Chapter 1: Introduction to Visualization

A picture is worth a thousand words –
Fred R. Bernard

Learning Objectives

- Learn about the history of data visualization.
- Understand the important concepts of data visualization.
- Learn about implementation of data visualization in various stages of data science.
- Understand the difference between exploratory and explanatory data visualization.
- Learn about various chart types such as bar chart, histogram, line chart, and word cloud.

What is Data Visualization?

- Pictorial format representing some form of collected data of the world which would help us in making decisions or navigating in the real world

- Transformation of the symbolic into the geometric (DeFanti, 1989)

History of Visualization

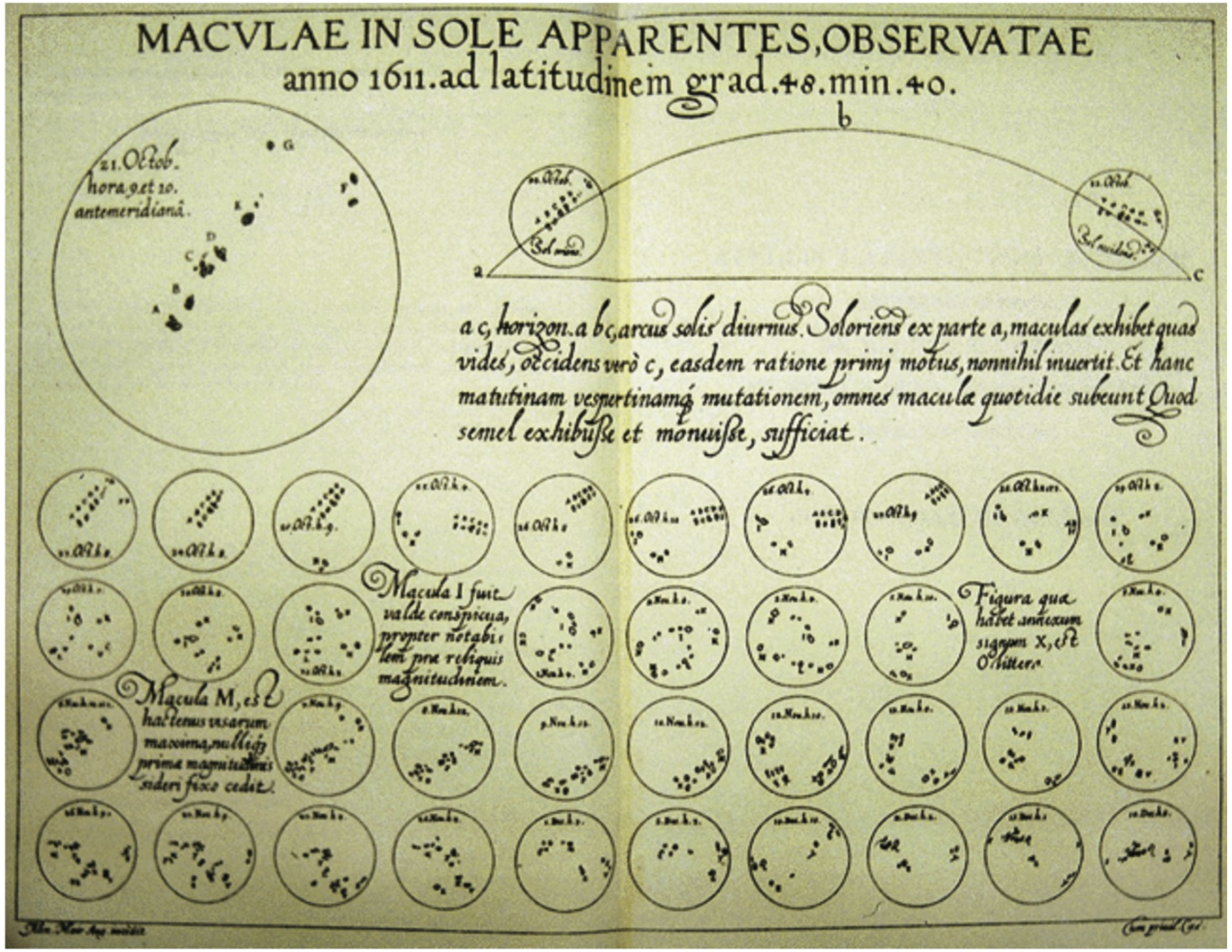

MACVLAE IN SOLE APPARENTES, OBSERVATAE
anno 1611. ad latitudinem grad. 48. min. 40.
21. Octob. hora 9. et 10. antemeridianâ.
a
b
c
Sol oriens
Sol occidens
a c, horizon. a b c, arcus solis diurnus. Sol oriens ex parte a, maculas exhibet quas vides, occidens verò c, easdem ratione primi motus, nonnihil inuertit. Et hanc matutinam vespertinamq. mutationem, omnes maculæ quotidie subeunt. Quod semel exhibuiſse et monuiſse, sufficiat.
Macula I fuit valde conspicua, propter notabilem præ reliquis magnitudinem.
Macula M, est hactenus visarum maxima, nulliq. primæ magnitudinis sideri fixo cedit.
Figura quæ habet annexum signum X, et O litteræ.

Figure 1.1 Sunspots map.
Source: http://galileo.rice.edu/sci/observations/sunspots.html

- These are *Sunspots map* from Galileo Galilei (Galilei, 1613).
- Previously, it was believed that spots are moons or planets appearing in front of the sun.
- By analyzing these maps, Galileo correctly inferred that the spots are characteristics of the Sun

History of Visualization

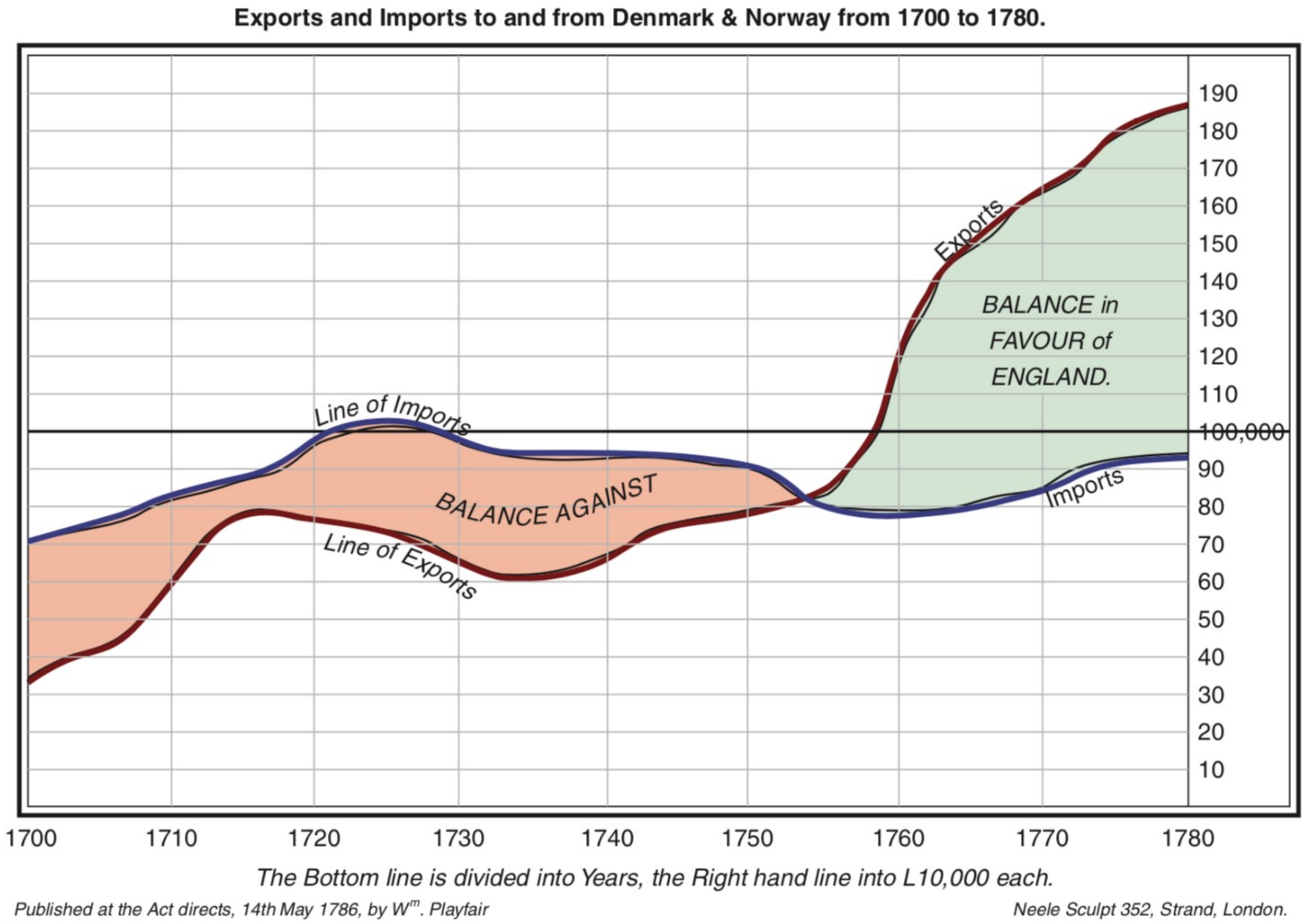


Figure 1.2 Timeseries chart on trade balance (published in *Commercial and Political Atlas*, 1786).
Source: https://en.wikipedia.org/wiki/William_Playfair

- In this line chart, we can see different trade data for England contrasted with imports and exports to Denmark and Norway.
- A line is used to show the change over time of these different quantities.
- Coloring has been used to show nominal value of “balance against” and “balance in favor” of England.

History of Visualization

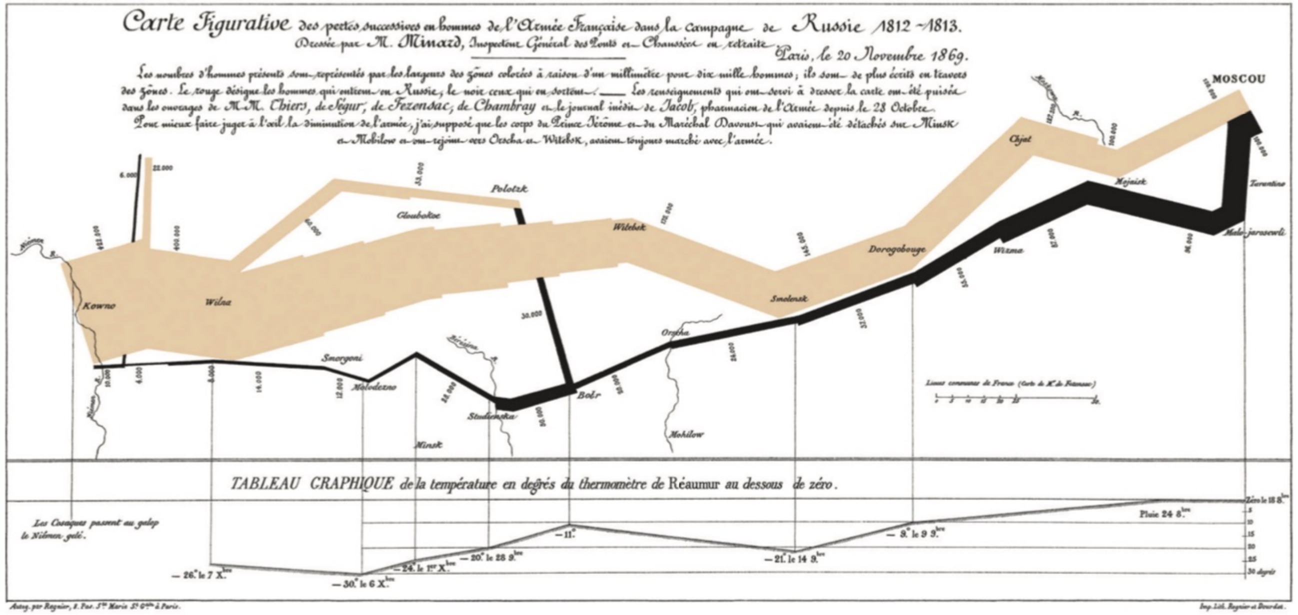


Figure 1.3 Napoleon's disastrous Russian campaign of 1812.
Source: https://en.wikipedia.org/wiki/Charles_Joseph_Minard

- This chart shows the ruthless ambition of Napoleon led to a disastrous loss of human life.
- In this graph Charles Minard cleverly combined many dimensions, such as
  - size of the army,
  - loss of life,
  - time,
  - location,
  - temperature,
  - geography, and
  - path taken in one chart

History of Visualization

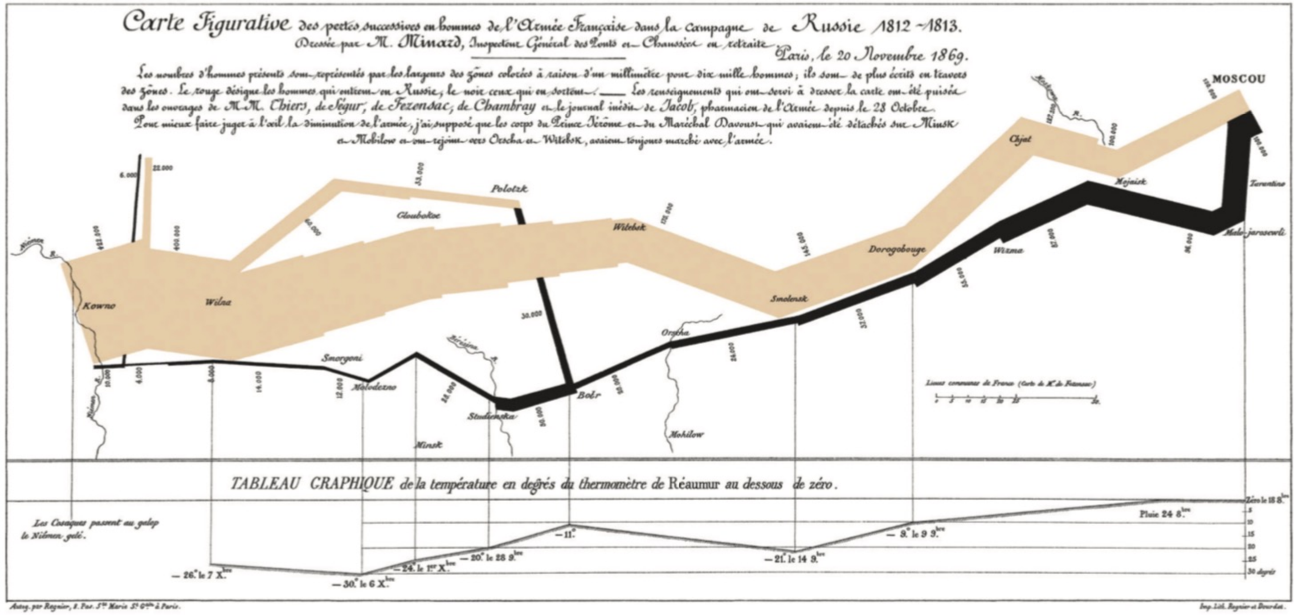


Figure 1.3 Napoleon's disastrous Russian campaign of 1812.
Source: https://en.wikipedia.org/wiki/Charles_Joseph_Minard

- Multiple dimensions have been represented such that it enables comparison
- Latitude and longitude, along with the size of the army as well as its direction represented in color and then correlated with temperature.
- Figure 1.3, referred to by Edward Tufte as the best statistical graphic ever made, is one of the best-known charts of all time.[1]

[1] ***Source:*** https://www.edwardtufte.com/tufte/posters

Why Do We Have to Visualize Data?

