Visualising numerical and non-numerical data

5.1 Visualising numerical data

Introduction

In the previous weeks, you have explored the need to make sense of data and the methods of translating it into meaningful visualisations. As you process through great amounts of data, your job is to filter the inaccurate data and present it in a significant way.

Before you investigate how to visualise data, you must understand what kind of data (Variables) you are handling.

In this section, you will learn more about Categorical and Numerical variables.

a. Categorical
have values that can only be placed into categories such as yes and no. Categorical data represent features such as gender, marital status, hometown, or shopping preferences. Categorical data can take on numerical values (such as “1” indicating male and “2” indicating female), but those numbers don’t have a mathematical meaning as you cannot add them together.

How to visualise **categorical variables**?

- Using bar chart, pie chart, Pareto chart, or stacked chart.

b. Numerical
have values that represent measures. Numerical variables are further identified as being either discrete or continuous variables:

- **Discrete variables** have numerical values from a counting process; their possible values cannot be counted and can only be described using intervals on the real number line. The list of possible values may be fixed (also called finite), or it may go from 0, 1, 2 to infinity (making it countably infinite).
- **Continuous variables** produce numerical responses that arise from a measuring process. For example, the exact amount of petrol purchased at the pump for cars with 10-liters tanks would be continuous data from 0 litres to 10 litres, represented by the interval [0, 10], inclusive. You might pump 7.40 litres, or 7.41, or 7.413463 litres, or any possible number from 0 to 10.

How to visualise **numerical variables**?

- For **Two Numerical Variables**: Using scatter plot, area charts, map charts and time-series plot.

Activity: Differentiating categorical and quantitative variables

The FAA monitors airlines for safety and customer service. The carrier must report the type of aircraft, flight number, number of passengers, and whether the flights departed and arrived on schedule for each flight. What variables are reported for each flight, and are they quantitative or categorical?
Sort the following variables as Quantitative or Categorical.

Quantitative	Categorical
Number of Passengers	Type of aircraft Arrived / Departed on schedule Flight number

Visualising numerical data

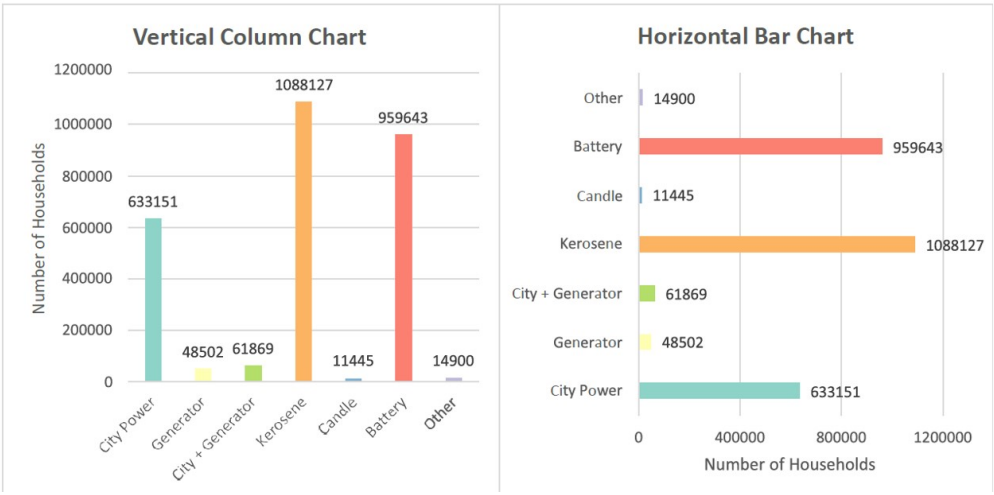
Stacked graphs

Our visualisations should start with the response, understanding the traits of its variables, and then build outward from the extra information provided in the data.

Bar charts are useful for ranking categorical data by analysing how two or more values or groups compare to each other in relative degree at a given point in time.

The figure below shows both a vertical column and a horizontal bar chart representing the same data. The vertical column chart measures the categorical data (household light source) at one time. It ranks the categorical data so that it is easy to compare values between the various light sources in 2008. This horizontal bar graph represents the same data but shows an alternative method for visualising categorical data at one point in time. (Talmadge & Gale, 2018)

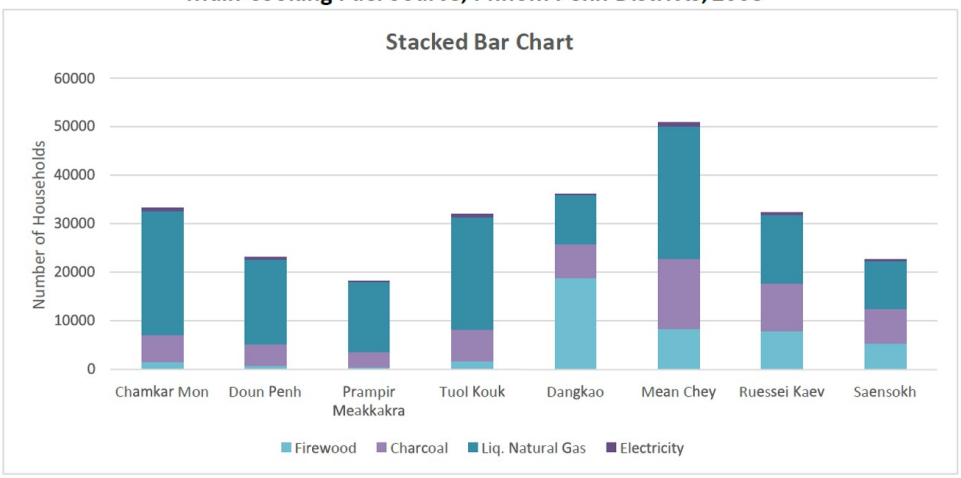
Cambodian Households' Main Source of Light, 2008[1]



Source: Cambodian Households' Main Source of Light, 2008 (Tufts Data Lab, 2016)

Stacked bar charts are useful when the sum of all the values is as important as the individual categories/groups. Stacked bar charts show multiple values for individual categories and the total for all the combined categories.

