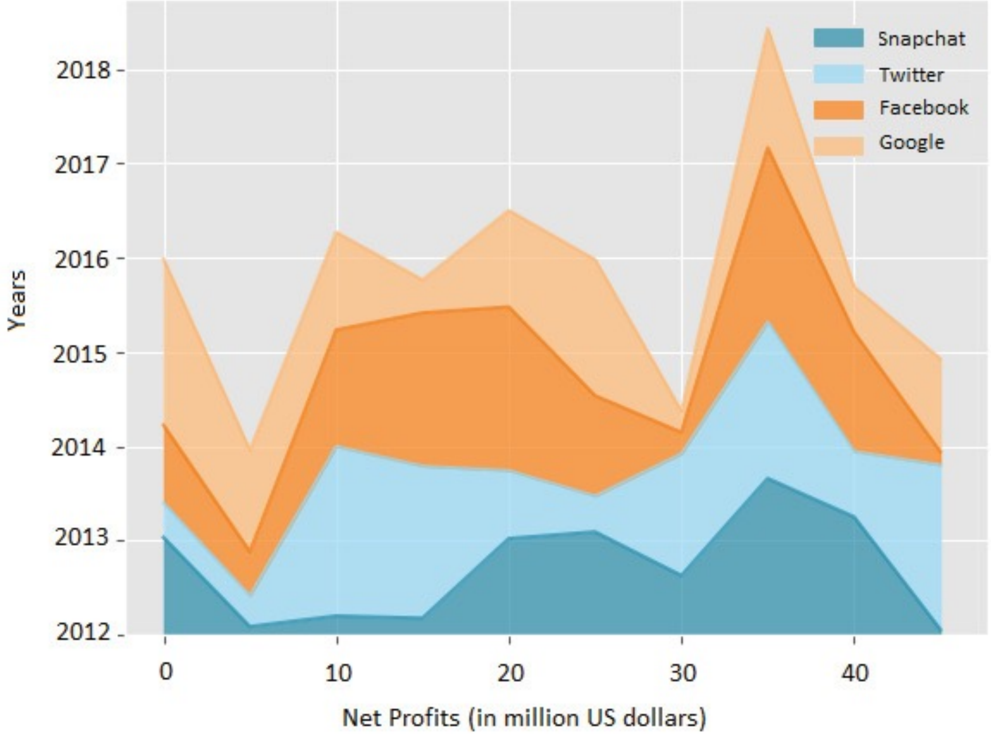
- A 100% stacked bar chart makes it easier to see relative differences between quantities in each group.



=> bổ sung thêm trend và miền => thấy đc sự liên tục => khó nhìn nên ít dùng

Stacked Area Chart

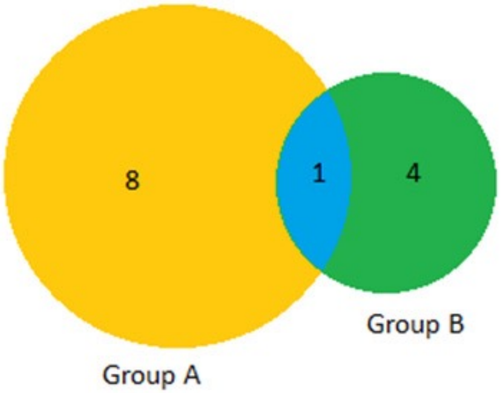
- Stacked area charts show trends for part-of-a-whole relations.
  - The values of several groups are illustrated on top of one another.
  - It helps to analyze both individual and overall trend information.



Venn Diagram

=> tập hợp và tập hợp ko giới hạn
=> so sánh các tập hợp xem chúng có giao nhau ko và giao bn %
=> có sự giao nhau giữa các tỷ lệ

- Venn diagrams, also known as set diagrams, show all possible logical relations between a finite collections of different sets.
  - Each set is represented by a circle.
  - The circle size illustrates the importance of a group.
  - The size of an overlap represents the intersection between multiple groups.



Content

- Introduction
- **Types of Visualization**
  - Comparison Plots
  - Relation Plots
  - Composition Plots
  - Distribution Plots => Lược đồ phân phối
  - Geo Plots

Distribution Plots

- Distribution plots give a deep insight into how your data is distributed.
  - Histogram
  - Density plot
  - Box plot
  - Violin plot