- A research by Dr. Richard Felder on engineering students in the 1980s suggests that visuals are powerful drivers for learning and receiving information (Felder, 1988):

1. 65% of us are visual learners (Romih, 2016)
2. We retain
   a) 80% of what we see compared to
   b) 20% of what we read
   c) 0% what we hear
3. It takes just 13 milliseconds for our brain to process an image.

Why Do We Have to Visualize Data?

- Data visualization typically represents some form of collected data of the world represented as picture which would help in
  1. Decision making
  2. Uncovering patterns and trends
  3. Presenting arguments or telling a story

- As the father of data visualization, Edward Tufte, says, “There are two goals when presenting data: convey your story and establish credibility.” (Tufte, 1983)

Why Do We Have to Visualize Data?

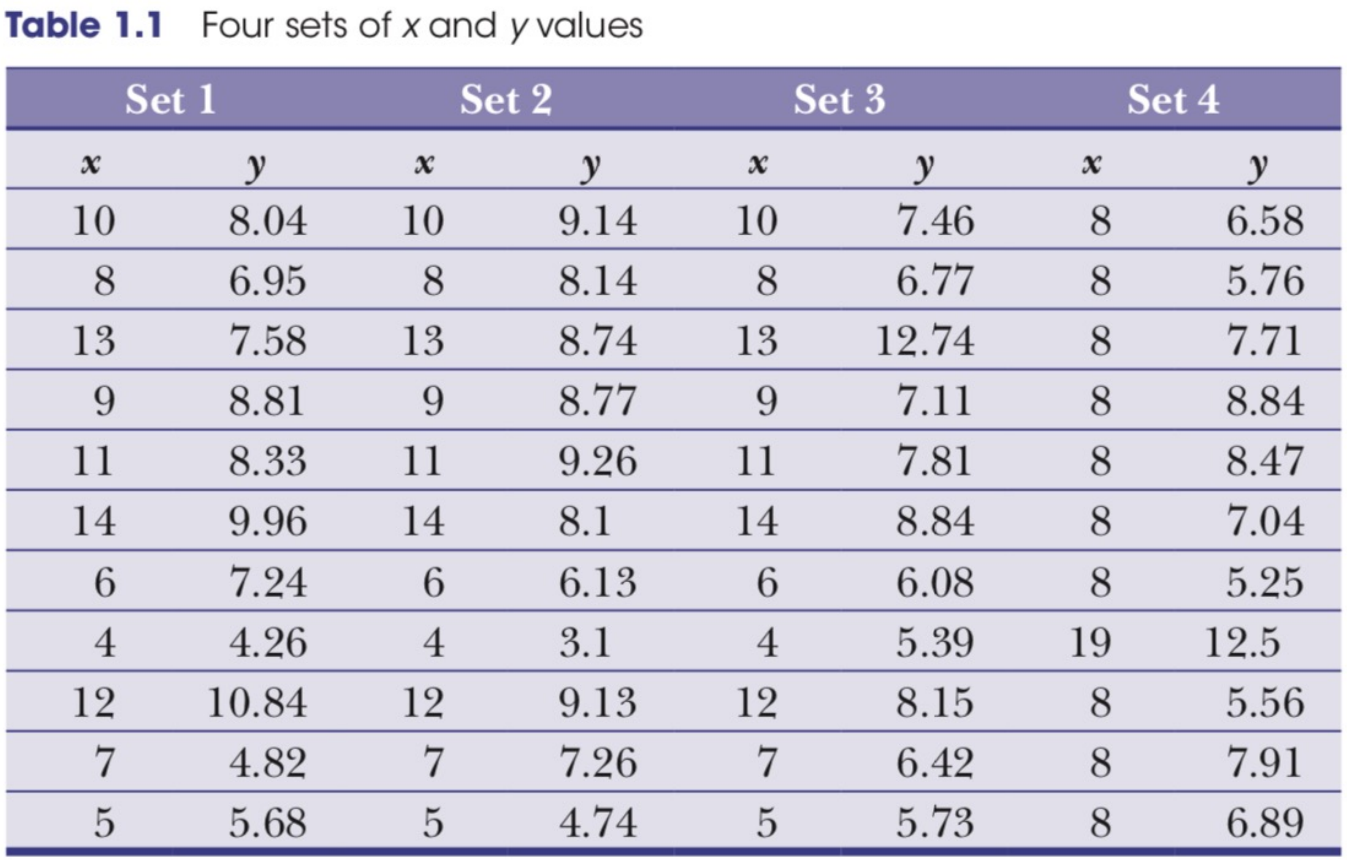
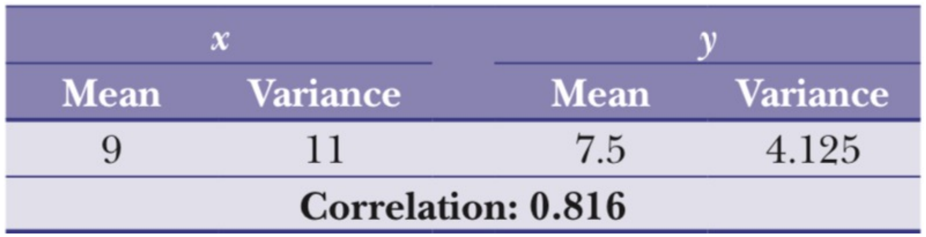


Table 1.1 Four sets of *x* and *y* values

Set 1		Set 2		Set 3		Set 4	
x	*y*	*x*	*y*	*x*	*y*	*x*	*y*
10	8.04	10	9.14	10	7.46	8	6.58
8	6.95	8	8.14	8	6.77	8	5.76
13	7.58	13	8.74	13	12.74	8	7.71
9	8.81	9	8.77	9	7.11	8	8.84
11	8.33	11	9.26	11	7.81	8	8.47
14	9.96	14	8.1	14	8.84	8	7.04
6	7.24	6	6.13	6	6.08	8	5.25
4	4.26	4	3.1	4	5.39	19	12.5
12	10.84	12	9.13	12	8.15	8	5.56
7	4.82	7	7.26	7	6.42	8	7.91
5	5.68	5	4.74	5	5.73	8	6.89



x		*y*	
Mean	Variance	Mean	Variance
9	11	7.5	4.125
Correlation: 0.816			

- Descriptive statistics are same for all the four sets of data.

- Does this mean all these datasets are identical?

Why Do We Have to Visualize Data?

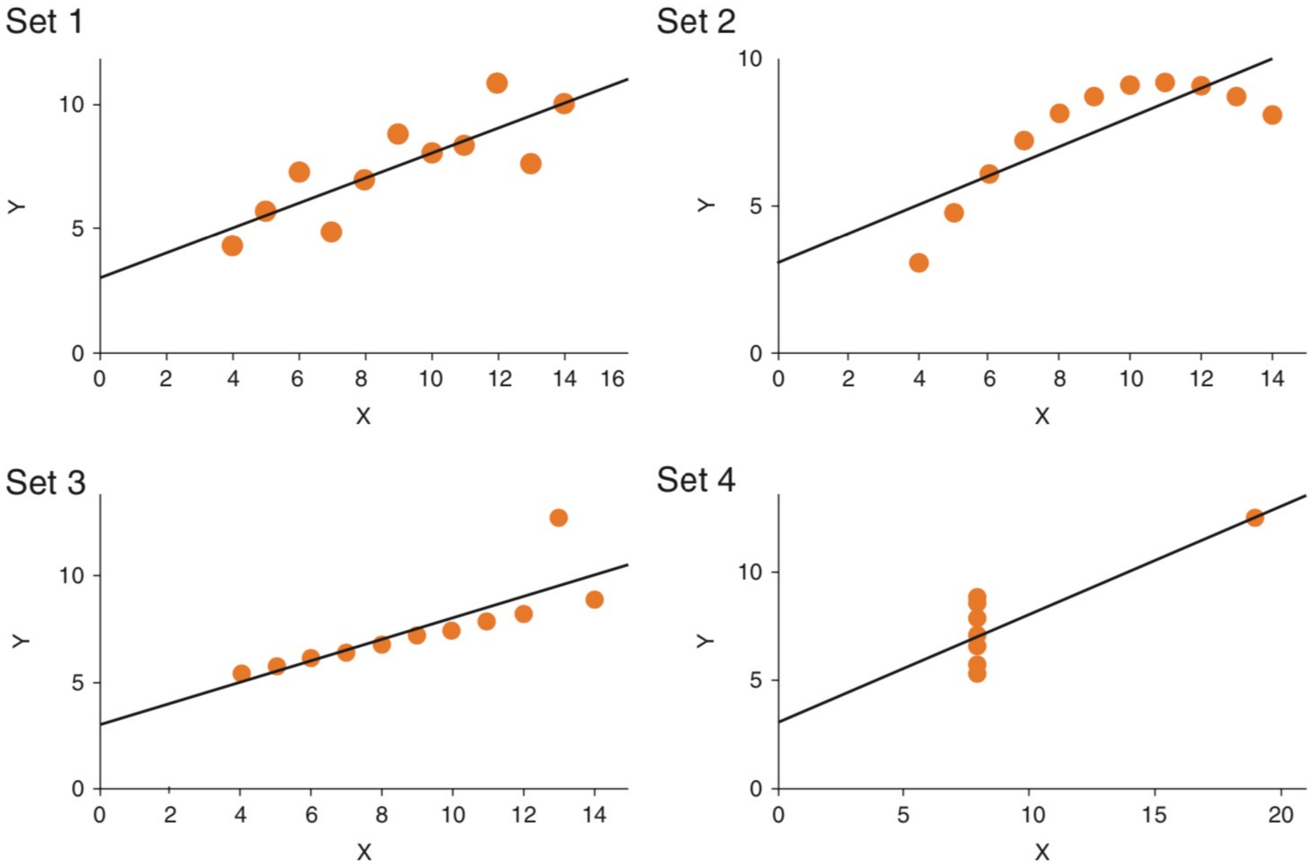


Figure 1.4 Visualizing Anscombe's quartet.

- When we plot the datasets, we can see they have different distributions.

- This example is popularly known as Anscombe's quartet (Anscombe, 1973)

Why Do We Have to Visualize Data?

- Reasons to create visualization can be (Healy and Moody, 2014; Solis, 2019)
  - Visualization compresses large volume of data into easy-to-understand visuals
  - Enables business intelligence
  - Effective for exploration of the data as well communication of insights from the data
  - Discovering answers to questions from the underlying data
  - Enabling data aid decision making
  - Understanding the data in a context
  - Finding hidden patterns in the data
  - Presenting an argument or telling a story
  - Inspiring or persuading with data story

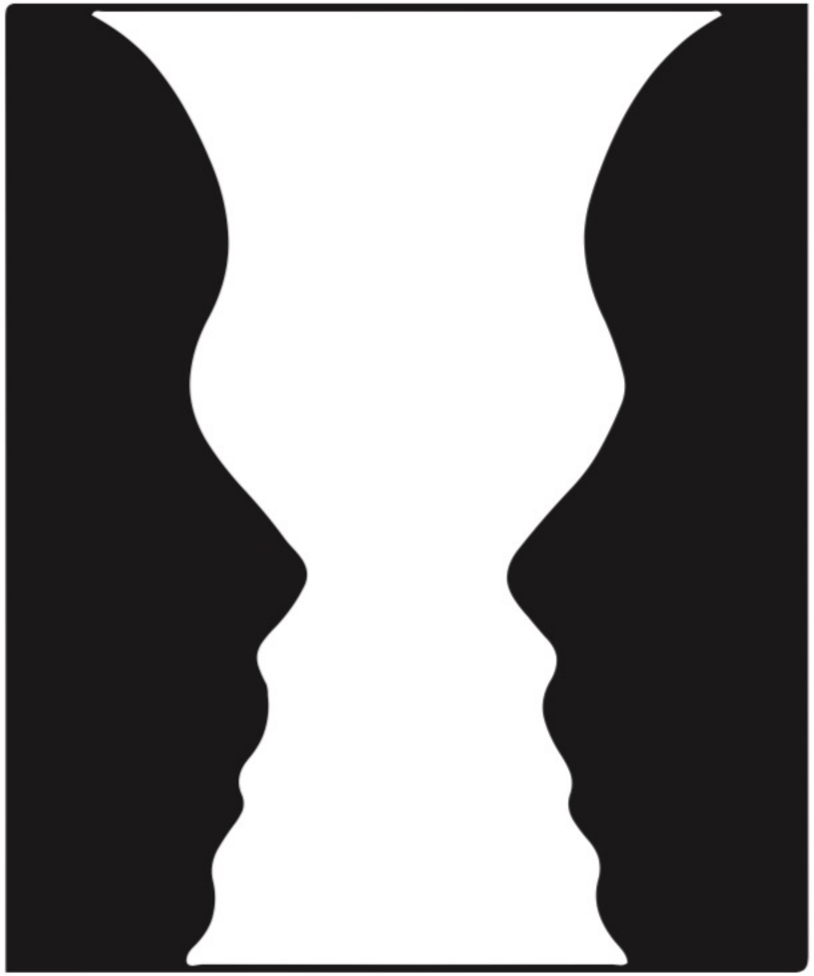
How Do We Visualize?