Main Cooking Fuel Source, Phnom Penh Districts, 2008[1]



Source: Main Cooking Fuel Source, Phnom Penh Districts, 2008 (Tufts Data Lab, 2016)

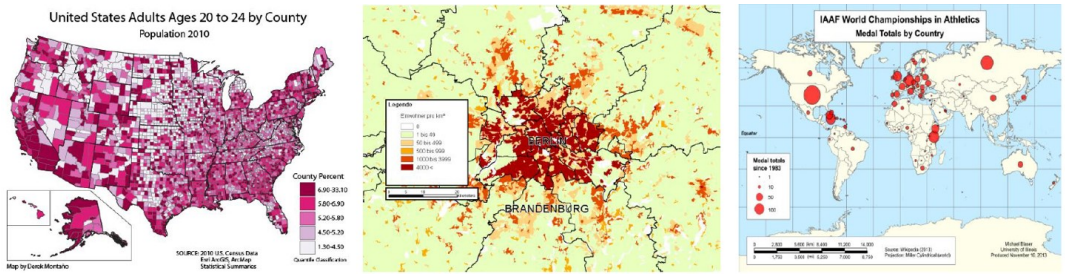
Although stacked graphs help convey multiple meanings simultaneously, they also have some limitations. While it's easy to interpret the values for the full bar and the first group of the bar, it is challenging to quantify the values for subsequent groups (strips) in the same bar or compare the groups within the same bar.

Clustered Bar Charts display categorical data next to each other, rather than stacked in the same bar, to easily compare values between groups. (Talmadge & Gale, 2018)

You may explore more about Stacked Bar Charts in Tableau (https://www.rigordatasolutions.com/post/stacked-bar-chart-in-tableau) to further your understanding.

Mapping

Maps possess a visual appeal and can help demonstrate how data is distributed spatially.



Source: Maps (PwC, 2016)

When to use maps

1. To categorise regions into four or fewer groups. More than four groups can be difficult to interpret. A U.S. map showing states in red or blue for Republicans vs. Democrats are a good example.
2. The closeness of the geographic locations adds meaning to the data, such as a map showing virus outbreaks.
3. The client understands the environment mainly in a geographic sense.
4. The geographic data is multidimensional, such as sales by state and each city within the state.

The most effective best practice relating to maps is deciding whether to use one. Beyond that, several rules apply only to maps:

1. Use different colours to identify different states (Republican vs. Democrat).
2. Use the same colour shades to indicate the degree or size of the same value (such as population size). However, do not use more than three shades if an accurate shade determination is necessary.
3. Use hovers to provide additional data on-demand.
4. The readability of the smallest regions will determine the overall size. If space doesn't allow that, consider a different visualisation type.
5. If you have location data like coordinates, country names, state names or abbreviations, or addresses, you can plot a map. Map charts are good for giving your numbers a geographical perspective to instantly identify the best and worst-performing areas, trends, and outliers.

6. Maps won't be very good for comparing exact values because map charts are usually colour scaled, and it's better to use overlay bubbles or numbers if you need to convey exact numbers or enable comparison.

In a nutshell, map charts are most useful when you want to:

- display quantitative information on a map
- present spatial relationships and patterns
- require a regional context for your data
- obtain an overview of the distribution across geographic locations.

In his Tableau Public "Viz of the Week" (https://public.tableau.com/app/profile/justindavis/viz/AtMinimum/AtMinimum), Justin Davis demonstrates the percentage of all US hourly workers that earn minimum wage or less. Each map in the set of small multiples provides a yearly snapshot of minimum wage workers going as far back as 2002. When you hover over a state, the percentage of workers for that area is noted and looking year-by-year, state-by-state, you can see if there is an upward or downward trend.

5.1.2 Tufte's design rules

According to Edward Tufte, a notable pioneer and statistician, one must strategically use data visualisation to communicate patterns and insights with 'clarity, precision and efficiency.

He notes that numeric or quantitative results are opportunities to tell strong narratives. Tufte, a statistician and artist, has written, designed, and self-published four books on data visualisation, including 'Beautiful Evidence', a well-respected book on visualising numerical findings.

Tufte's principles serve as a guide to accurately interpret data using visual elements:

1. Comparisons
   Show data by comparisons (bar charts and the like) to depict contrasts and differences between dependent variables.
2. Causality
   Demonstrate how one or more independent variables impact or influence dependent variables.
3. Multivariate
   Various data are combined so an audience can easily interpret an otherwise complex narrative.
4. Integration
   Incorporate various modes of information (texts, maps, calculations, diagrams, etc.) to show evidence of source data-to-findings.

5. Documentation

   For credibility, include attribution, detailed titles, and measurements (scales).

6. Context

   Describe or depict the before and after state. Show trend lines to hint at results in the future.

Colour

Colour is one of the most powerful appealing features because it's the first thing you notice, and it can immediately emphasise specific insights or identify outliers. The following are three key points to orient your data visualisation towards when using colours:

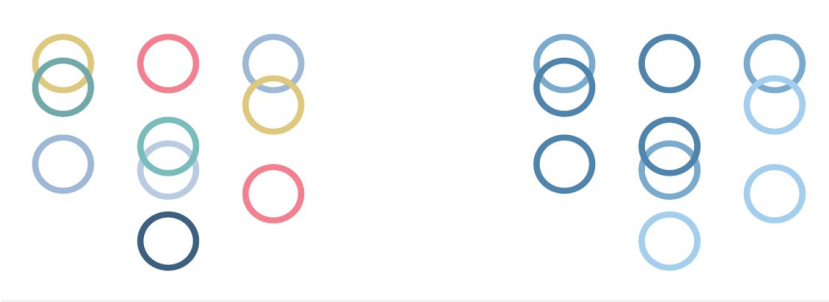
a. Colour for differentiation

   Do not use similar colours, or too many colours. Do not re-use colours for different dimensions or measures on the same dashboard.