Histogram

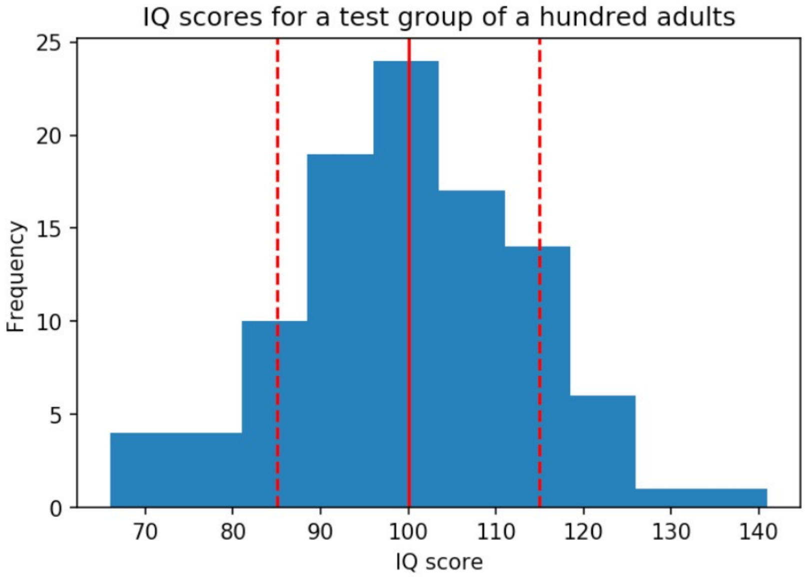
=> phân bố của biến đơn
=> no ko phải là bar chart vì ko thể hoán đổi giữa các cột , phải xếp theo range

- A histogram visualizes the distribution of a single numerical variable.
  - Each bar represents the frequency for a certain interval.
  - Either plot a histogram with absolute frequency values or alternatively normalize your histogram.
  - To compare distributions of multiple variables, use different colors for the bars.
- Histograms help get an estimate of statistical measures.
  - Identify where values are concentrated
  - Easily detect outliers.

Histogram Example

=> BD tứ phân vị

- The distribution of the Intelligence Quotient (IQ) for a test group.
  - The solid line indicates the mean and the dashed lines indicate the standard deviation