- German psychologists Max Wertheimer, Wolfgang Kohler, and Kurt Koffka developed Gestalt principles (Guberman, 2015).

- Gestalt principles is a collection of visual perception principles that aid us in understanding how human perception works.

- These principles explain how humans tend to interpret structure, logic, and pattern around them.

How Do We Visualize? (Gestalt Principles)

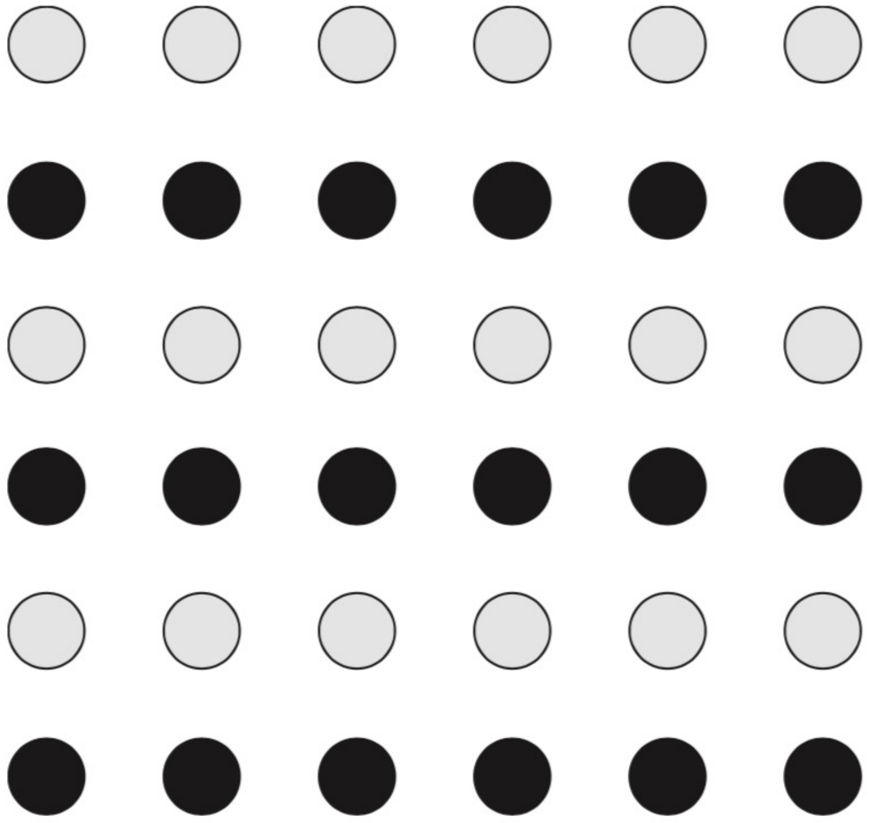


(a) Principle of figure and ground.

Principle of Figure and Ground :

- This states that when we look at a visual, our brain instinctively distinguish the objects as either they are in the foreground or in the background.
- It has been famously known as the Rubine vase faces (Rubin, 1915).
- It helps you to focus on the main elements when you develop charts and put those elements in the foreground.

How Do We Visualize? (Gestalt Principles)

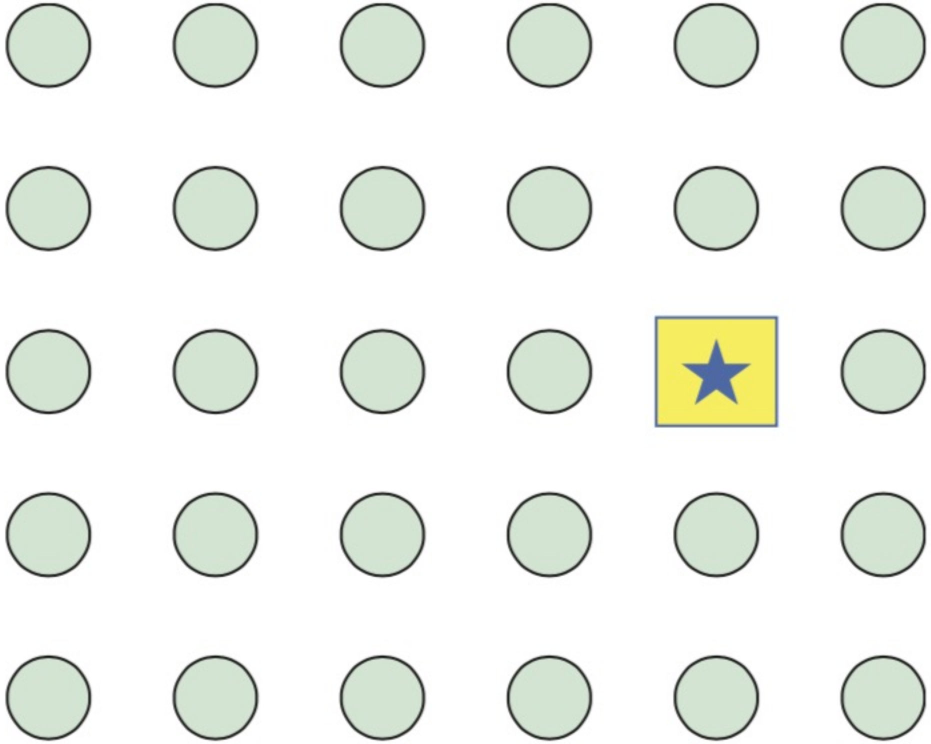


(b) Principle of similarity.

Principle of Similarity:

- This states if we group objects together then they appear to be similar.
- The circles appear to be two separate groups based on their color (black & white)
- Several visual elements can be used to show the similarity among different groups such as shape, size, or color.

How Do We Visualize? (Gestalt Principles)

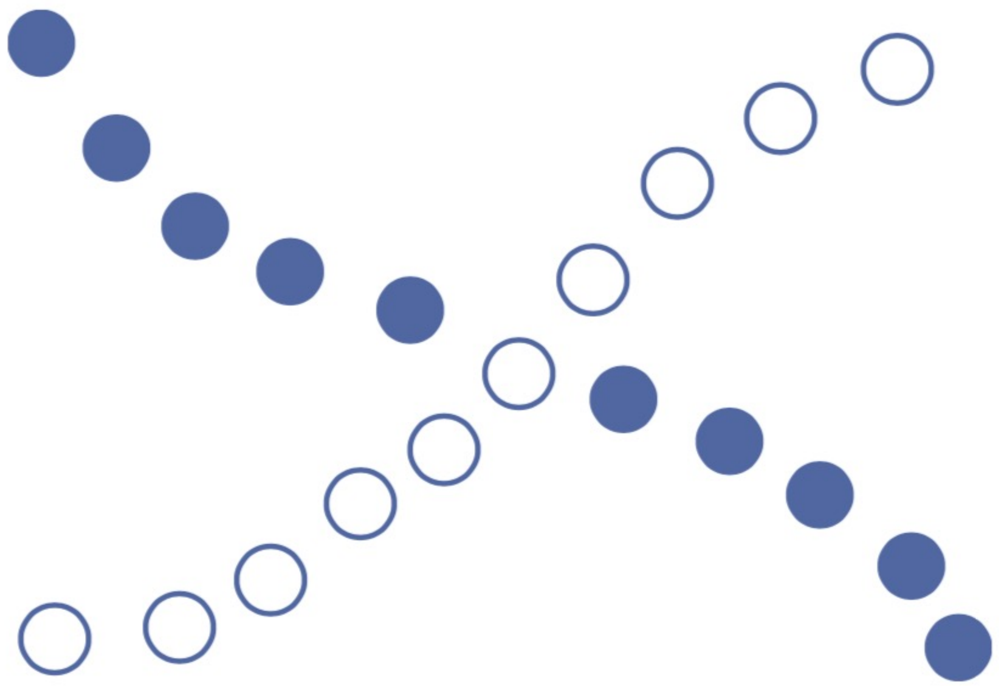


(c) Principle of focal point.

Principle of focal point:

- This states that whatever is different or distinct will be the focus point that catches the viewer’s attention.
- In the image, among multiples circles of same size and shape, the yellow squared box containing blue star catches your attention immediately.

How Do We Visualize? (Gestalt Principles)

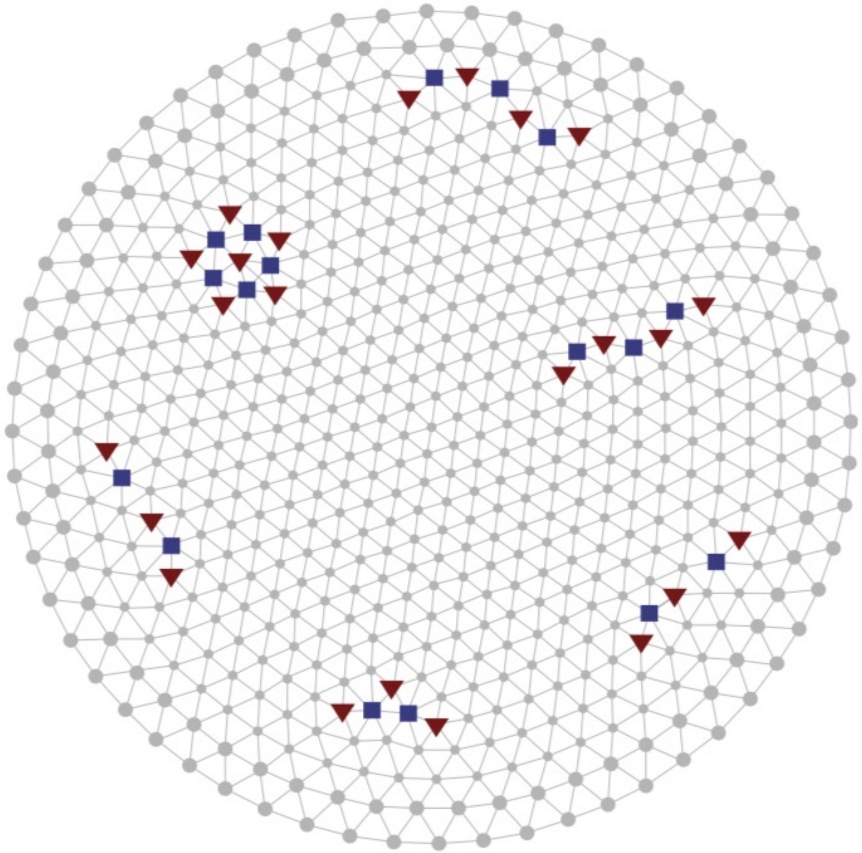


(d) Principle of continuity.

Principle of continuity:

- This states that human brain is more likely to perceive a continuous, smooth path when they visualize a line or a curve, regardless of their shape and color.

How Do We Visualize? (Gestalt Principles)

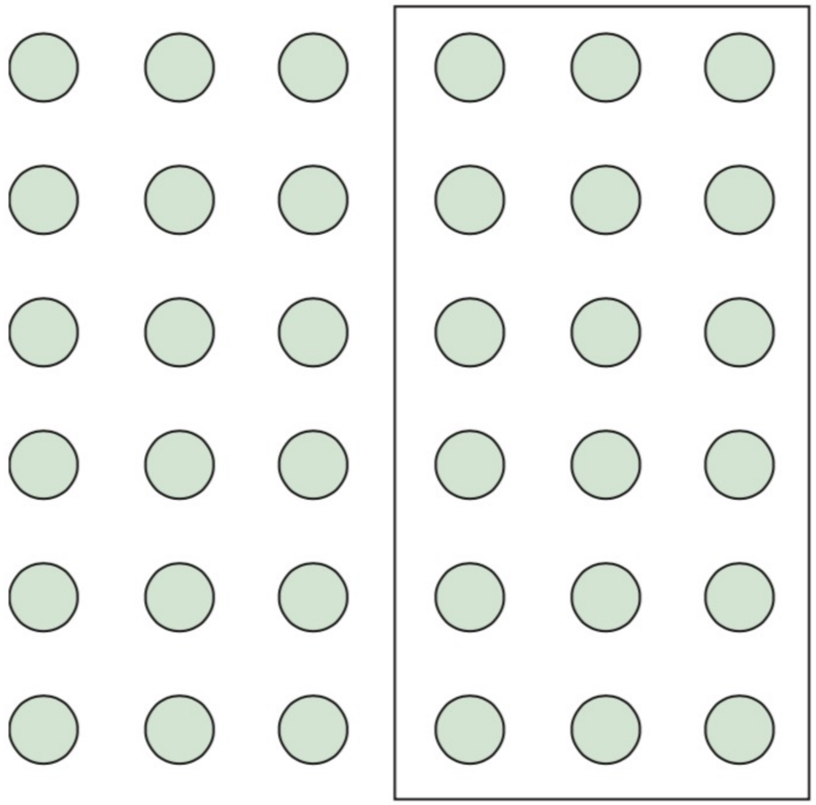


(e) Principle of proximity.