Source: Good enough to great (Tableau, n.d.)

b. Measurable colours

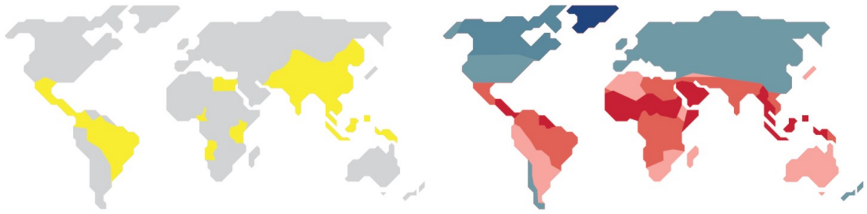
   Does the colour scale match my data? Does the colour move from light to dark, or is it positioned to best symbolise what you are measuring?



Source: Good enough to great (Tableau, n.d.)

c. Relatable colours

   Semantically resonant colors help people process information faster. Example, using red to represent heat. ColorBrewer offers a diagnostic tool for evaluating the robustness of individual color schemes. You may explore it to better understand the importance of colour in data visualisation.



Source: Good enough to great (Tableau, n.d.)

The following videos are recommended for further understanding.

- Watch this 2-minute 41-second video to learn how to create a stacked bar chart using multiple measures in Tableau. https://youtu.be/voNQa6jwAyo *Source: (Tableau, 2018)*

- Watch this 4-minute 13-second video to explore more about the types of data. https://youtu.be/DUcXZ08IdMo *Source: (365 Data Science, 2019)*

Read the following to extend your understanding.

- Cava, B. (October, 2015). *Airbnb San Francisco Analysis*. Tableau Public. https://public.tableau.com/app/profile/brit4337/viz/AirbnbSanFranciscoAnalysis/Airbnb
- Chou, L. (2019). Top 10 Map Types in Data Visualization. Towards Data Science. https://towardsdatascience.com/top-10-map-types-in-data-visualization-b3a80898ea70
- Eagan, J. (n.d.). Tufte's Design Principles. Telecom Paris Tech. https://perso.telecom-paristech.fr/eagan/class/ces-ds/notes/4%E2%80%94Tufte-Design.pdf

Analyse numerical visualisation

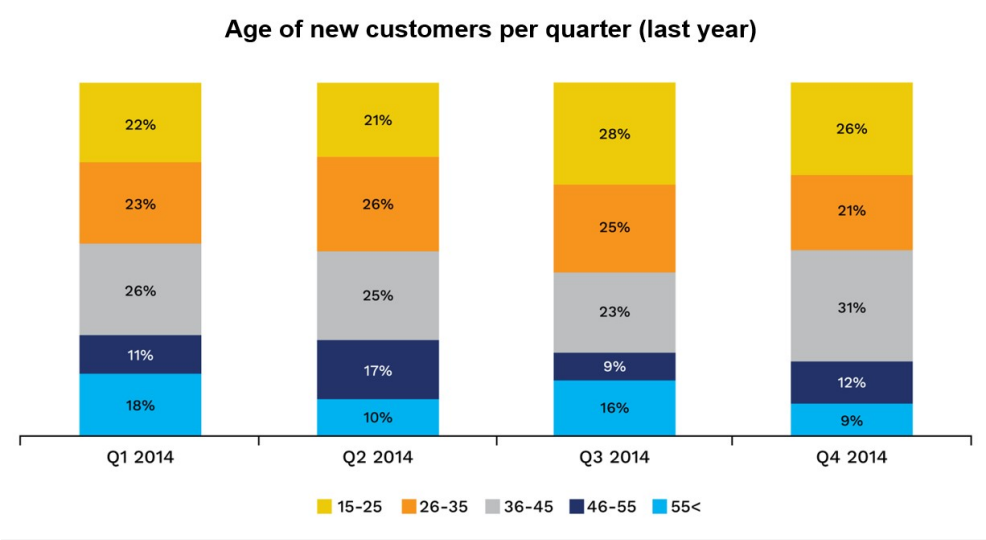
Activity: Stacked bar charts and maps
Time: 30 minutes

Purpose: To investigate and analyse numerical visualisation

Task : Evaluate and complete the two tasks below

Question 1:

Stacked charts handle part-to-whole relationships. This is when you are comparing data to itself rather than seeing a total – often in the form of percentages. Observe the following graph and analyse the story it is telling. Explain your findings in brief.

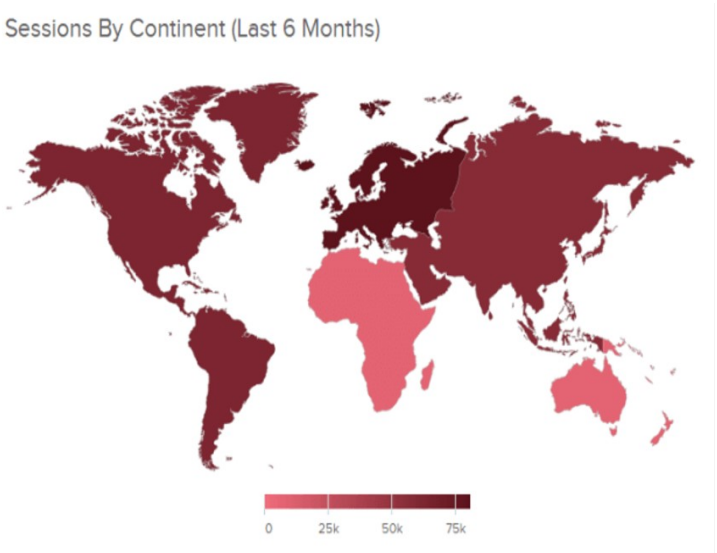


Source: Sandra Designing Charts and Graphs: How to Choose the Right Data Visualization Types (Durcevic, 2019)