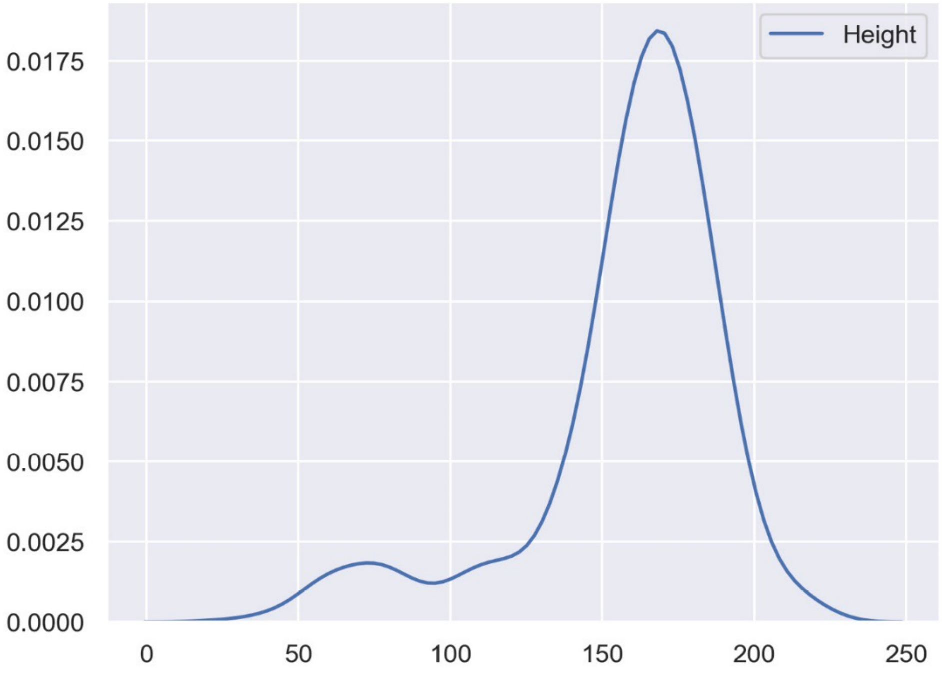


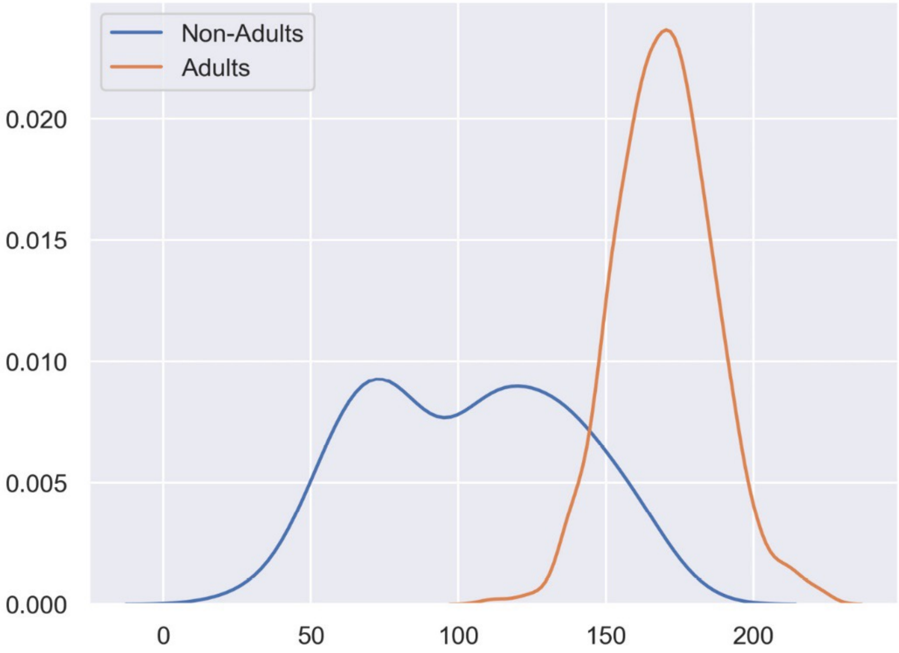
Density Plot

=> cũng là histogram nhưng mượt hơn để ta chiếu đg thì nó khớp vs mô hình toán học nào
=> coi xem hàm toán học có đáp ứng đc BD đó ko

- A density plot shows the distribution of a numerical variable.
  - It is a variation of a histogram that uses kernel smoothing, allowing for smoother distributions.
  - An advantage they have over histograms is that density plots are better at determining the distribution shape, since the distribution shape for histograms heavily depends on the number of bins (data intervals).

