Principles and Practices of Data Science

Lecture 4

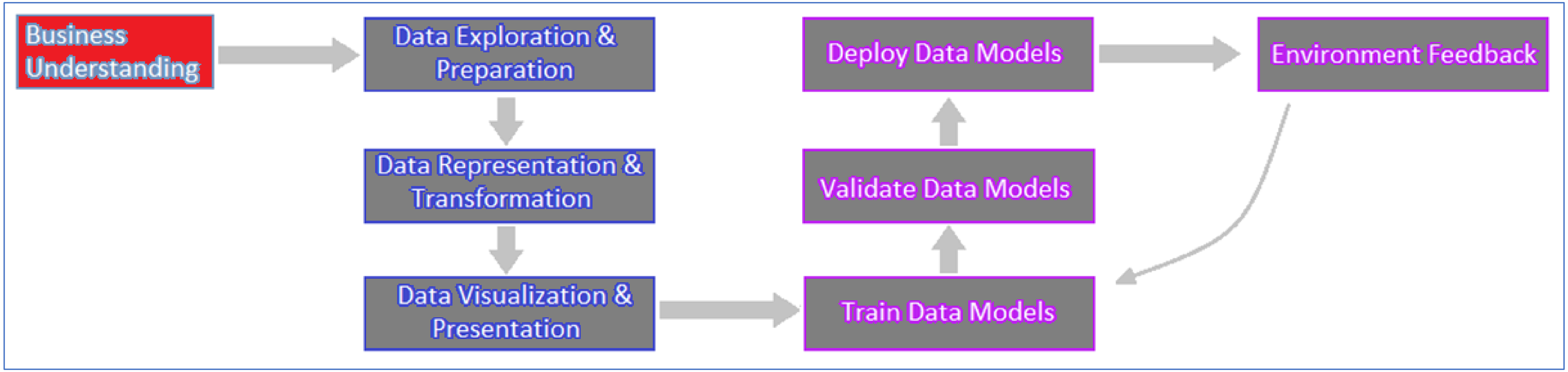
Melvin Ayala

Lecture 4: Data Collection, Requirements and Visualization

Sections

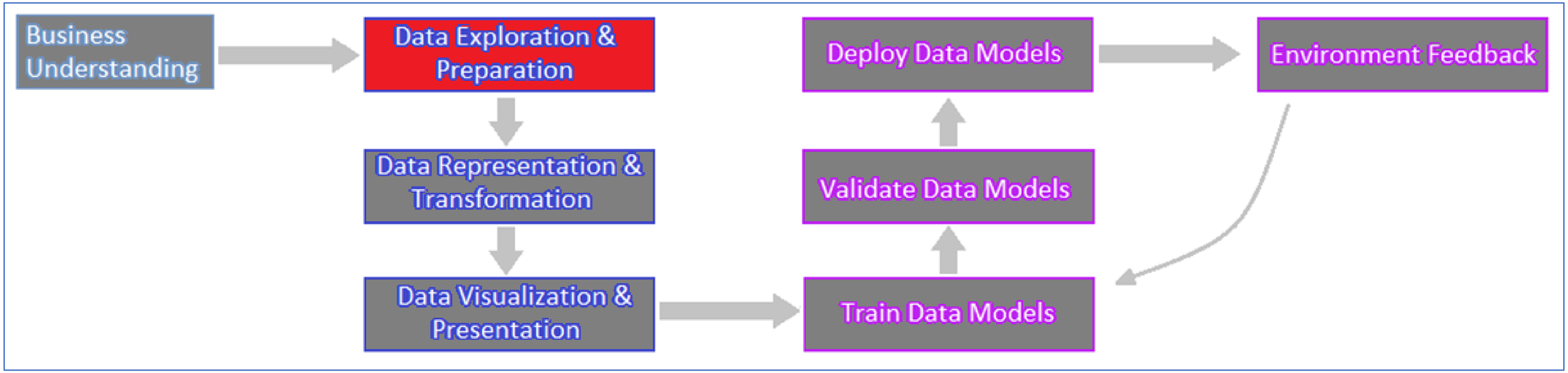
1. Business Understanding
2. Explore Data
   - 2.1. Data Exploration
   - 2.2. Accessing Data
   - 2.3. Ways Data Scientists Retrieve Data: Static File
   - 2.4. Ways Data Scientists Retrieve Data: Internet
   - 2.5. Ways Data Scientists Retrieve Data: Database
   - 2.6. Ways Data Scientists Retrieve Data: Unstructured Data
3. Prepare Data
   - 3.1. Process Details for Data Preparation
   - 3.2. Data Cleansing
   - 3.3. Missing Data
4. Understand Data
   - 4.1. About Data Types
   - 4.2. Finding Hidden Insights within Raw Data
5. Statistics and Representation Techniques
6. Data Transformation
7. Representing and Transforming Unstructured Data
8. Data Transformation Tools
9. Decision-Centered Visualization
10. Fundamentals of Visualization
    - 10.1. Perception
    - 10.2. Intervals and Ratios
    - 10.3. Manage Tricky Situations
    - 10.4. Comparable Scales
    - 10.5. Use Reference Points
    - 10.6. Create Interactions Between Multiple Visualizations
    - 10.7. Foster Iterative Data Interpretation
11. Common Graphs
    - 11.1. Introduction
    - 11.2. Bar, Line, and Pie Charts
    - 11.3. Maps
    - 11.4. Scatter Plots
    - 11.5. Bubble and Flow Charts
    - 11.6. Heat and Radar Maps