Source: Sandra Designing Charts and Graphs: How to Choose the Right Data Visualization Types (Durcevic, 2019)

Question 2:

The following map shows the most recent Zika outbreak. Based on the map, can you list the benefits of using mapping techniques for this kind of sample data? You can also visit the following site to investigate further to support your answer, https://wwwnc.cdc.gov/travel/page/zika-travel-information



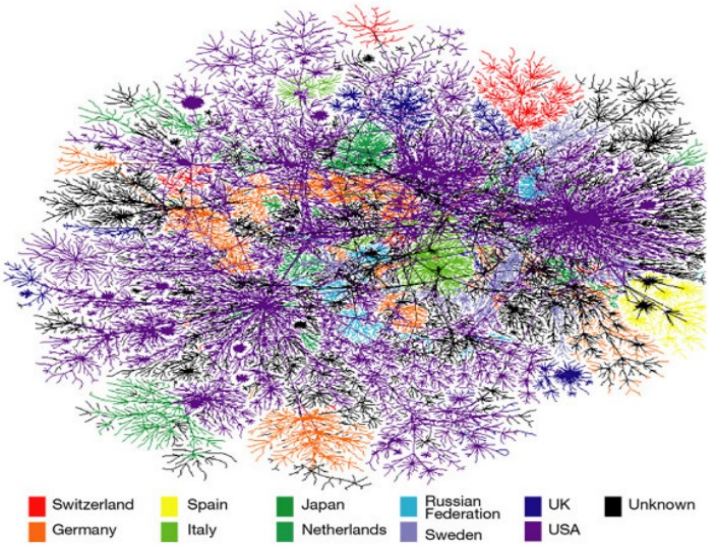
Source: Sandra Designing Charts and Graphs: How to Choose the Right Data Visualization Types (Durcevic, 2019)

5.2 Visualising non-numerical data

In this section, you will explore more about networks. But, what is a web network?

“Any collection of objects in which some pairs of these objects are connected by links” (Easley and Kleinberg, 2011)

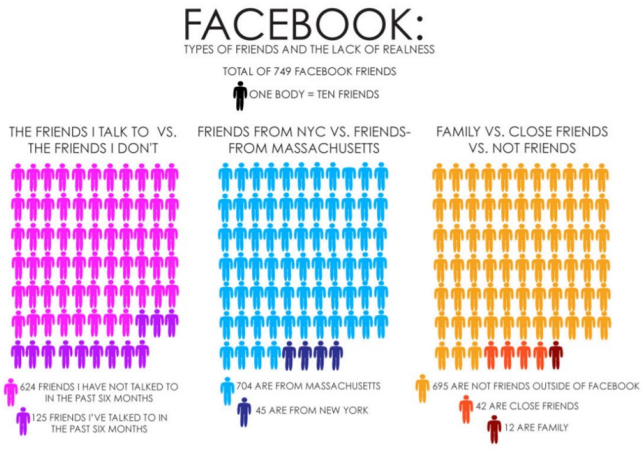
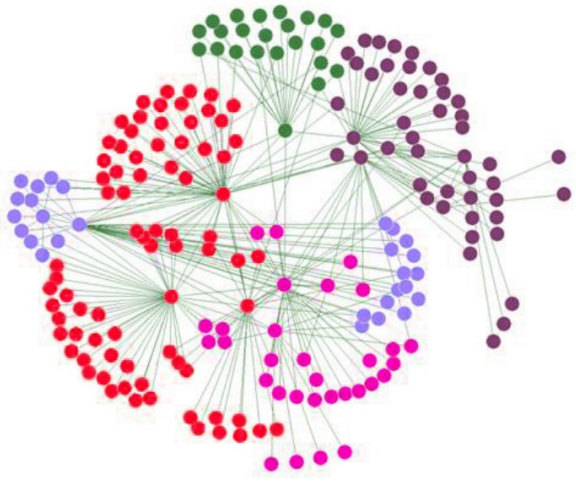
The image on the right demonstrates the Email paths on the Internet.



Source: Graphs, Trees, and Networks (SFU, 2016)

Activity: Identifying networks

Observe the two figures below and identify if they are a network. Justify your answer in brief.

Figure 1	Figure 2
	
Source: Personal Network Infographic (Kokos, n.d.)	*Source: Graphs, Trees, and Networks (SFU, 2016)*

Visualising non-numerical data

Explore: Graphs and network analytics

A graph (network) is a set of objects that are connected. Graph theory provides the formal foundation for network analysis across domains and provides a familiar language for describing the structure of networks.

Network Visualisation (also called Network Graph) is often used to visualise intricate relationships between many elements. A network refers to an object composed of elements and relationships or connections between those elements. A network visualisation displays undirected and directed graph structures.

This type of visualisation illustrates relationships between entities. Entities are displayed as round nodes, and lines show their relationships. The colourful display of network nodes can highlight non-trivial data inconsistencies that may be otherwise overlooked. The following are important terms to know when exploring graphs and network visualisation:

1. Objects = nodes, vertices
2. Connections = links, edges
3. A path is the/a set of connections from node a to node b
4. Topology is a schematic description of the arrangement of a graph/network, including nodes and links

Source: Graphs, Trees, and Networks (SFU, 2016)

Network analysis is a collection of techniques for examining the **relationships** between entities and depicting the **structure** of those relationships. Network analysis spans several domains, including social networks, bibliometrics, epidemiology, bioinformatics, complex systems, and text analysis.

Components of a network

Vertex - a set of objects (also called nodes) that are connected.