Density Plot Example





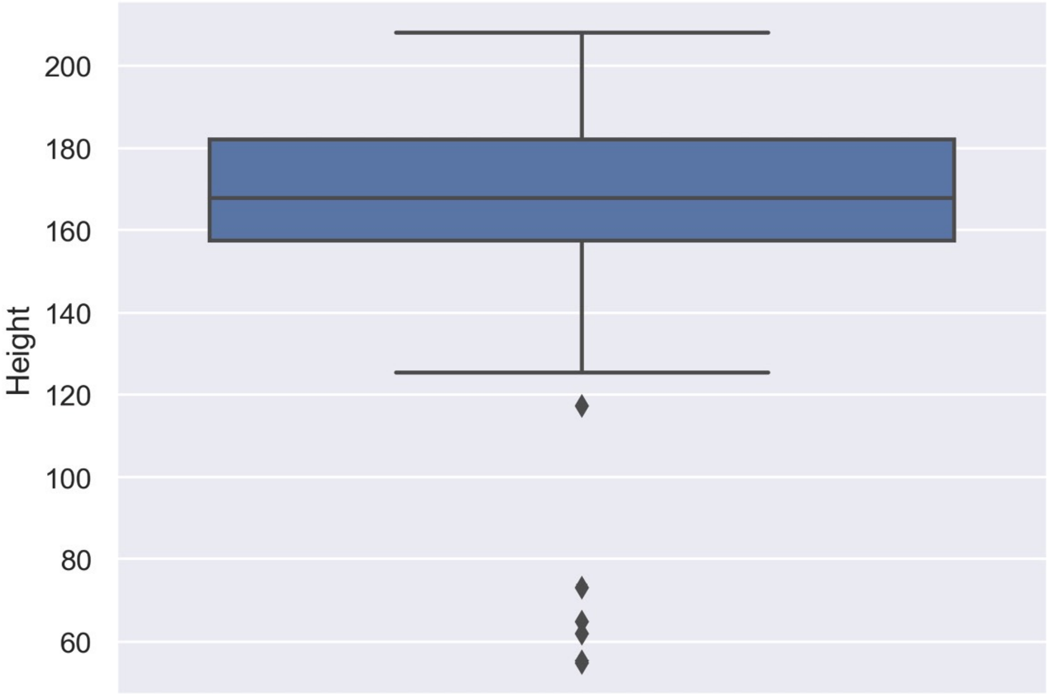
Box Plot

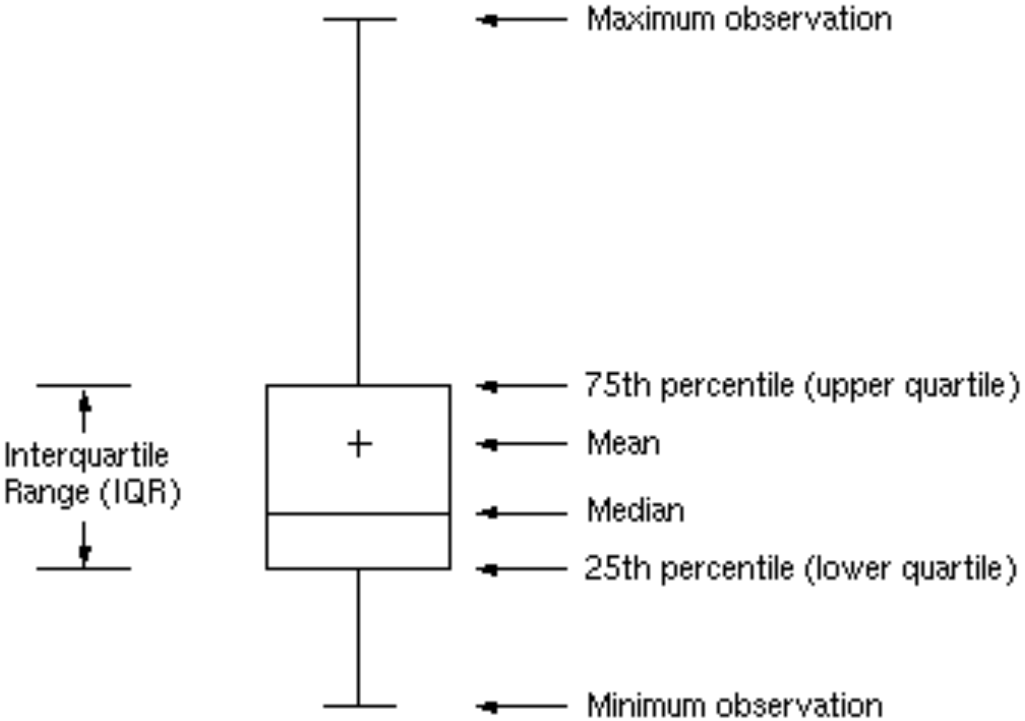
=> ko cần biểu diễn nhiễu

- The box plot shows multiple statistical measurements.
  - The box extends from the lower to the upper quartile values of the data, thus allowing us to visualize the interquartile range.
  - The horizontal line within the box denotes the median.
  - The whiskers extending from the box show the range of the data. It is also an option to show data outliers, usually as circles or diamonds, past the end of the whiskers.