1. Business Understanding


Business Understanding
Data Exploration & Preparation
Data Representation & Transformation
Data Visualization & Presentation
Train Data Models
Validate Data Models
Deploy Data Models
Environment Feedback

2. Explore Data

2.1. Data Exploration



Business Understanding
Data Exploration & Preparation
Data Representation & Transformation
Data Visualization & Presentation
Train Data Models
Validate Data Models
Deploy Data Models
Environment Feedback

Data Requirements:

The chosen analytic approach determines the data requirements.

Data Collection:

Data Scientists identify, gather and curate the available data resources relevant to the problem domain.

2.2. Accessing Data

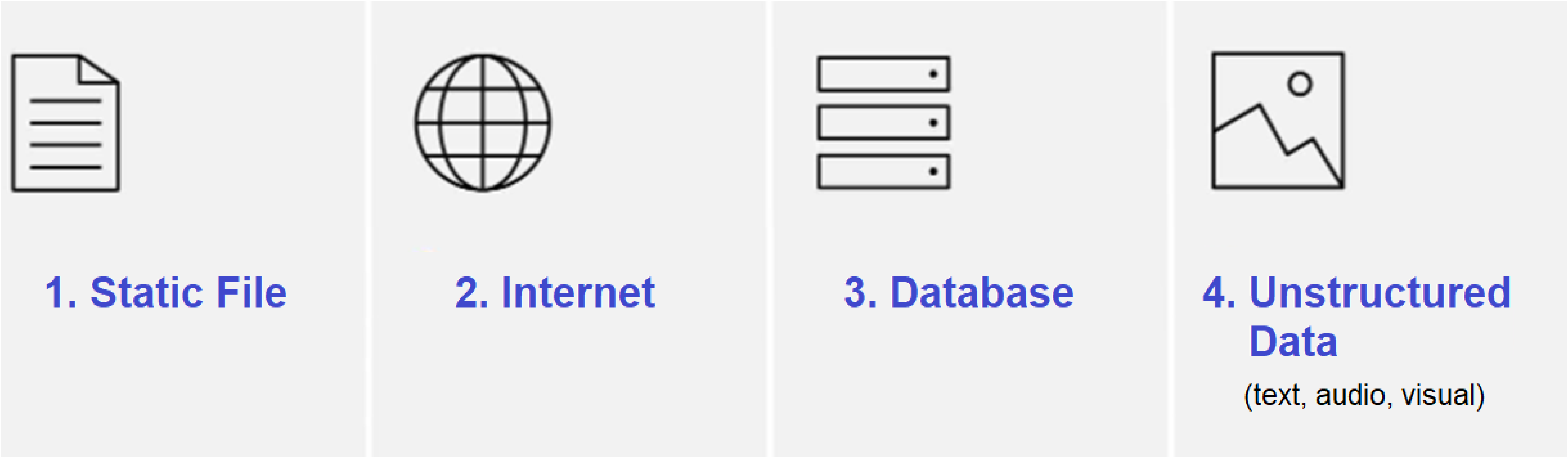
Your data might come from:

- your line of business applications
- data warehouses
- external sources

Access data from:

- static data (your local files)
- communities (internet)
- database repositories