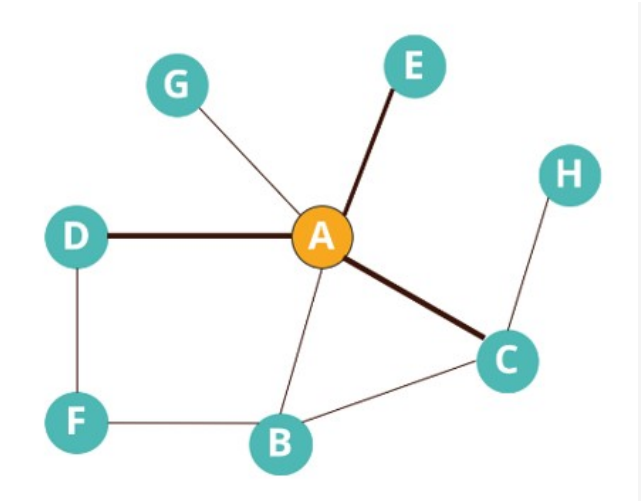
Vertex attributes define a vertex based on its characteristics. E.g. For airline routes, if Vertex is a city, attributes could be the city's population.

Edge - The connections between the nodes are called edges or links.

If the edges in a network are directed, i.e., pointing in only one direction, the network is called a directed network.

If all edges are bidirectional or undirected, the network is undirected.

The thickness of the edge determines the relationship between the two related vertices.



Source: Sample Network (The Math Company, 2018)

Examples of nodes and their components (The Math Company n.d)

Network	Vertices	Vertex attributes	Edges	Edge attributes
Airlines network	Airports	Footfall, terminals, staff, city population, international/domestic, freight, hangar capacity	Airplanes/Routes	Frequency, a passenger, plane type, fuel usage, distance covered, empty seats
Banking network	Account holders	Name, demographics, KYC document, products, account status, balance, and other details	Transactions	Type, amount, authentication (pass/OTP), time, location, device
Social network	Users	Name, demographics, # connections, likes, circles belong to, subscriptions	Interactions	Medium (like, comment, direct message), time, duration, type of content, topic
Physician network	Doctors	Demographics, speciality, location,	Patients	Demographics, diagnosis history, visit

		affiliation (type and size), weekly patient intake		frequency, purpose, referred to, insurance
Supply chain network	Warehouses	Location, size, capacity, storage type, connectivity, manual/automated	Trucks	Load capacity, # wheels, year of make, geographical permit, miles travelled. Maintenance cost, driver experience

Tree Maps

Tree maps are a hierarchical data visualisation. Each category is divided into segments that represent a whole. Each branch is a rectangle within a tree map, which is then associated with smaller rectangles (or sub-branches). The rectangles, or sub-branches, are sized proportionally to the data.

Tree maps are ideal for displaying large amounts of hierarchically structured (tree-structured) data. The space in the visualisation is split up into rectangles that are sized and ordered by a quantitative variable. The levels in the hierarchy of the tree map are visualised as rectangles containing other rectangles. Each set of rectangles on the same level in the hierarchy represents a column or an expression in a data table. Each rectangle on a level in the hierarchy represents a category in a column. For example, if you have a system of measurement that represents total car sales for three different car sales agents, the agent with the least number of sales would have the smallest rectangle.

Tree maps are an excellent way to visualise and drill down your data into layers to show the hierarchical relationship between items. Tree maps should be used to:

- see patterns in the branches due to the correlation between colour and size
- display large data sets simultaneously while making efficient use of space.

Multidimensional scaling

Structured data typically consists of data studies represented by rows and features or data attributes represented by columns. Each column can also be called a specific dimension of the dataset.

Most common data types include continuous, numeric, and discrete, categorical data. Hence any data visualisation will depict one or more data attributes in an easy to understand visual like a

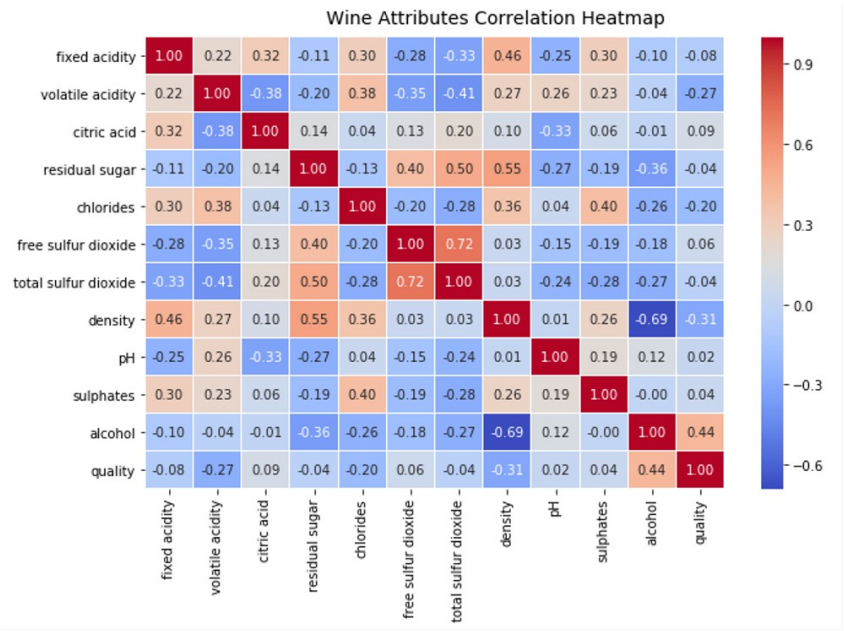
scatter plot, histogram, boxplot and so on, i.e., both univariate (one-dimension) and multivariate (multi-dimensional) data visualisation strategies.

a. **Multivariate analysis**

Multivariate analysis analyses multiple data dimensions or attributes (two or more). The multivariate analysis involves testing distributions and possible relationships, patterns, and correlations amongst these attributes. You can also perform inferential statistics and hypothesis testing.

b. Visualising data in two dimensions (2-D)

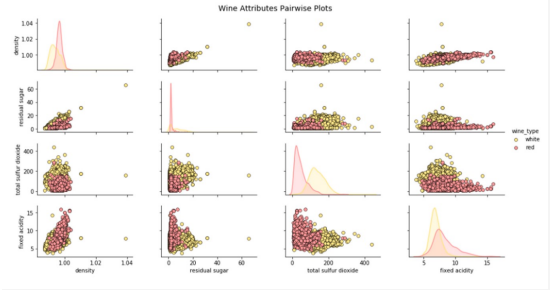
One of the best ways to decipher potential relationships or correlations amongst the different data attributes is to leverage a pair-wise correlation matrix and depict it as a heatmap.