Box Plot Example

=> vẽ đơn giản hơn so vs histogram
=> có 2 biến thể: có dấu + và ko có

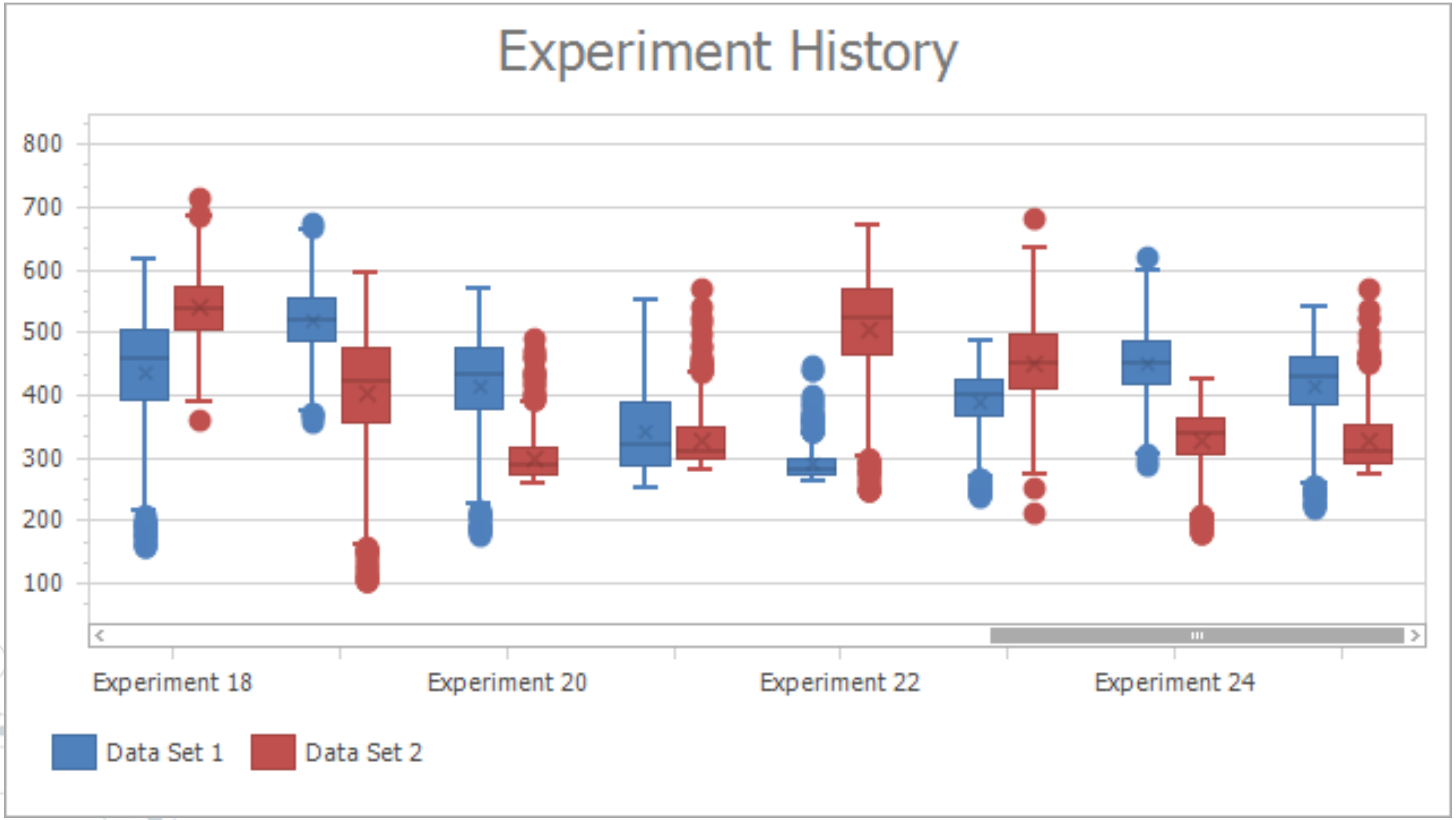



Maximum observation
75th percentile (upper quartile)
Mean
Median
25th percentile (lower quartile)
Minimum observation
Interquartile Range (IQR)

=> chỉ dùng khi có nhiều biến, khi so sánh nhiều phân phối

Box Plot Example

- If you want to compare statistical measures for multiple variables or groups, simply plot multiple boxes next to one another.

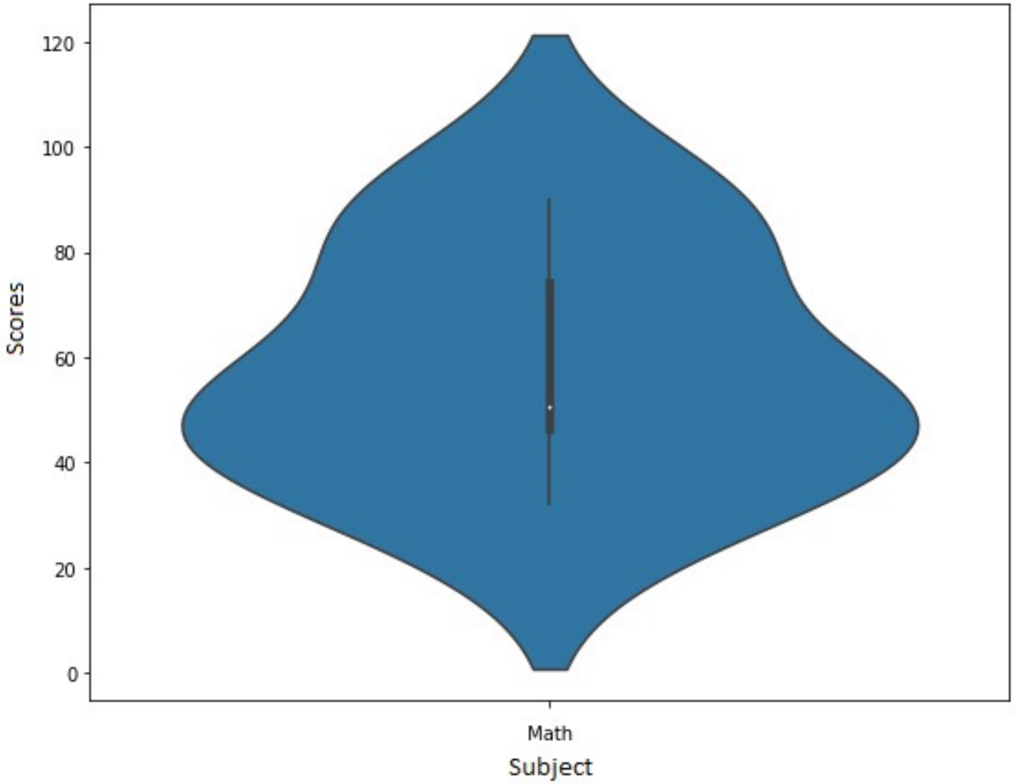


Violin Plot

- Violin plots are a combination of box plots and density plots.
  - Both the statistical measures and the distribution are visualized.
  - The thick black bar in the center represents the interquartile range, the thin black line shows the 95% confidence interval, and the white dot shows the median.
  - On both sides of the center-line, the density is visualized.