Principle of proximity:

- This states that closer elements are more likely to be similar to one another.
- This principle is so powerful that it overrides the similarity factor that might differentiate a group of things that have same color, shape, or size.

How Do We Visualize? (Gestalt Principles)

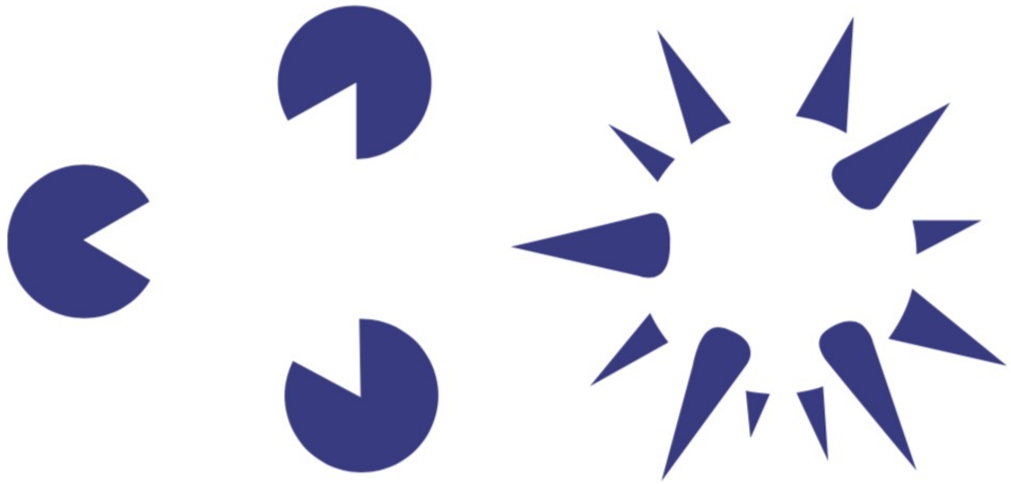


(f) Principle of common region.

Principle of common region:

- This states that things that belong to the same closed region are perceived as a group.
- By creating a boundary, we can create a perception that within that closed region things function similar.

How Do We Visualize? (Gestalt Principles)



(g) Principle of closure.

Principle of closure:

- This states that human brain tends to find a single known pattern whenever it looks at a complicated, incomplete arrangement of visual elements.
- In first image we see a triangle, but three Pac-man are arranged a distance.
- In the second image, multiple cones are placed in a way that looks like a sphere.

How Do We Visualize? (Gestalt Principles)

- These principles play an important role while building effective visualization.
- One can incorporate these principles in their visualizations to grab the audience’s attention immediately.

Seven Stages of Visualizing Data

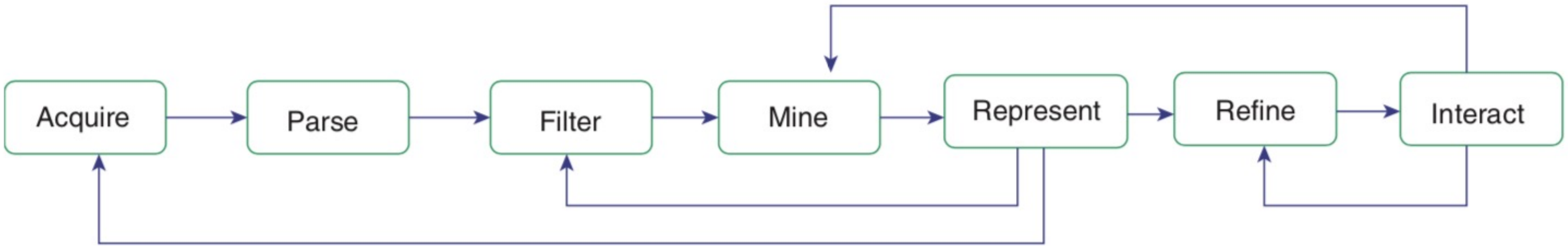


Figure 1.8 Steps to visualize data.
Source: http://media.espora.org/mgoblin_media/media_entries/1633/Visualizing_Data.pdf

1. **Acquire**: It is about how we obtain the data.
2. **Parse**: in this step, we work toward providing the required structure for the data.
3. **Filter**: Here, we remove noise and filter the data of interest.
4. **Mine**: It is about exploring and uncovering hidden patterns in the data.
5. **Represent**: start with basic visualization to discern complex relations and patterns.
6. **Refine**: Improve on the representation to turn your data into information
7. **Interact**: Provide information of your analysis to your stakeholder to interact and learn more on the underlying analysis.

Data Science Process

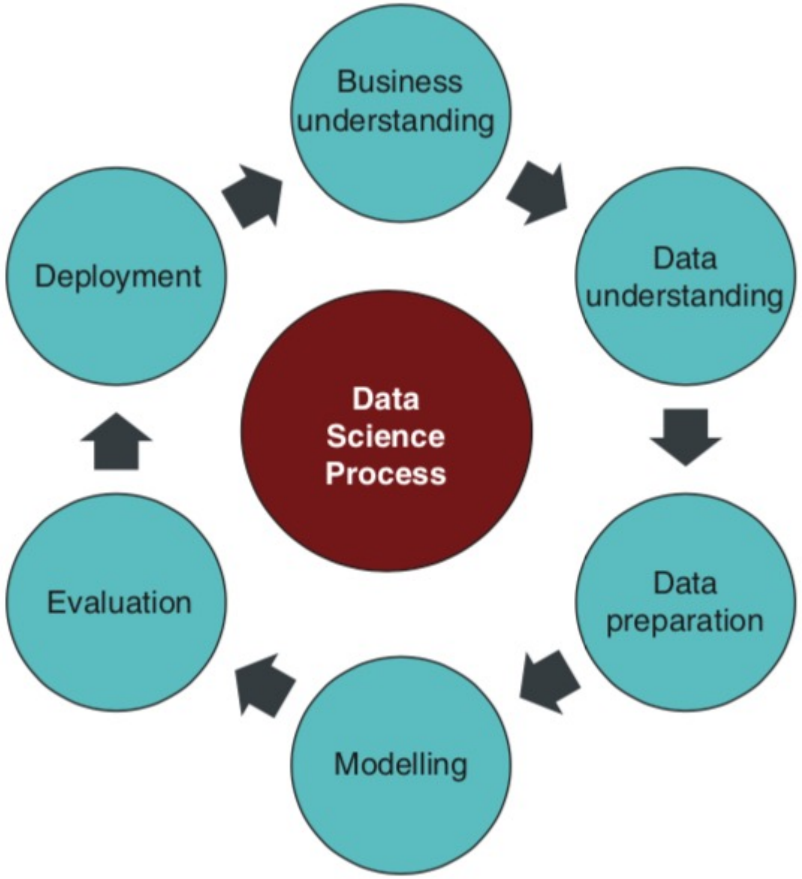


Figure 1.7 Data science process.

- Data science is the paradigm which will help us derive value from data.
- A typical data science implementation would involve processes as shown in the figure.
- At every step of the data preparation process, we can use data visualization.

Data Understanding Process

- The data understanding process involves data collection, understanding the features of data by exploratory analysis, and verifying data quality.
- Based on the structure, there are two classification of data:
  1. **Structured:** Data arranged in a tabular form or a matrix form with labelled rows and columns. For example, marks of students arranged with student ID.
  2. **Unstructured:** Any data which is not originally in matrix form is unstructured data. For example, click stream data, images, text data, etc.

Data Preparation Process

- The data preparation process involves coherently combining raw data from multiple sources.
- It can then be used for analysis and extracting valuable business insights.
- Based on the structure, there are two classification of data:
    1. Cleaning the data where we identify corrupt data, irrelevant or inaccurate records and then use methods to replace these inaccurate values.
    2. Applying domain knowledge and build a strategy on how to handle missing values. This cleaned data would then be needed further for our process.
    3. Formatting the data and building integrating data pipelines.

Modelling Process

- Data modelling involves selecting modelling techniques and assessing the models.

- Data visualization helps build our confidence in the model, and it will equip us to explain the model to others.

Evaluation Process

- This process involves evaluating results of multiple models and review processes.

- We use visualization in combination with numeric scores to build better intuition around model performance.

Deployment Process

- This involves the final product deployment, monitoring, and maintenance.

- Data visualization is utilized to continuously track the health of the model against a set of key indicators.

Usage of Visualization

- If we want to draw important insights from our data, the first step must be to analyse it thoroughly.
- The first approach should be to have a better understanding of the data using different analytics approaches, such as *Explore* or *Explain*.
- **Data *exploration*** consists of the science of drilling the data till the point it speaks.
- **Explanation** is the art of communicating those results to your data-driven audience in such a way that it makes an impact in their business.

Exploratory Analysis

Key rules that can be used in the art of exploring the data:

1. **Ask questions:** In a data exploration process, we need to dive deep into the data by asking a lot of questions. Data exploration helps us identify whether the data are right to answer our questions.
2. **Diversify your data:** A common trap people generally fall into is of not having diversified or voluminous data. Using large data from various sources let us explore many options to find an optimal answer to our questions.
3. **Look for insights:** The main aim of exploring data is to find meaningful insights that can further lead to actionable items. The outcome of data exploration would identify the key items and attributes which will answer what happened and why it happened.

Explanatory Analysis

- Explanatory analysis is the process of explaining the whys and how's in an analysis.

- Explaining the outcome in a data, which can be either visualized or modelled, to our relevant audience is called *explanatory analysis*.

- The focus of an explanatory analysis is to communicate the important insights of the data analysis process to our audience/stakeholders.

Explanatory Analysis

- Cole Nussbaumer Knaffic in the book Storytelling with Data says,

“**It can be tempting to want to show your audience everything, as evidence of all of the work you did and the robustness of the analysis. Resist the urge. Concentrate on the information your audience needs to know**” [Knaffic, 2015].

Types of Charts

- To make data visualization more intuitive and meaningful, charts and graphs should be appropriate and easily understandable.

- Visualization must be in such a manner that it can bring out all the important information your data contains.

- To make data visualization visually compelling and coherent, we need to address the question of what we want to conclude from our visualization.

Comparison

- Visualization is perfect when we want to compare one or many variables in a data, such as
  1. comparing several variables to visualize the difference between them,
  2. ranking various data categories from best to worst.
- A few examples where comparison visualization can be used are as follows:
  1. Annual revenues from previous years to show which products performed better than the rest.
  2. Quarterly/monthly sales of a product to check seasonal trends.
  3. Total number of monthly footfalls in a supermarket, grouped by the area.

Bar Chart

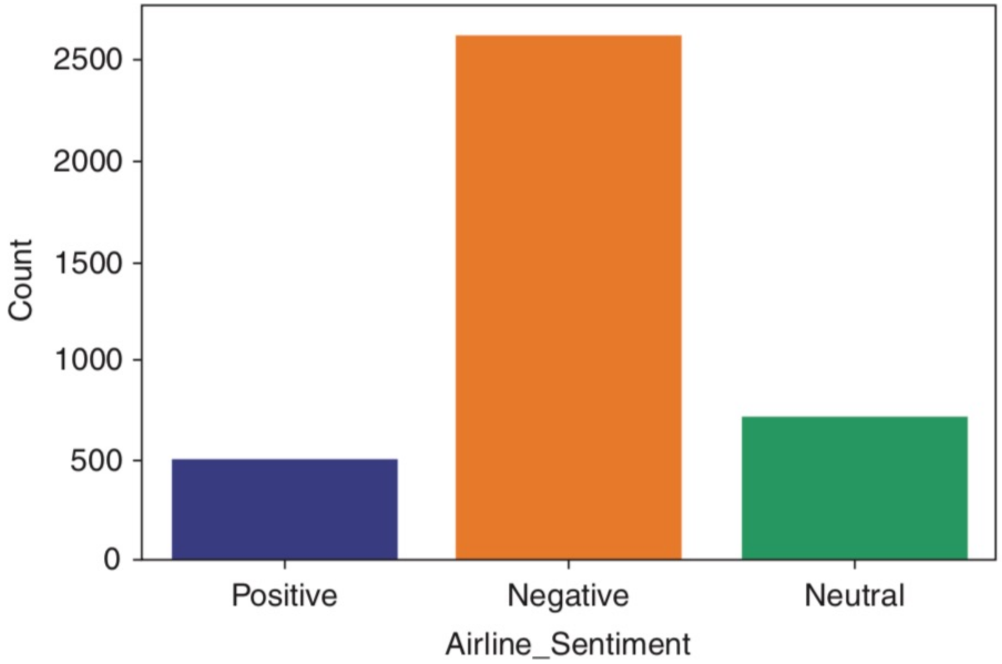


Figure 1.10 Column bar chart.

- The dataset is called "Twitter US Airline Sentiment"[2].
- The data consist of comments of passengers based on service provided by various airlines.
- The data contains the sentiment of the tweets as positive, negative, and neutral for one of the U.S. airlines, United.
- Refer Figure 1.10
- This chart provides you a clear vision of how the different variables can be compared among each other.