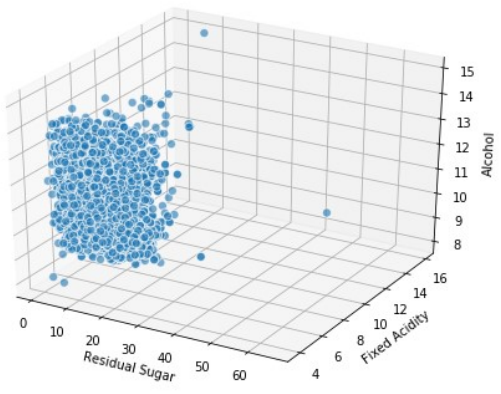
Source: Visualising two-dimensional data with a correlation heatmap (Sarkar, 2018)

c. Visualising data in three dimensions (3-D)

Considering three attributes or dimensions in the data, you can visualise them by considering a pairwise scatter plot and introducing the notion of colour or hue to separate values in a categorical dimension.



Source: Visualising three-dimensional data with scatter plots and hue (colour) (Sarkar, 2018)

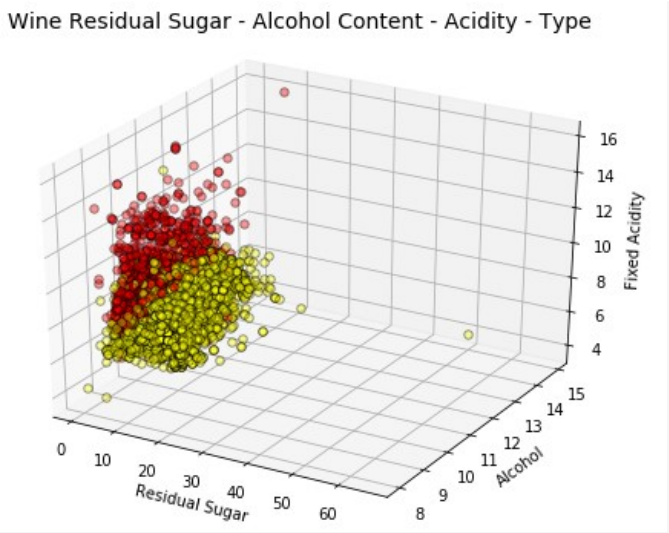


Source: Visualising three-dimensional numeric

	data by introducing the notion of depth (Sarkar, 2018)

d. Visualising data in four dimensions (4-D)

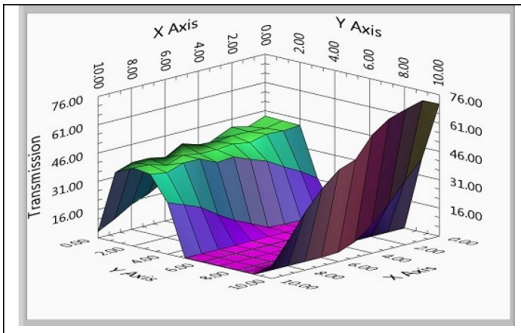
You can leverage various elements of the charts to visualise multiple dimensions. One way to visualise data in four dimensions is to use depth and hue as specific data dimensions in a conventional plot like a scatter plot. Visualising data in four dimensions leveraging scatter plots and the concept of hue and depth.



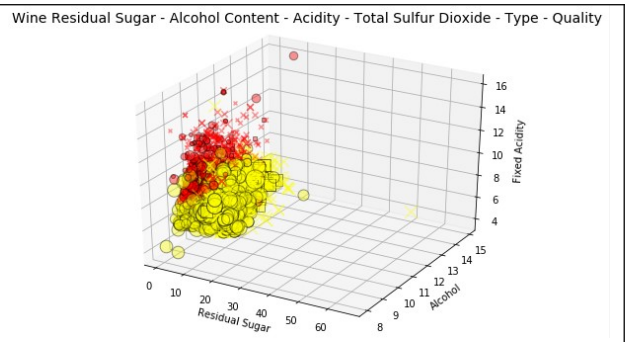
Source: Visualizing data in four-dimensions leveraging scatter plots and the concept of hue and depth (Sarkar, 2018)

e. Visualising data in five dimensions (5-D) & (6-D)

A similar strategy to visualise data in five dimensions follows where you pull various plotting components. Use depth, hue, and size to represent three data dimensions besides regular axes representing the other two dimensions.



Source: 3D mapping of LCD colour mask (Wiley Analytical Science, 2016)



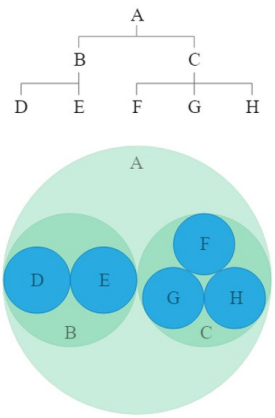
Source: Source: Visualizing data in five-dimensions leveraging bubble charts and the concept of hue, depth and size (Sarkar, 2018)

5.2.1 Packing

Introduction

Packing is a method to visualise large amounts of hierarchically structured data. Tangent circles represent brother nodes at the same level. To visualise the hierarchy, all children of a node are packed into that node and represent its size). The size of a leaf node can represent a random property, such as file size. An advantage of this algorithm is the fine outline of large data sets and the clear representation of groupings and structural relationships.

Circle Packing is a variation of a Tree map that uses circles instead of rectangles.