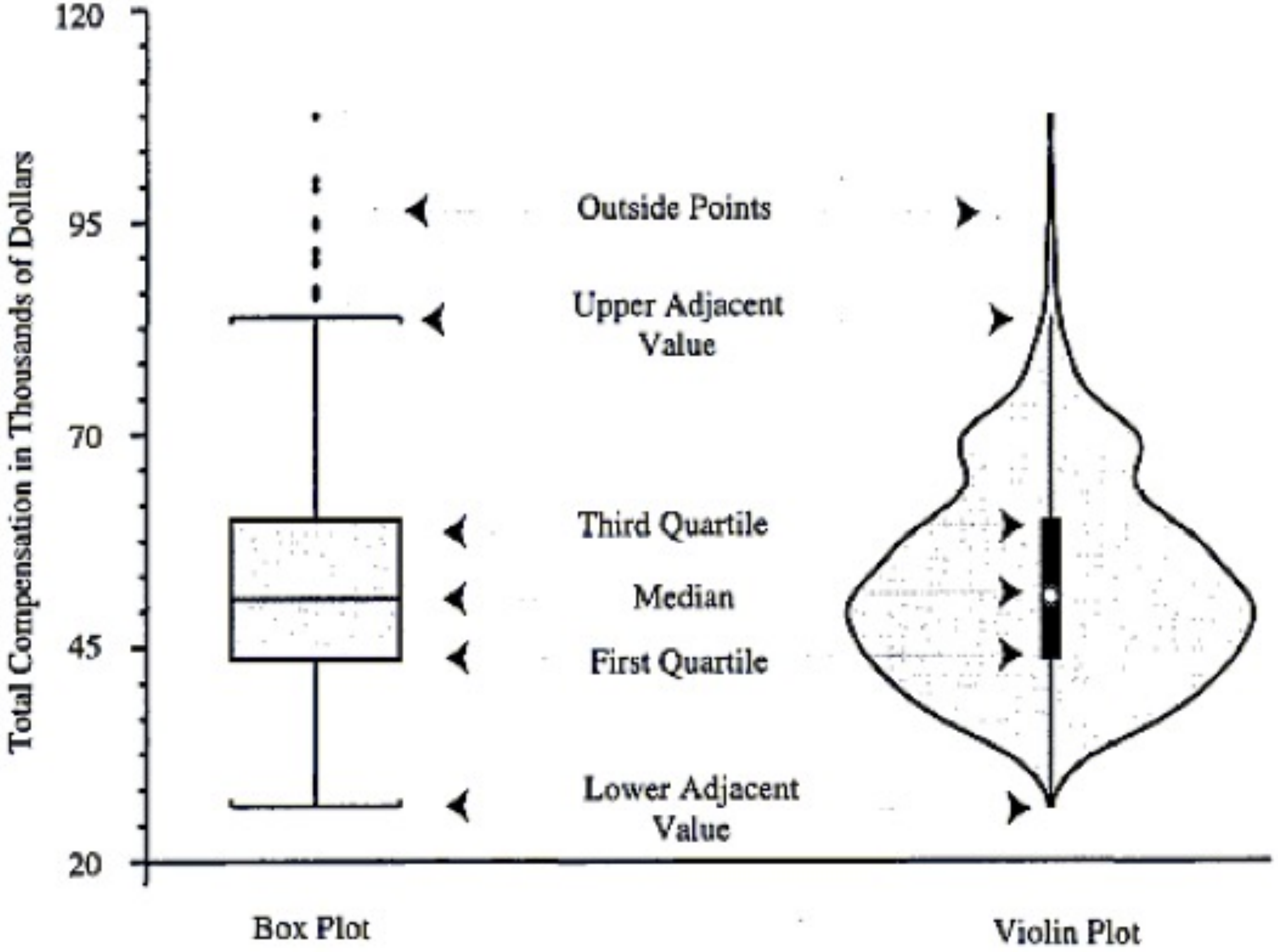
Violin Plot Example

- The following diagram shows a violin plot for a single variable and shows how students have performed in Math:



=> Violin phức tạp hơn

Box plot vs Violin Plot



Content

- Introduction
- **Types of Visualization**
  - Comparison Plots
  - Relation Plots
  - Composition Plots
  - Distribution Plots
  - Geo Plots => lược đồ bản đồ

Geo Plots

- Geological plots are a great way to visualize geospatial data.
  - Dot map => tương tự scatter plot
  - Choropleth map
  - Connection map