[2] ***Source:*** https://data.world/crowdflower/airline-twitter-sentiment

Bar Chart

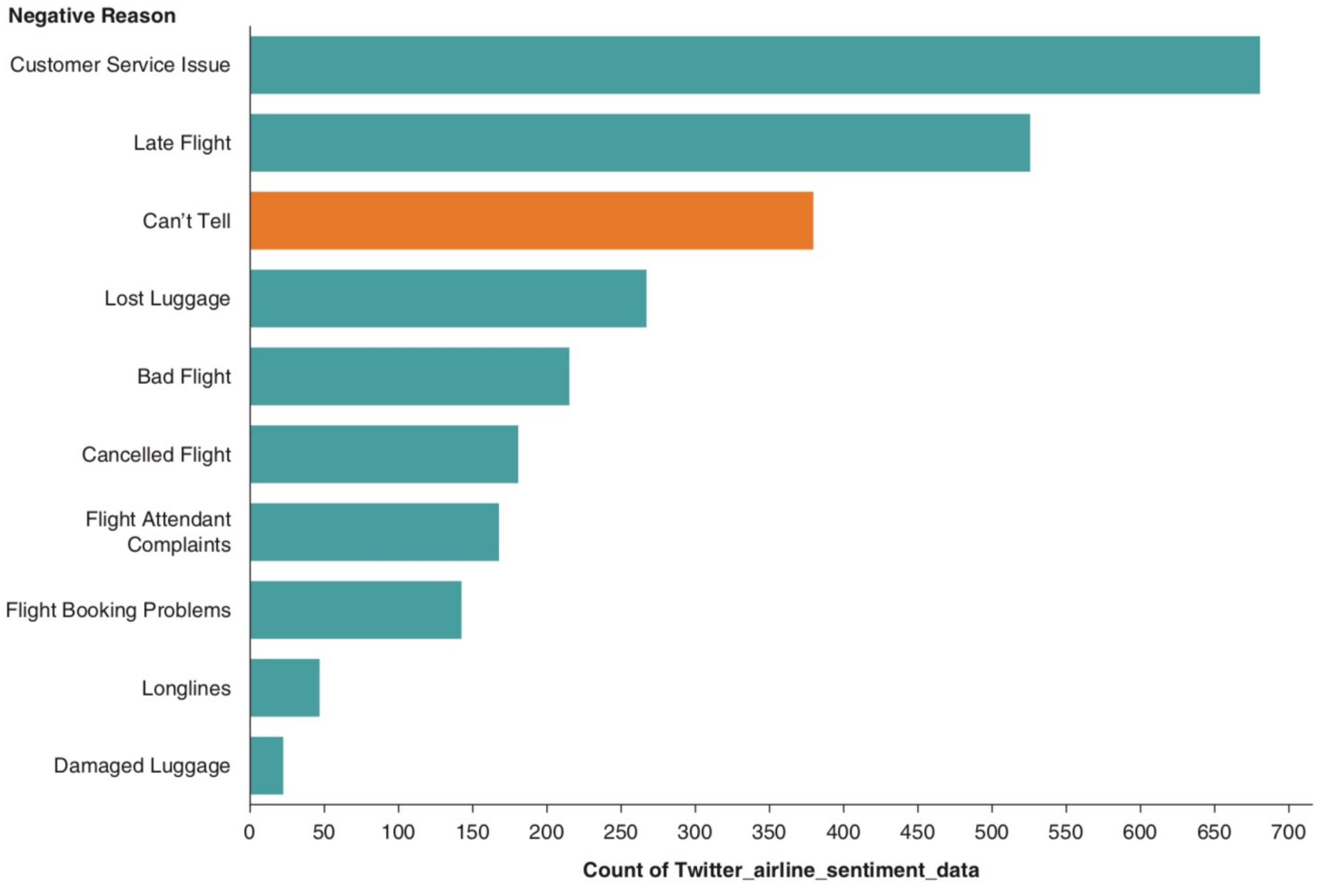


Figure 1.11 Horizontal bar chart.

- Refer Fig. 1.11
- The horizontal bar graph shows the plot for the negative sentiment reasons.

- As we can clearly see, customer service has the greatest number of reviews.

Bar Chart

- When one wants to compare similarities and dissimilarities among various variables, one can use a bar chart.
- **Horizontal bar graph:** This is a perfect way of showing a comparison among variables. But when we are dealing with too many variables, this type of graph may look cluttered.
- **Column bar graph:** When we are showing chronological data such as comparing data across various categories, column graphs are used.
- **Stacked bar graph: W**hen we are interested in comparing the data to itself rather than looking for a composition. It often shows the percentages.

Note – The stacked bar shows the percentage relative to the total and the categories may not always be mutually exclusive; hence the sum of the percentage in a bar may or may not be 100%.

Distribution

- Distribution charts help find the normal tendency of the data and in what range the data lie.
- It also helps to identify any outliers in the data.
- For example,
    - population distribution by age or sex,
    - distribution of grades of students in an exam, and
    - heights of children in a class.
- Scatter charts and histograms are most used charts for showing distribution analysis.
- They are good at representing data distribution or clustering trends and help in finding outliers and anomalies in data.

Distribution

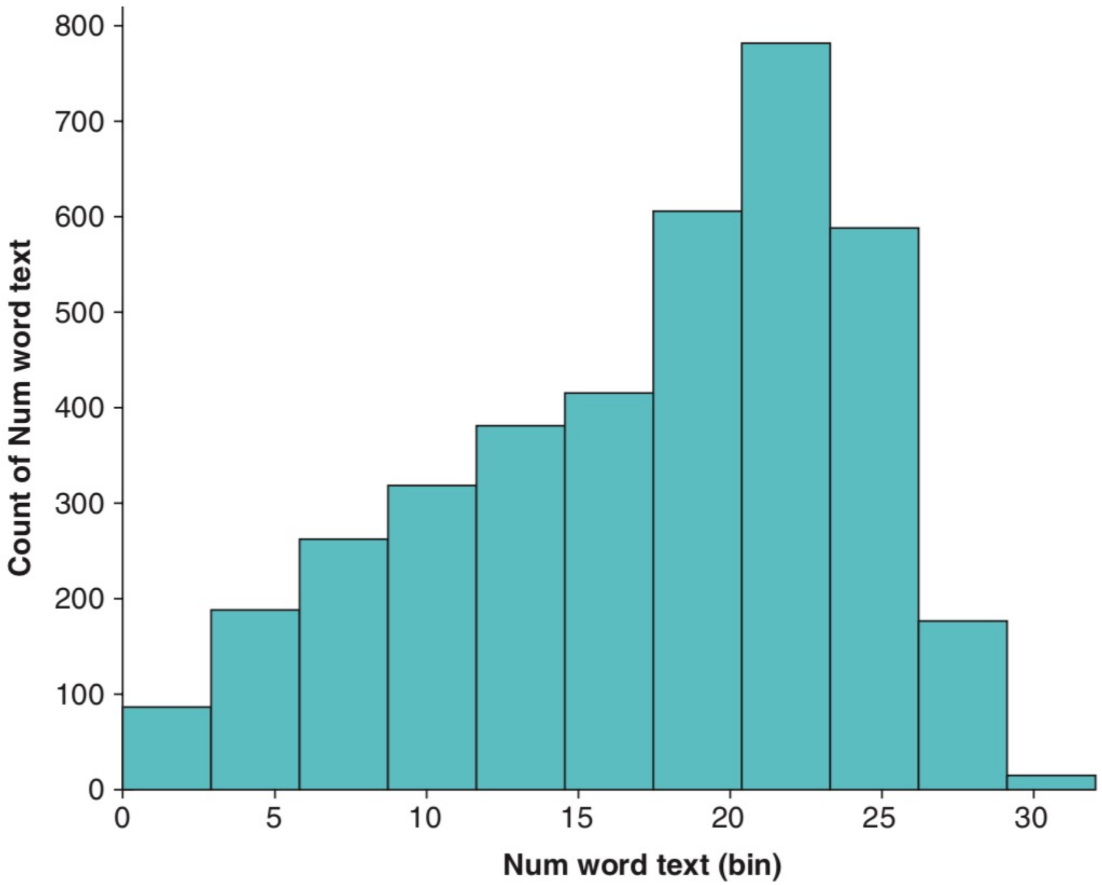


Figure 1.12 Distribution chart.

- Scatter charts and histograms are most used charts for showing distribution analysis.
- They are good at representing data distribution or clustering trends and help in finding outliers and anomalies in data.
- In Fig 1.12, we visualize how the number of words in the passenger reviews are distributed.
- It gives us an idea of how number of words in a tweet (reviews) are distributed.

Composition

- The main idea behind the visualizations that show composition in data is to understand how individual components of data comprise a whole.
- It focuses on showing the contribution of each part with respect to the total value.
- To show the part-to-whole analysis in the data, one can use pie chart, stacked charts, map-based graphs, etc.
- A few examples where composition visualization can be used are as follows:
  - Total population in a country,
  - composition based on religion or language.
  - Company market share in the market.

Composition

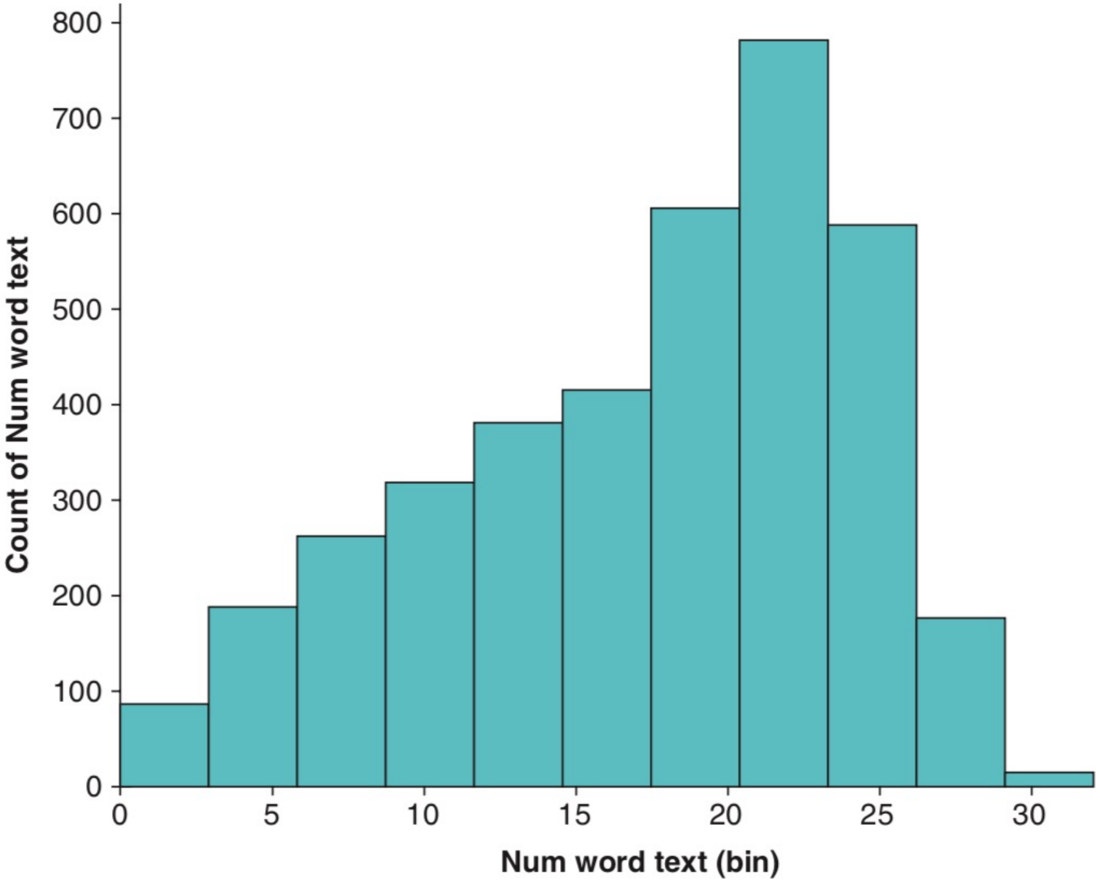


Figure 1.12 Distribution chart.

- In the Twitter dataset, the composition of the airline sentiment can be visualized using the pie chart.
- It shows the percentage of each kind of the sentiment (Fig. 1.13).
- It tells us that about 69% of all the sentiments are negative.

Relationship

- Relationship charts are used to show the relationship or correlation between different variables.
- The variables may be positively correlated, negatively correlated, or may have no correlation at all.
- Scatterplot, bubble chart, and line chart are a few charts used to show the relationship between different variables in data.
- A few examples where the relationship visualization can be used are as follows:
  - Relationship between sales and profit in marketing expense.
  - Testing hypotheses such as “Does salary depend on IQ level”, and
  - “Children who drink health drinks are taller than children who do not.”