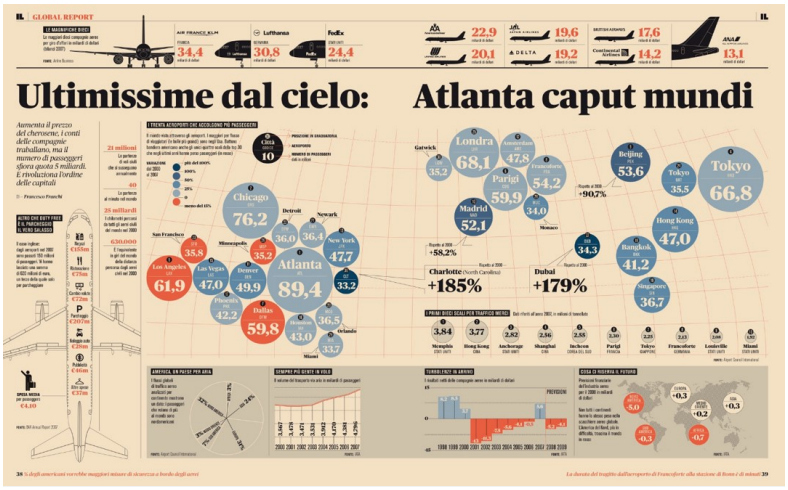
Source: Circle Packing (Data Visualisation Catalogue, n.d.)

Containment within each circle represents a level in the hierarchy. Each branch of the tree is represented as a circle, and its sub-branches are represented as circles inside it.

Containment within each circle represents a level in the hierarchy. Each branch of the tree is represented as a circle, and its sub-branches are represented as circles inside it.

As wonderful as Circle Packing appears, it's not as space-efficient as a Tree map, as there's a lot of space within the circles, but still, Circle Packing reveals hierarchical structure better than a Tree map.

Example 1



Source: Analisi grafica (Franchi, n.d.)

Example 2

Visual Capitalist (https://www.visualcapitalist.com/worlds-100-valuable-brands-2018/) created this circle packing diagram with tightly organised circles to display the World's 100 most valuable brands in 2018. The brand circles vary in size and colour according to their valuation and industry. Each circle features a clean, white-scale version of the company logo, making this chart a snap to read.

By organising the data in this way, you can see who the biggest players and industries are at-a-glance. You can also dig deeper into the data to analyse the total brand value for individual companies or whole industries. You can find the valuation for the companies who don't have it listed in their circle below in a regular table.

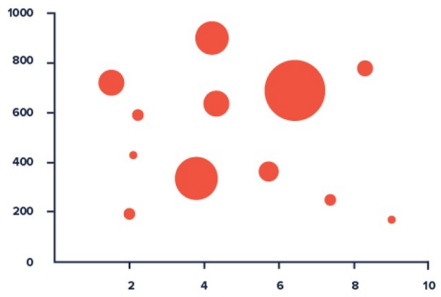
Source: The World's 100 Most Valuable Brands in 2018 (Visual Capitalist, 2018) 5.2.5 Bubble plot

Bubble Chart

A bubble chart is a type of chart that displays three dimensions of data.

Each entity with its triplet (v1, v2, v3) of associated data is plotted as a disk that expresses two values through the disk's xy location and the third through its size.

Bubble charts can facilitate understanding social, economic, medical, and other scientific relationships. Bubble charts can be considered a variation of the scatter plot, in which the data points are replaced with bubbles.



Source: Tableau your Data (Milligan, 2019)

The following are some of the best practices for bubble charts:

1. Identify all axes, including the third, which might need an additional label.
2. Use hovers to allow the user to view supplementary information about each data point.
3. If data points are close together, consider using an empty circle rather than a ‘dot’ so that overlapping points can be seen.
4. If time is one of the dimensions, place it on the x-axis for maximum clarity.
5. Using size for the third dimension makes a huge variation.
6. If differences are small but noteworthy, consider using colour or shape instead.

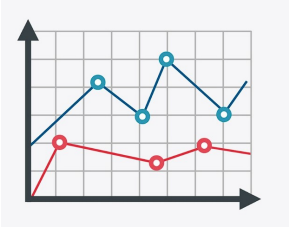
Design effective data visualisation for key stakeholders using appropriate sample data

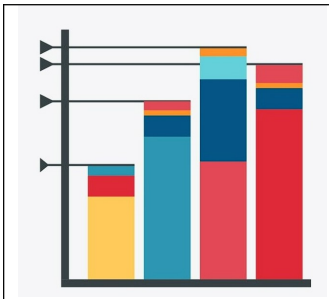
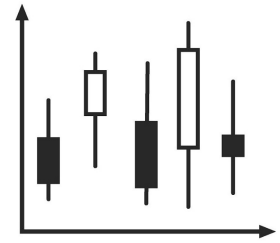
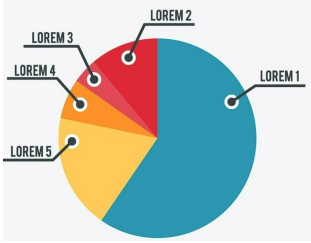
The table in this section displays the effectiveness of different visuals across data types. To better understand the table, you need to understand how variables (attributes from the data) can be categorised into different data types.

As discussed earlier, **Categorical variables** are the ones that don’t have any order, e.g., Gender, Grades, Marital Status, hair colour, etc. **Numerical Variables** are segmented into Ordinal and Quantitative variables.

Ordinal variables are categories that can be ranked. E.g., Satisfaction (Good, Bad, and Average), Potential (High, Medium, and Low), etc. **Quantitative variables** are the ones that can take any range of numeric values between -infinity to +infinity. E.g., Age, Salary, Revenue, Sales, etc.

The figure below illustrates how different graphs can be used to visualise designs in the data, considering the variable's data type.