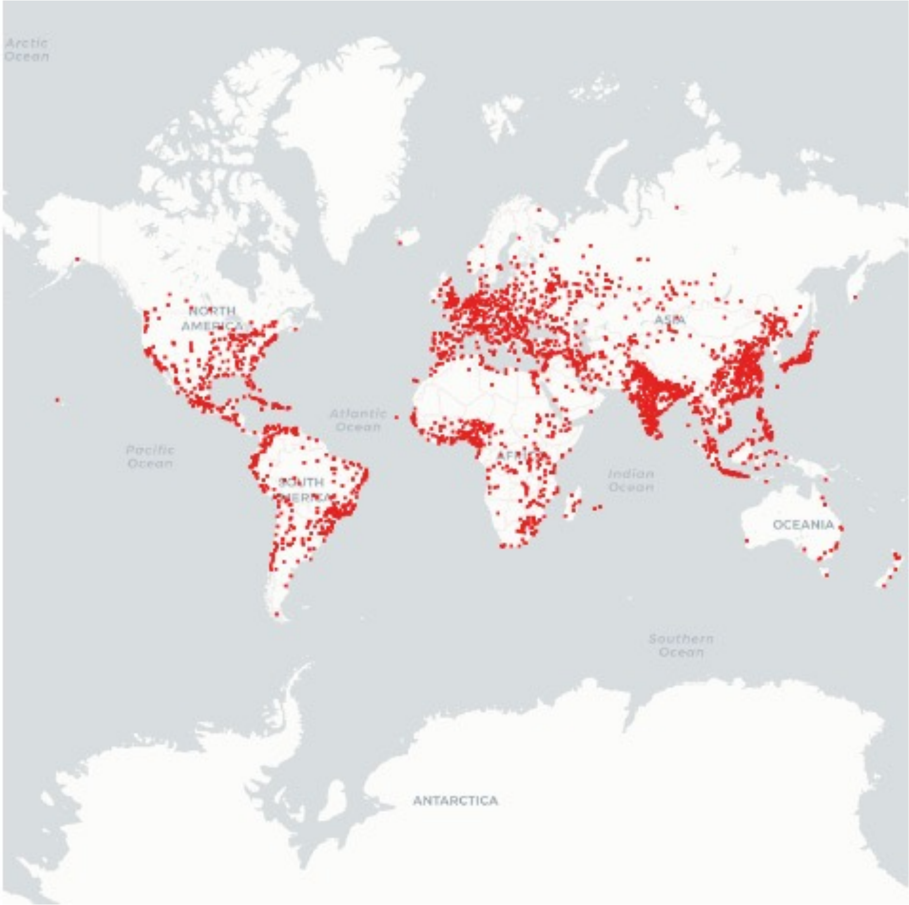
Dot Map

- In a dot map, each dot represents a certain number of observations.
  - Each dot has the same size and value (the number of observations each dot represents).
  - The dots are not meant to be counted—they are only intended to give an impression of magnitude.
  - The size and value are important factors for the effectiveness and impression of the visualization.
  - You can use different colors or symbols for the dots to show multiple categories or groups.

Dot Map Example

=> bản đồ có thêm những chấm

- The following diagram shows a dot map where each dot represents a certain amount of bus stops throughout the world