Relationship

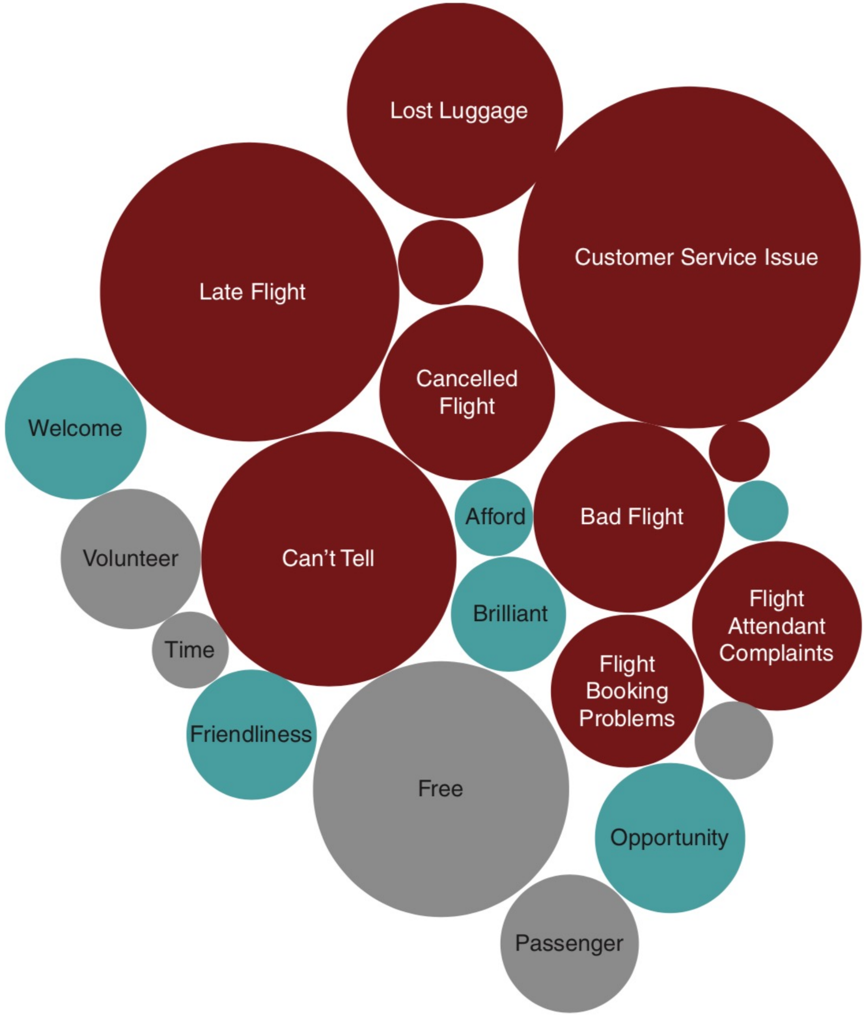


Figure 1.14 Bubble chart.

- Figure 1.14 is a bubble chart which shows most common words which appear in the passenger reviews.
- Size of the bubble indicates number of times the word has been used in reviews and colour hue indicates positive or negative sentiment.

Relationship

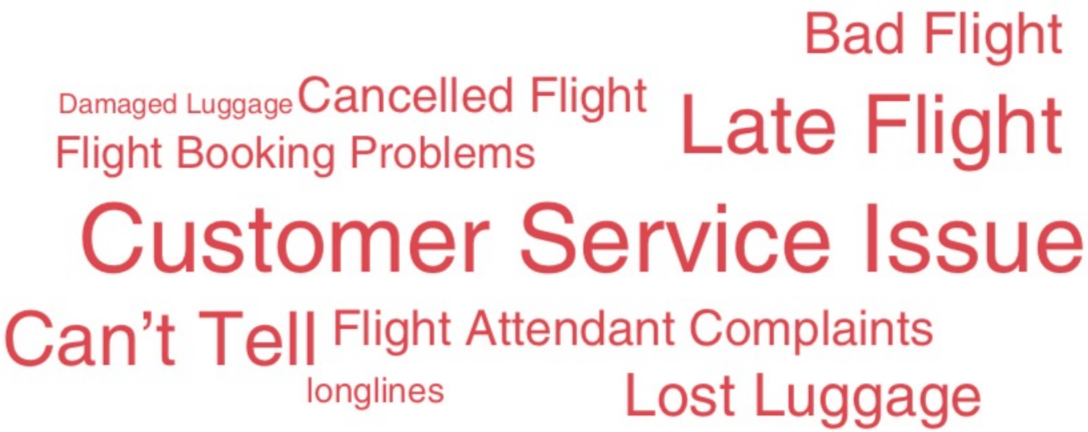


Figure 1.15 Word cloud.

- Another type of graph most commonly used when we analyse text data is word cloud (Fig. 1.15).
- It represents the relative frequency of the words to one another.
- The size of the word is directly proportional to the number of times the word is used in the text.

Chart Selection Guide

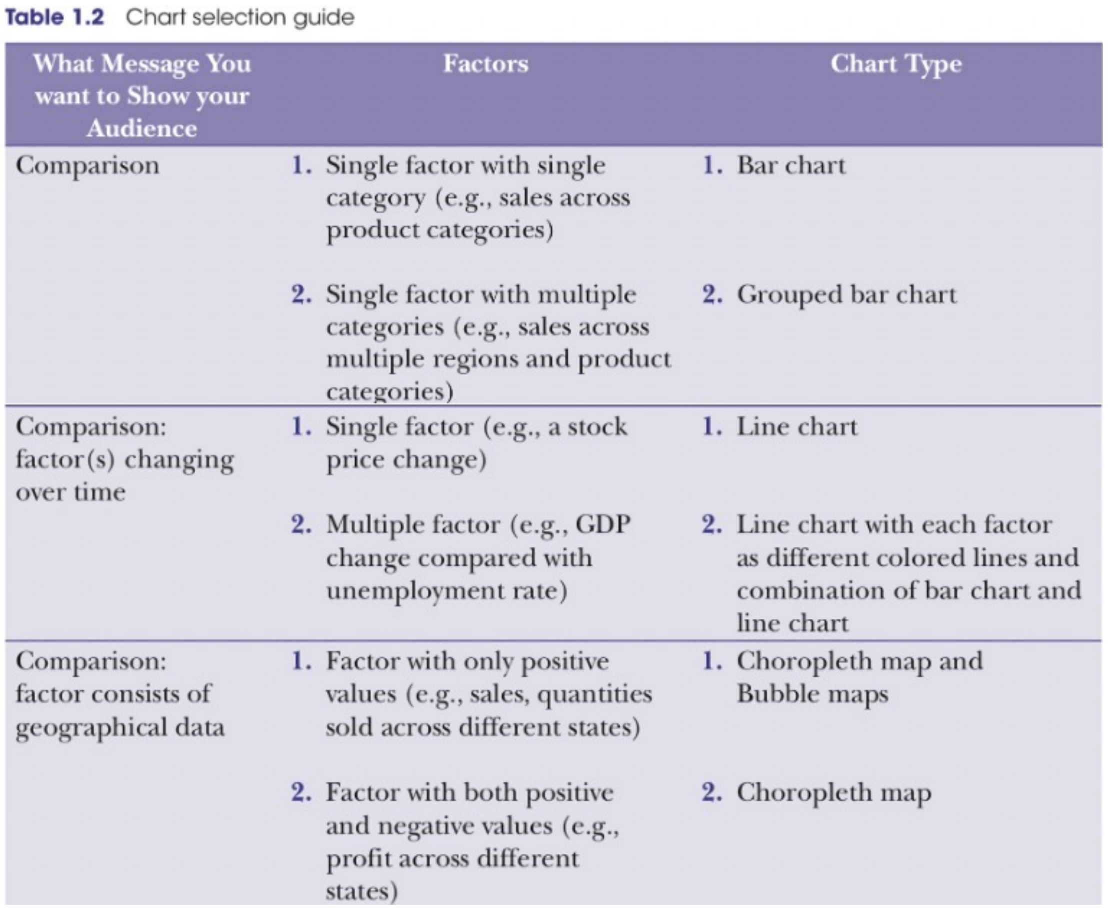
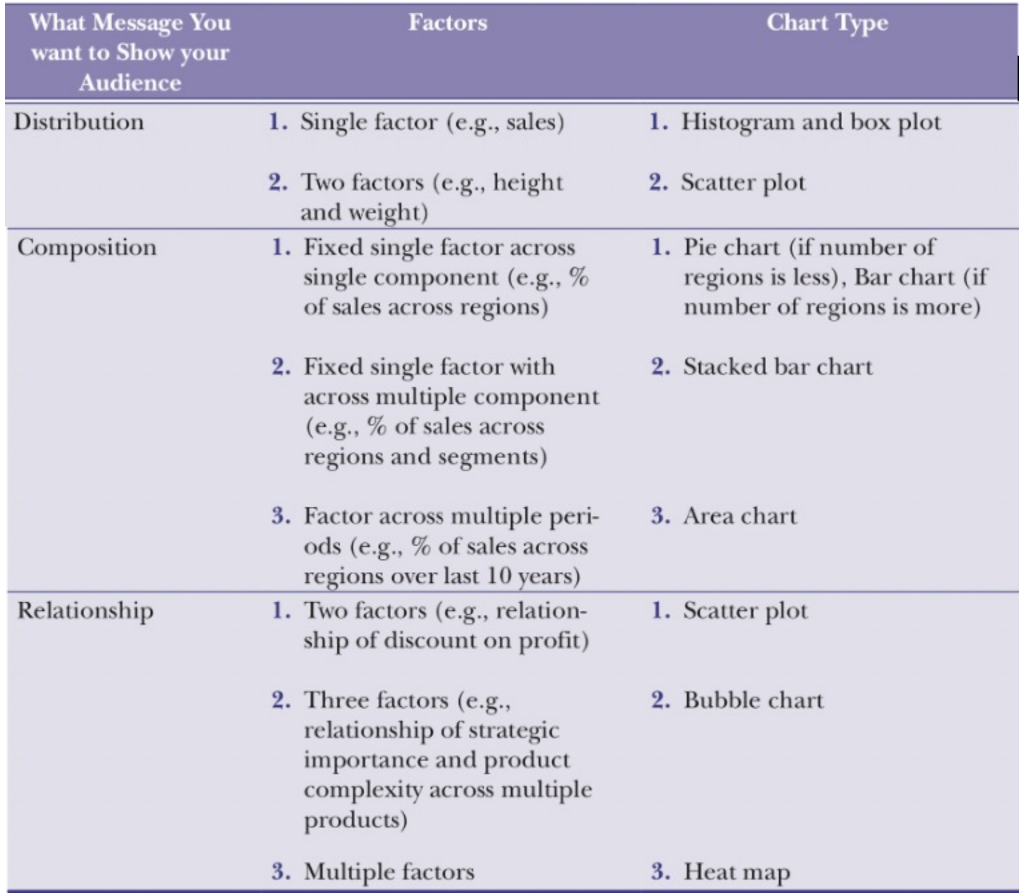


Table 1.2 Chart selection guide

What Message You want to Show your Audience	Factors	Chart Type
Comparison	1. Single factor with single category (e.g., sales across product categories)	1. Bar chart
	2. Single factor with multiple categories (e.g., sales across multiple regions and product categories)	2. Grouped bar chart
Comparison: factor(s) changing over time	1. Single factor (e.g., a stock price change)	1. Line chart
	2. Multiple factor (e.g., GDP change compared with unemployment rate)	2. Line chart with each factor as different colored lines and combination of bar chart and line chart
Comparison: factor consists of geographical data	1. Factor with only positive values (e.g., sales, quantities sold across different states)	1. Choropleth map and Bubble maps
	2. Factor with both positive and negative values (e.g., profit across different states)	2. Choropleth map



What Message You want to Show your Audience	Factors	Chart Type
Distribution	1. Single factor (e.g., sales)	1. Histogram and box plot
	2. Two factors (e.g., height and weight)	2. Scatter plot
Composition	1. Fixed single factor across single component (e.g., % of sales across regions)	1. Pie chart (if number of regions is less), Bar chart (if number of regions is more)
	2. Fixed single factor with across multiple component (e.g., % of sales across regions and segments)	2. Stacked bar chart
	3. Factor across multiple periods (e.g., % of sales across regions over last 10 years)	3. Area chart
Relationship	1. Two factors (e.g., relationship of discount on profit)	1. Scatter plot
	2. Three factors (e.g., relationship of strategic importance and product complexity across multiple products)	2. Bubble chart
	3. Multiple factors	3. Heat map

Common Chart Selection Questions

- How can we ensure that the visualization communicates the message correctly to our audience?
- Let us look at few common questions about how to select a chart that conveys the message in the simplest manner possible.
- We will be using the sample superstore dataset (superstoredata.xls), this is a data source which has been curated and made available by Tableau.
- This dataset contains information about customer orders along with details such as state, region, date of purchase, product category, sales, and profit.

Contrast Between Stacked Bar Chart and Grouped Bar Chart

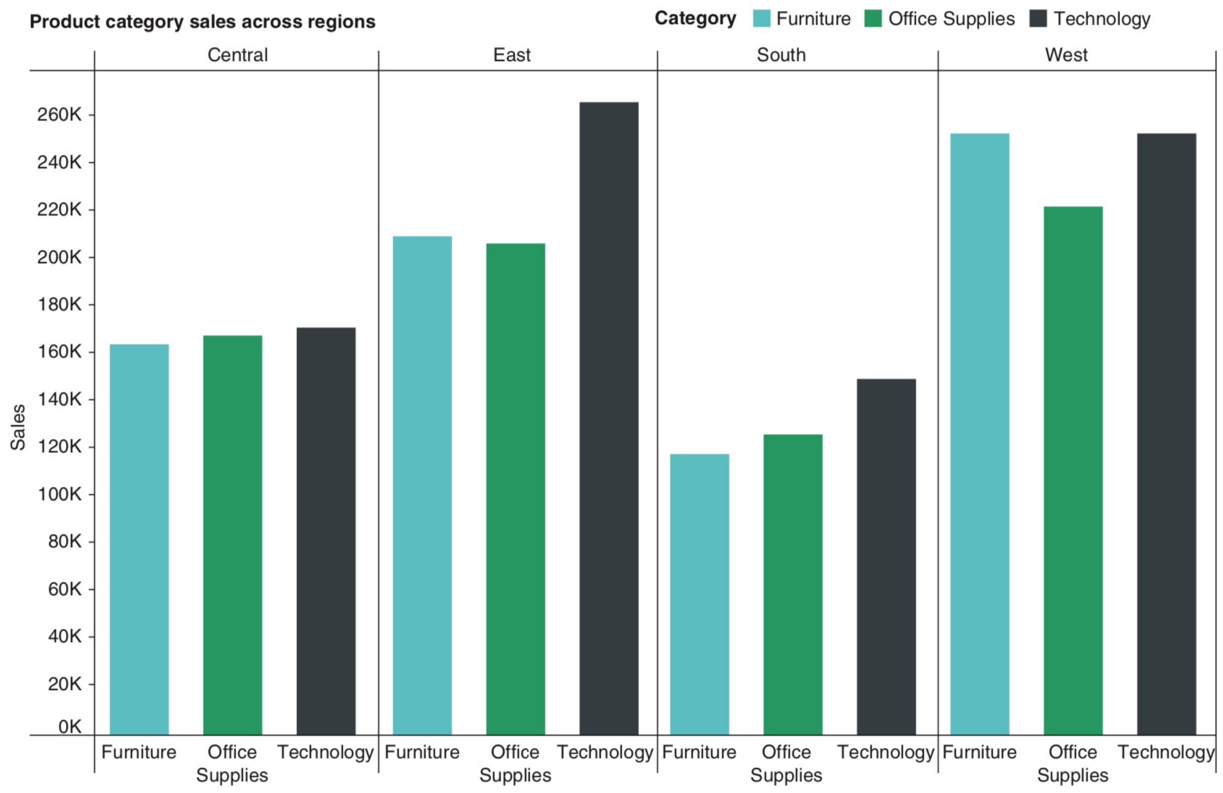


Figure 1.16 Grouped bar chart.