Chart	Name	Analysis	Examples
	**Chart**: Scatter plot/Line plot **X-axis**: Continuous **Y-axis**: Continuous	• Understanding linear, non-linear relationship between two variables • Trend analysis, change in KPI over time	• How does the heart range change with age. • How does the sales of a company vary over time.
	**Chart**: Bar graph **X-axis**: Categorical/Discrete Continuous **Y-axis**: Continuous	How Y (can be any performance indicator) varies across different categories?	How sales in 2019 varied for different mobile phone brands. i.e. Mobile phone brand is the category and sales is the KPI.

	**Chart:** Stack bar graph **X axis:** Categorical **Y axis:** Continuous	Relative comparison of multiple categories within a category	Comparison of revenue generated by Apple, Samsung and Xiaomi across different products like mobile phones, laptops, televisions and headsets.
	**Chart:** Box plot **X-axis:** Continuous **Y-axis:** Continuous	• Outlier detection • Analysing data distribution across Median and Inter Quartile range	How different sales figures across a year are distributed.
	**Chart:** Pie chart **X-axis:** Categorical and Continuous **Y-axis:** Categorical and Continuous	Relative comparison of different categories for one single entity in terms of proportion/ percentages	What percentage of sales in 2019 is constituted by different products under Apple?
	**Chart:** Histogram plot **X axis:** Continuous **Y axis:** -	How does the distribution of values of x vary across different range buckets?	Distribution of income across income buckets for developing countries.

The following videos are recommended for further understanding.

- Watch this 3-minute 37-second video to learn how to create a packed circle chart. https://youtu.be/1rnU9Sn1xnA *Source: (Toan Hoang, Tableau Visionary, 2019)*

- Watch this video of 1 minute 25 seconds to learn more about how Circle Packing charts work. https://youtu.be/hDAA_YiOnac *Source: (The DataViz Cat, 2017)*

Read the following to extend your understanding.

- *Yau, N. (n.d.). Network Visualization. FlowingData. Retrieved March 15, 2022, from* https://flowingdata.com/category/visualization/network-visualization/
- *Fitzgerald, B. (February, 2016). Which Graph Should You Use? Data Visualization. Retrieved March 15, 2022, from* https://hampdatavisualization.wordpress.com/category/networksgraphs/

5.2.2 Map charts

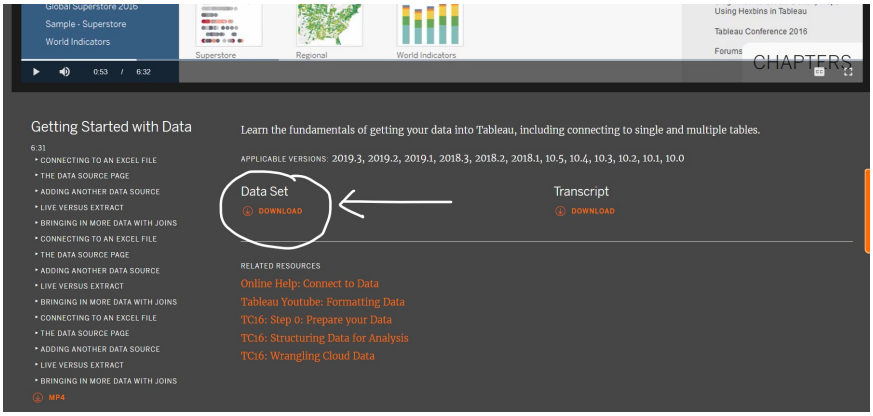
Activity: Map charts

Time: 30 minutes

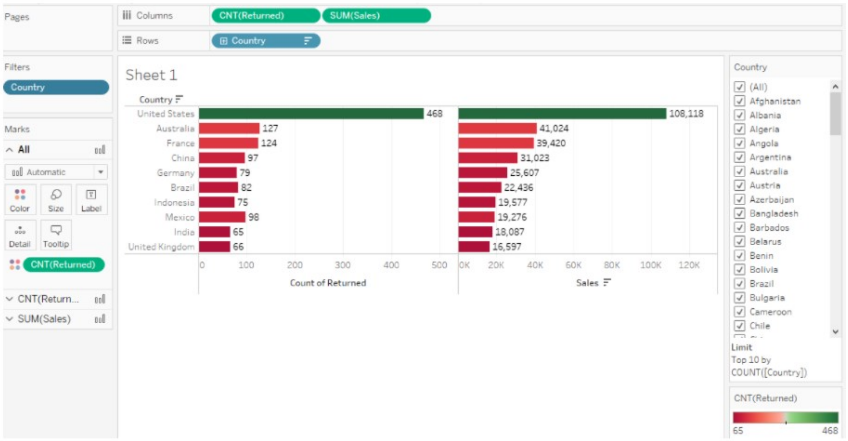
Purpose: To create a map chart from the given dataset

Task:

1. Download the dataset from Tableau Tutorials https://www.tableau.com/learn/training



2. Generate a list of orders returned from customers and compare it to the original sales, sort the visualisation in ascending order in terms of returned order, for top 10 countries in terms of refunds.



3. Map the top 10 countries with most of the refunds on the world map.

4. Now,

- Predict the refund for next one year category-wise.

- Predict the returned sale with lowest and actual forecast of the data. [Drag "Order Date" on Column's shelf -> Expand the year hierarchy to Month ->Drag "Category" dimension and "Sales" measure to Row shelf -> Drag "Returned" dimension to Filters shelf-> add filter. Click on "Analysis" on the top shelf ->click on "Forecast" -> select "Show Forecast".]

5. Analyse the dataset to identify a problem question where a Tree map would be the best fit.
6. Take a screengrab of your workbook output and add it to the discussion forum.

Feedback: Students must submit the output in the form of a screengrab to the discussion forum.

5.2.3 Activity: Industry survey 2021

Activity: Industry survey 2021

Time: 20 minutes

Purpose: To analyse a visualisation

Task: Analyse a Data Visualisation and describe the steps involved, the character of the dataset used and thought process behind the idea in 50-100 words.

Feedback: To help further the conversation, respond to the posts of at least 2 of your peers. You can ask them questions for clarification on their approach or provide them with feedback. Your facilitator will moderate this discussion.