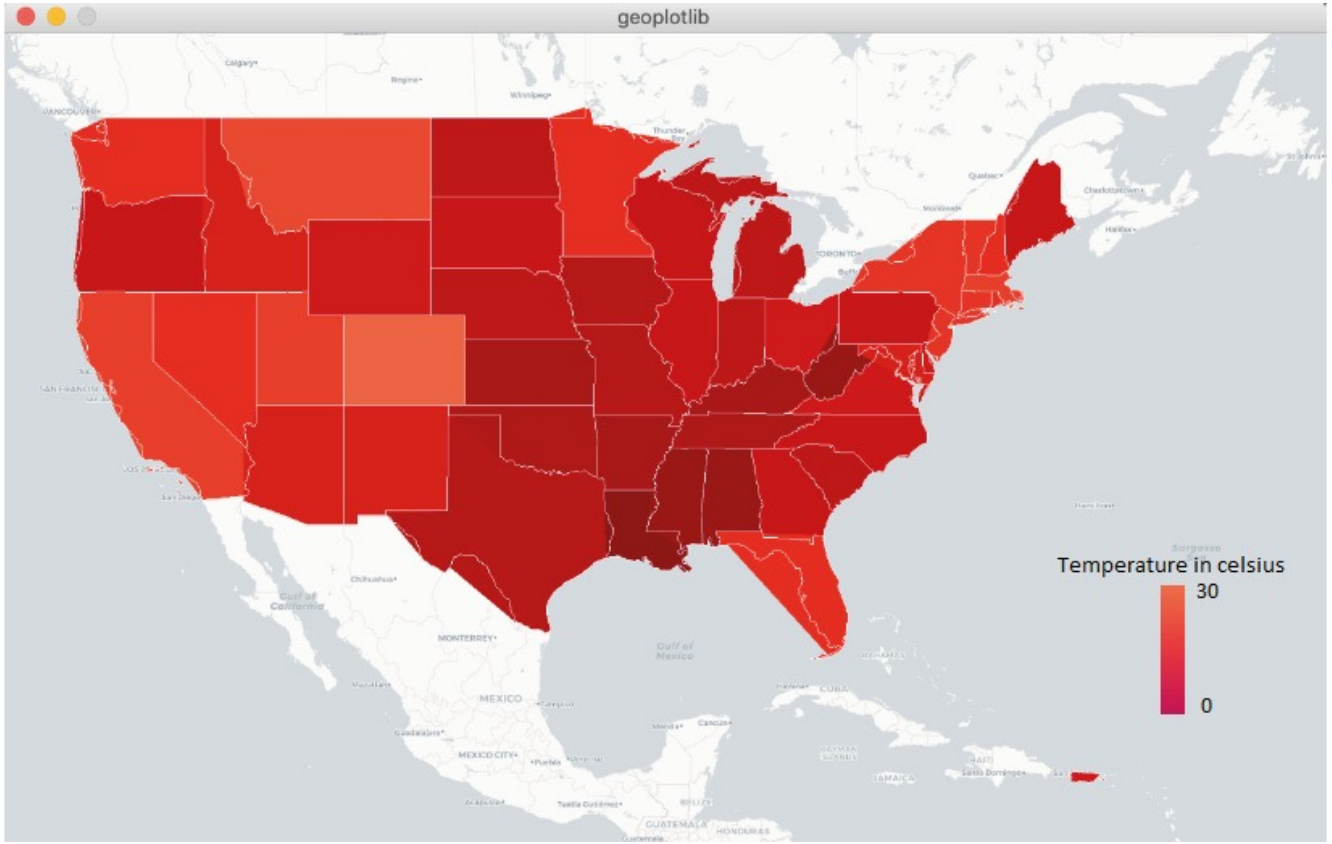
Choropleth Map

=> lai giữa heatmap và bản đồ => dùng nhiều

- In a choropleth map, each tile is colored to encode a variable.
  - A tile represents a geographic region for, for example, counties and countries.
- Choropleth maps provide a good way to show how a variable varies across a geographic area.

Choropleth Map Example

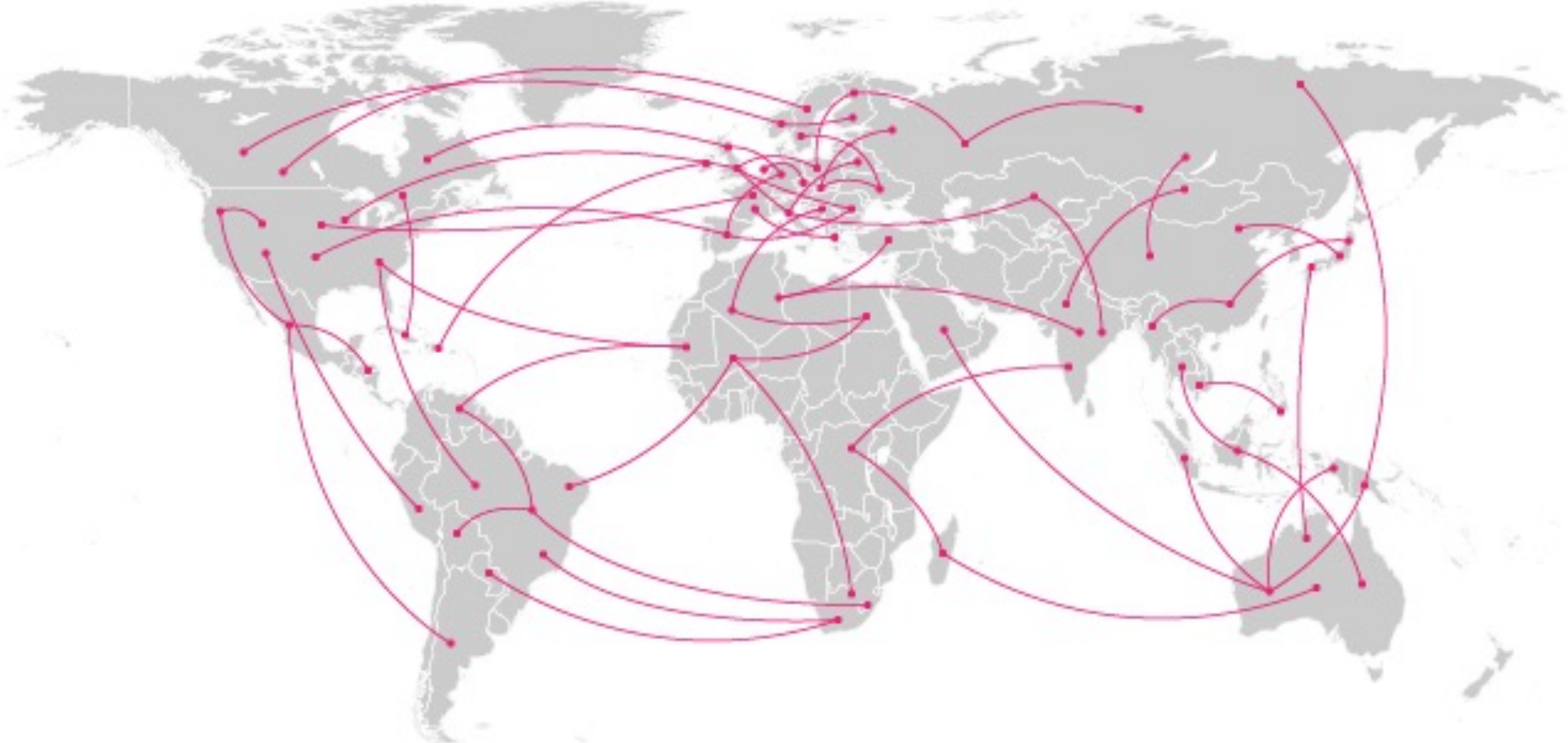
- The following diagram shows a choropleth map of a weather forecast in the USA:



Connection Map

=> mối tương quan giữa 2 vùng địa lý

- In a connection map, each line represents a certain number of connections between two locations.
  - The link between the locations can be drawn with a straight or rounded line representing the shortest distance between them.



Thank you for listening