- We can use stacked bar charts and grouped bar charts when we have multiple groups of items to compare.
- But if we are interested in comparing composition of each element, then a stacked bar chart will enable this comparison (refer Fig. 1.17).
- As we can see, stacked bar chart will not help us compare and estimate differences between individual categories across regions.

Contrast Between Stacked Bar Chart and Grouped Bar Chart

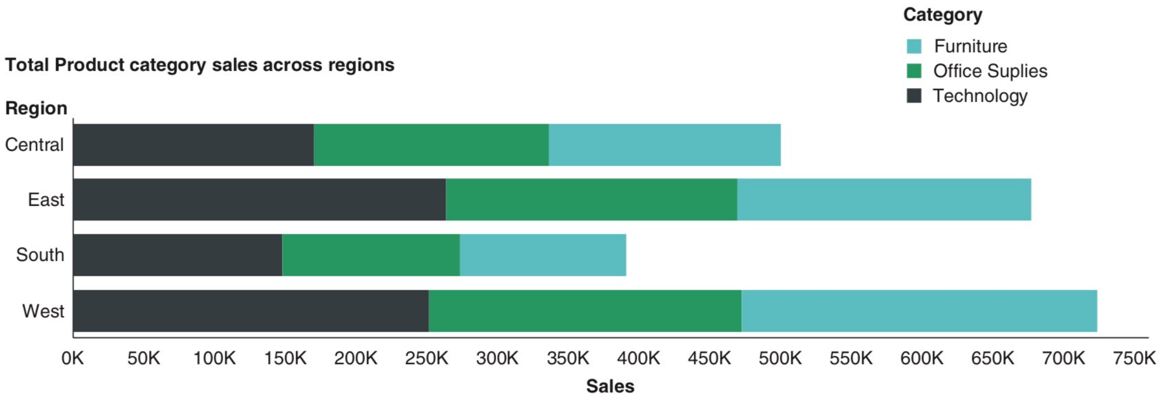


Figure 1.17 Stacked bar chart.

- But if we are interested in comparing composition of each element, then a stacked bar chart will enable this comparison (refer Fig. 1.17).
- As we can see, stacked bar chart will not help us compare and estimate differences between individual categories across regions.

Contrast Between Line Chart and Area Chart

- Line chart and Area chart both enable us to analyse how a quantitative variable has changed over time.
- So, can these charts be used interchangeably?
- Even though both charts enable similar analysis, they have different functional uses.
- As an example, we will analyse segment-wise monthly sales trends in our superstore dataset.

Contrast Between Line Chart and Area Chart

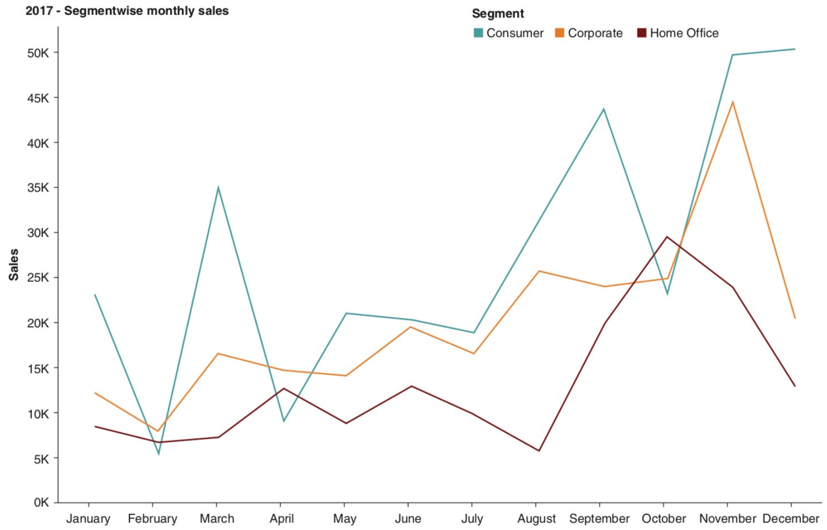


Figure 1.18 Line chart.

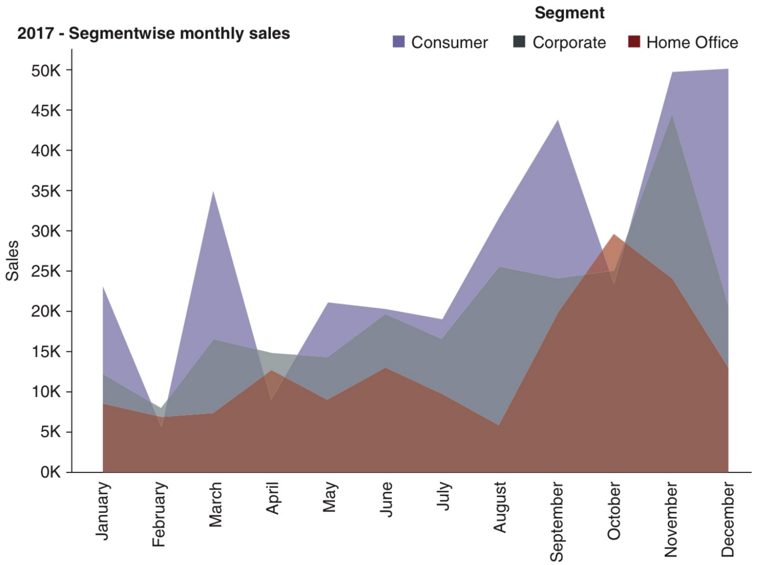


Figure 1.19 Area chart.

- Here, we want our user to compare trends across different segments then Line chart enables this function better than an Area chart.
- The overlapping areas of each segment make it difficult to compare even with set transparency.
- Thus, if our goal is to compare groups and analyse trends, we should use a line chart.

Contrast Between Line Chart and Area Chart

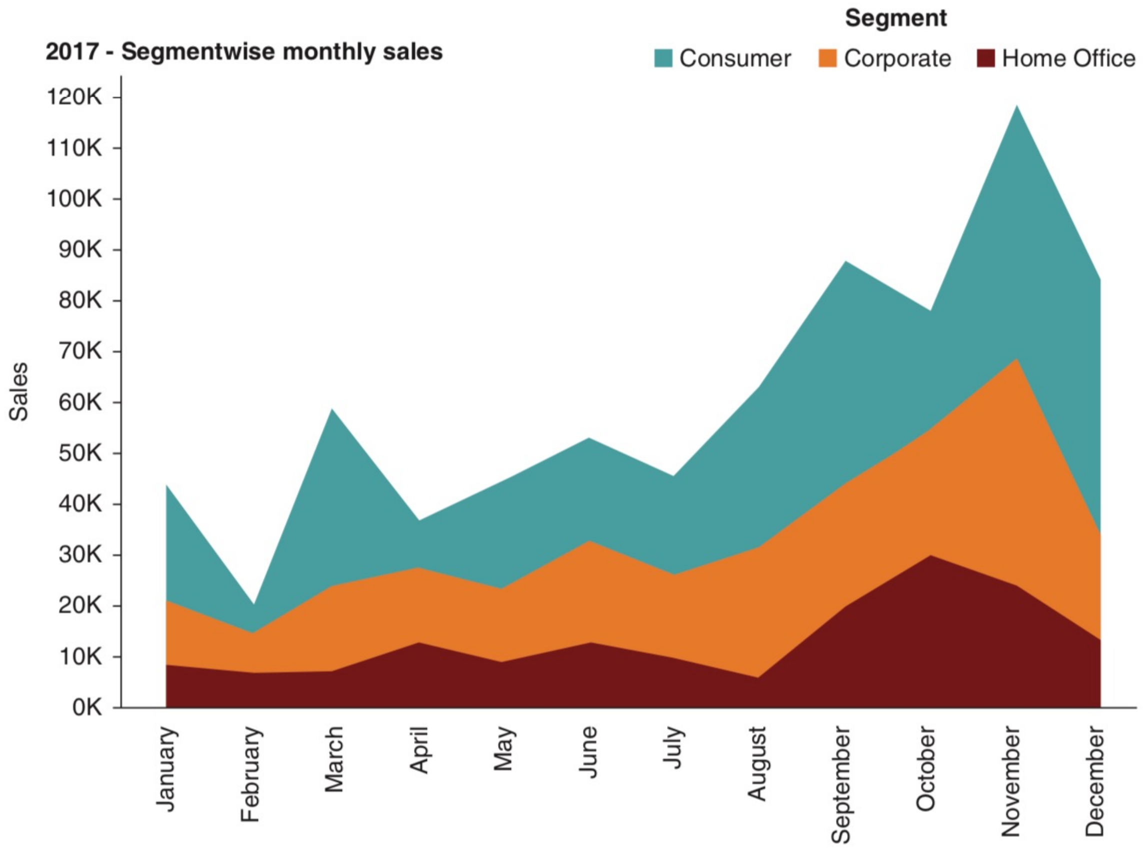


Figure 1.20 Stacked area chart.

- As a part of trend analysis, if we are looking to convey part of a whole relationship, then we can use a Stacked area chart

Hierarchical Data – Part of a Whole Analysis

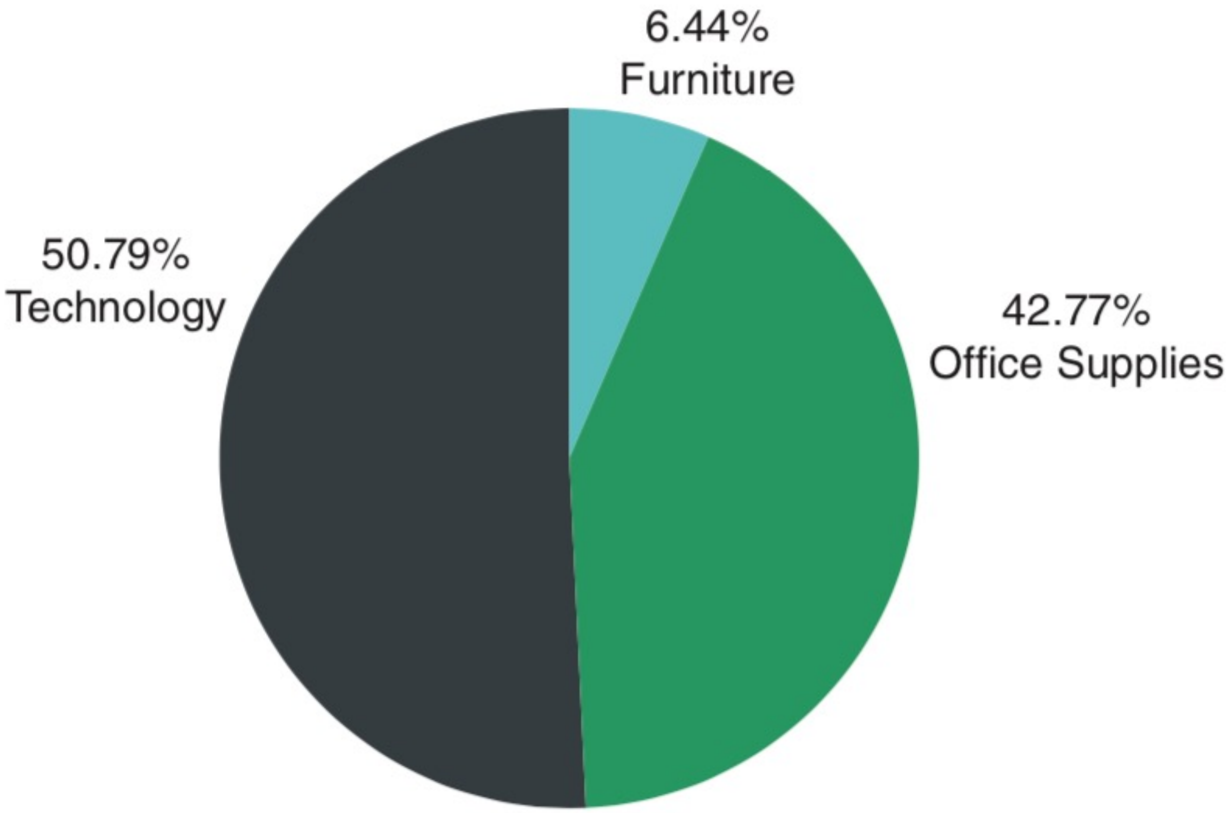


Figure 1.21 Pie chart.

- if we want to know what profit is contributed by each product category then we can use pie chart (Fig. 1.21).

Hierarchical Data – Part of a Whole Analysis

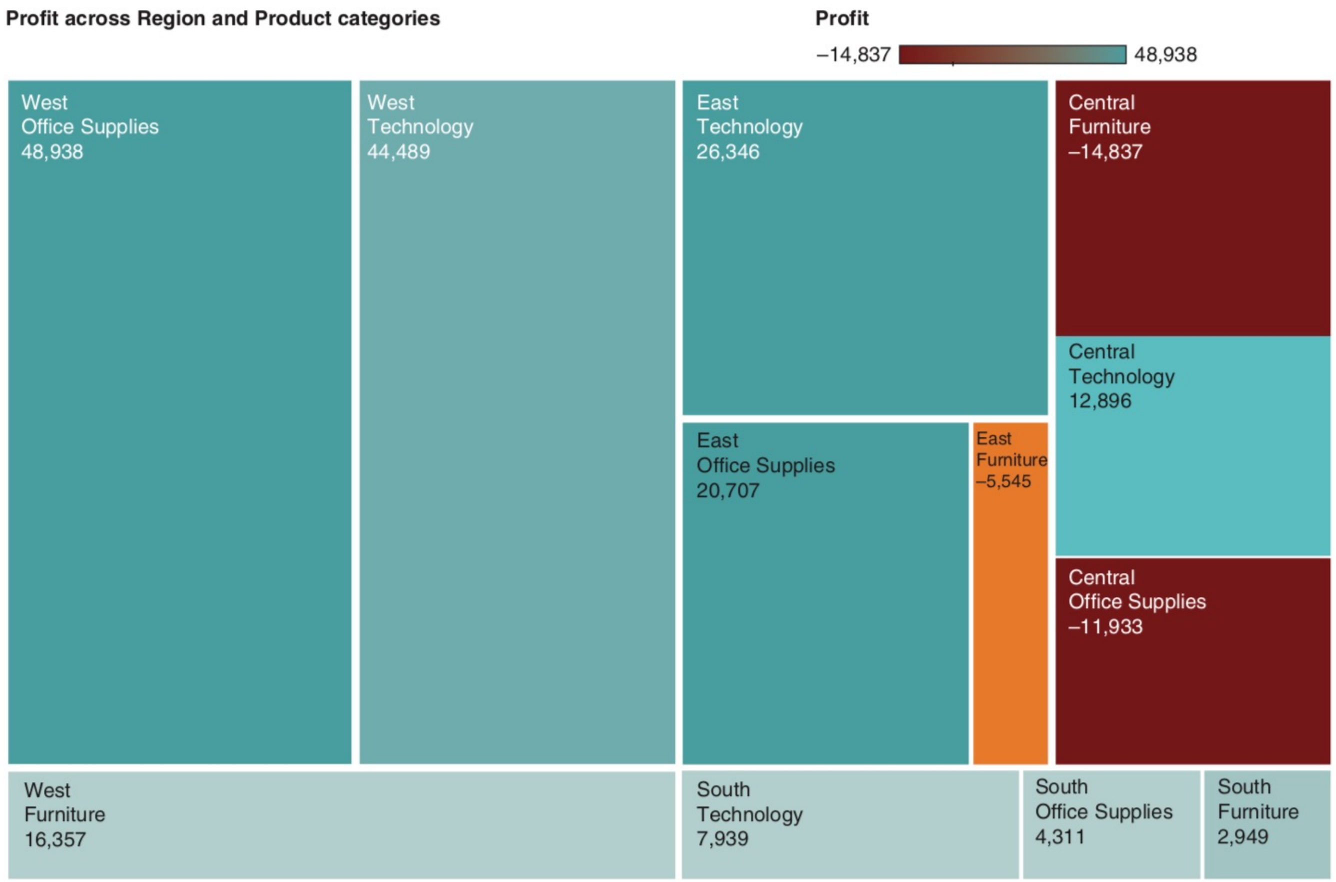


Figure 1.22 Tree map.

- If we want to show this in the context of each region, then we can use Treemap (Fig. 1.22).

Hierarchical Data – Part of a Whole Analysis

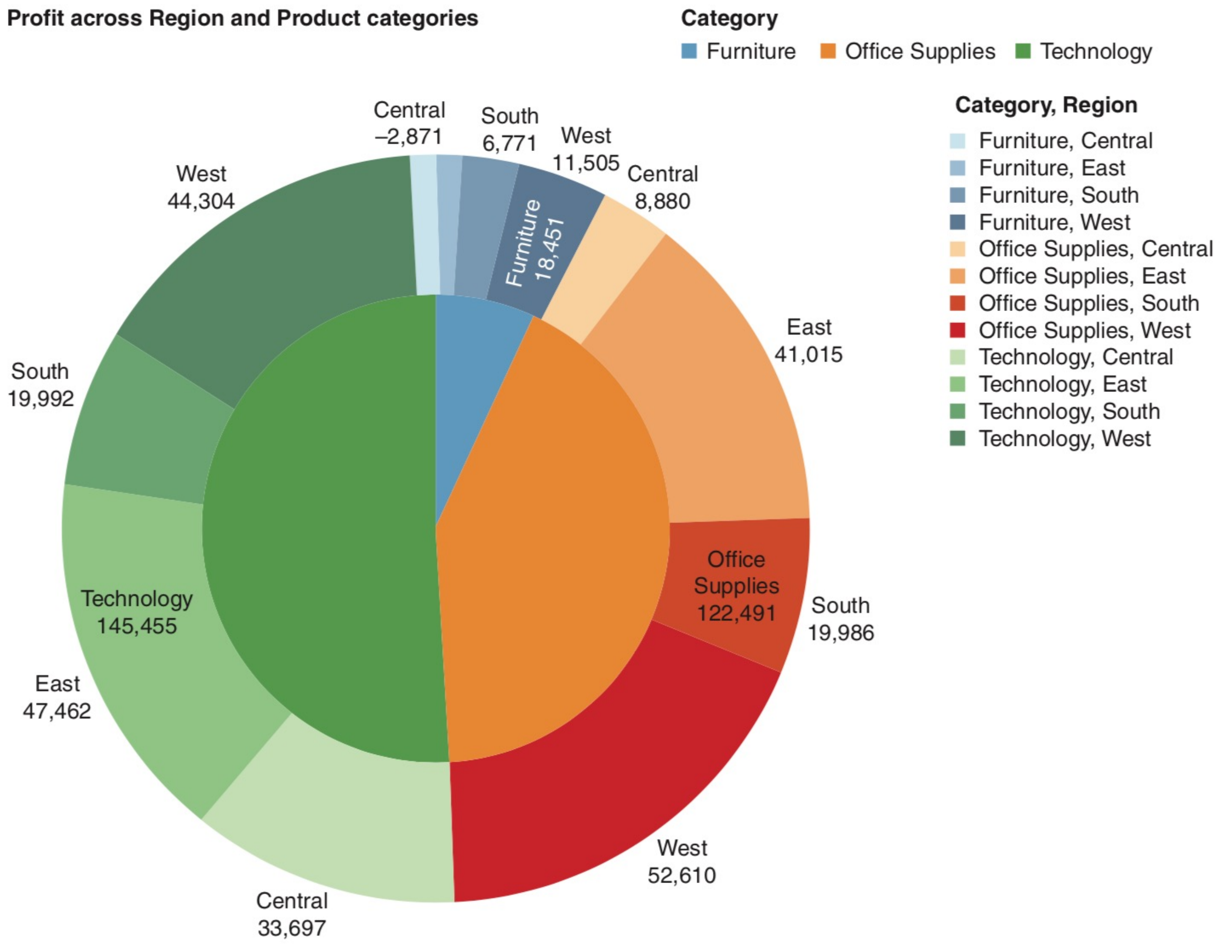


Figure 1.23 Sunburst chart.

- We can also use a Sunburst chart (Fig. 1.23) which is a radial view of Treemap.

Combination Chart

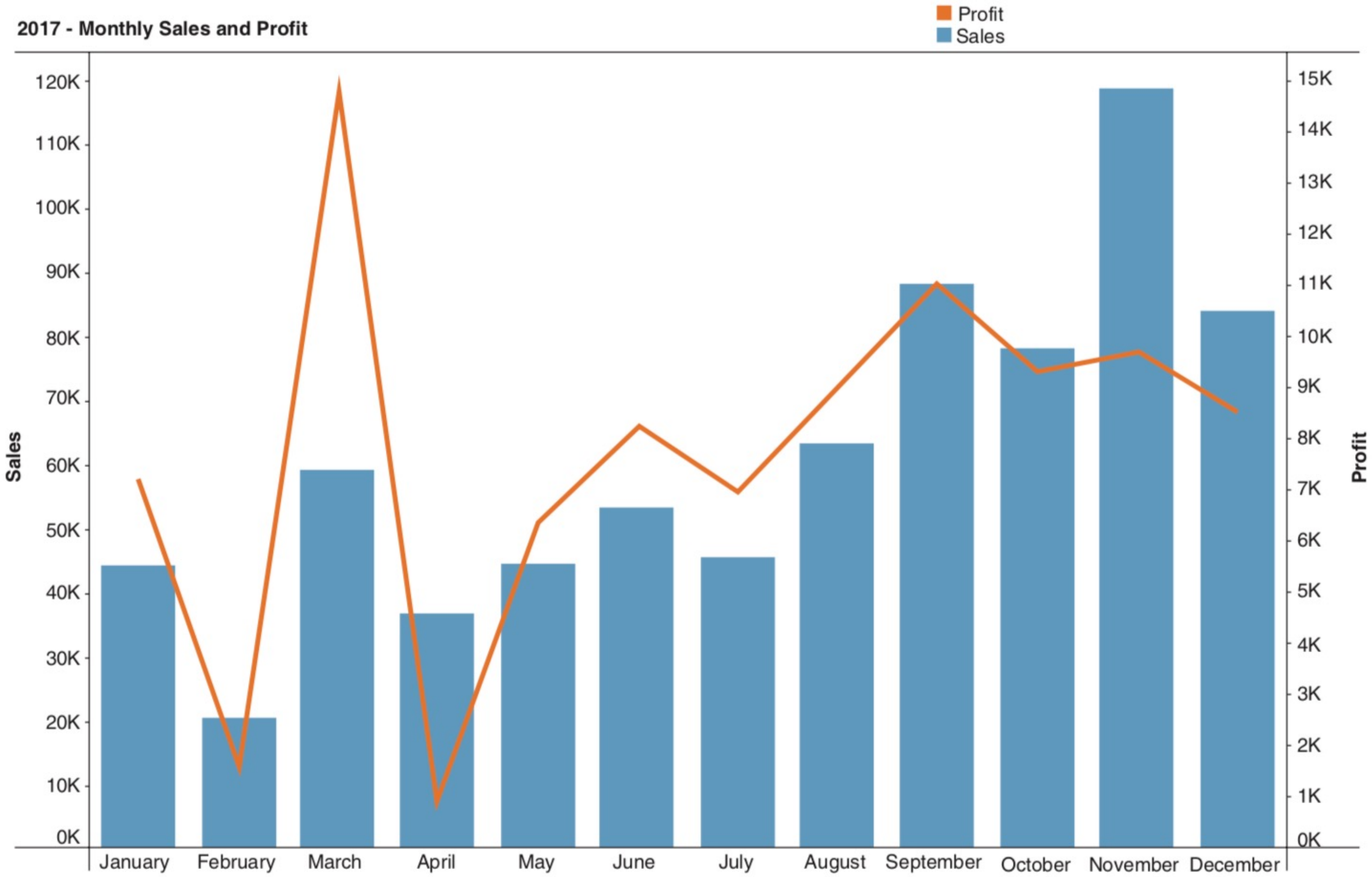


Figure 1.24 Combination chart.

- Consider an example where you would like to present sales data of superstore for each month along with profit.
- We need to compare month on month sales along with profit for each month.
- Now we have two data measures sales and profit with different scales.
- Here we can use combi- nation chart or dual axis chart to enable this analysis

Mistakes to avoid when designing data visualizations

- A chart with too much information will create clutter and discourage your audience from seeking insights.
- Avoid putting a lot of data points in one chart rather use multiple visualizations to convey the story.
- Always label the axes and provide chart headers. This informs our audience of the context of our analysis.
- Depending on the data, horizontal, and vertical axes should be scaled appropriately.
- Select the chart type wisely so that the audience can easily understand your analysis.

References

- [DeFanti, 1989] – T. A. DeFanti, M. D. Brown, and B. H. McCormick 1989, "Visualization: *Expanding Scientific and Engineering Research Opportunities*". *Computer* 22, 8, 12–25, [DOI: http://doi. org/10.1109/2.35195].
- [Galilei, 1613] – G. Galilei (1613), "Letters on sunspots (Istoria e Dimostrazioni intorno alle Macchie Solari)", British Library, Shelfmark: Egerton MS 48, last accessed May 25, 2021.
- [Felder,1988]–R.M.FelderandL.K.Silverman(1988),"Learning and Teaching Styles in Engineering Education", *Journal of Engineer- ing Education* 78(7): 674–681.
- [Romih, 2016] – T. Romih, (2016), "Humans are Visual Creatures", available at https://www.seyens.com/humans-are-visual-creatures/, last accessed May 27, 2021.
- [Tufte, 1983] – E. Tufte (1983), "The Visual Display of Quantitative Information, Edward Tufte, available at https://www.edwardtufte. com/tufte/books_vdqi.
- [Anscombe, 1973] – F. J. Anscombe (1973), "Graphs in Statistical Analysis". *American Statistician.* 27(1): 17–21. [Doi:10.1080/000313 05.1973.10478966. JSTOR 2682899].
- [Healy, 2014] – K. Healy and J. Moody (2014), "Data Visualization in Sociology", *Annual Review of Sociology,* 40, 105–128.
- [Solis, 2019] – J. Solis (2019), "Data Visualization is King", *The Journal of Private Equity,* 22(3), 102–107.
- [Knaffic, 2015] – C. N. Knaffic (2015), "Storytelling with Data: *A Data Visualization Guide for Business Professionals*", Wiley.

Thank You!