There are several ways data scientists retrieve data



1. Static File

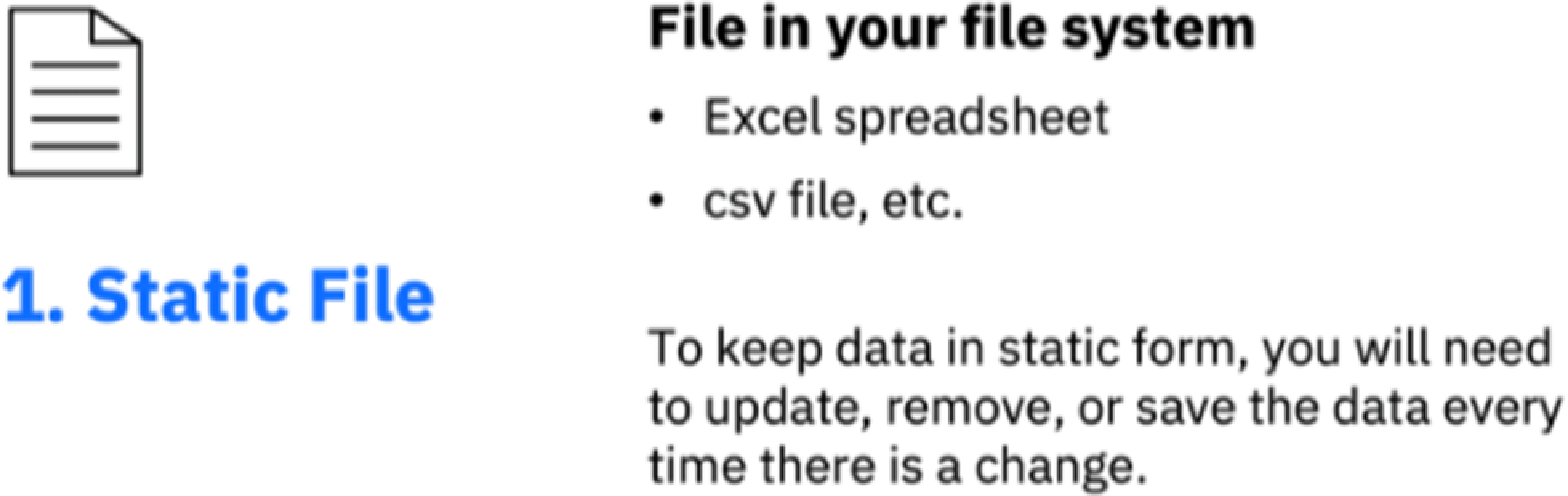
2. Internet

3. Database

4. Unstructured Data

(text, audio, visual)

Ways Data Scientists Retrieve Data: Static File



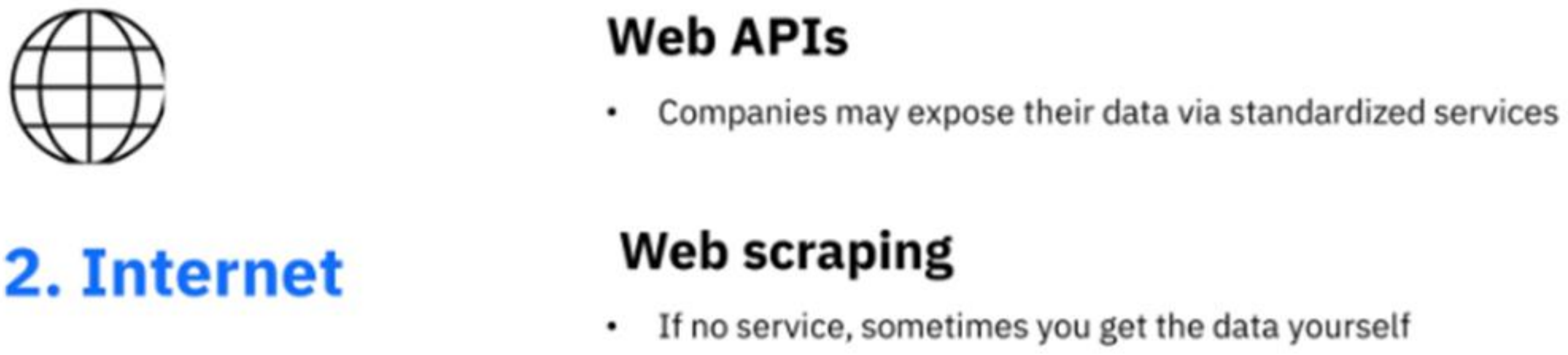
1. Static File

File in your file system

- Excel spreadsheet
- csv file, etc.

To keep data in static form, you will need to update, remove, or save the data every time there is a change.

Ways Data Scientists Retrieve Data: Internet



2. Internet

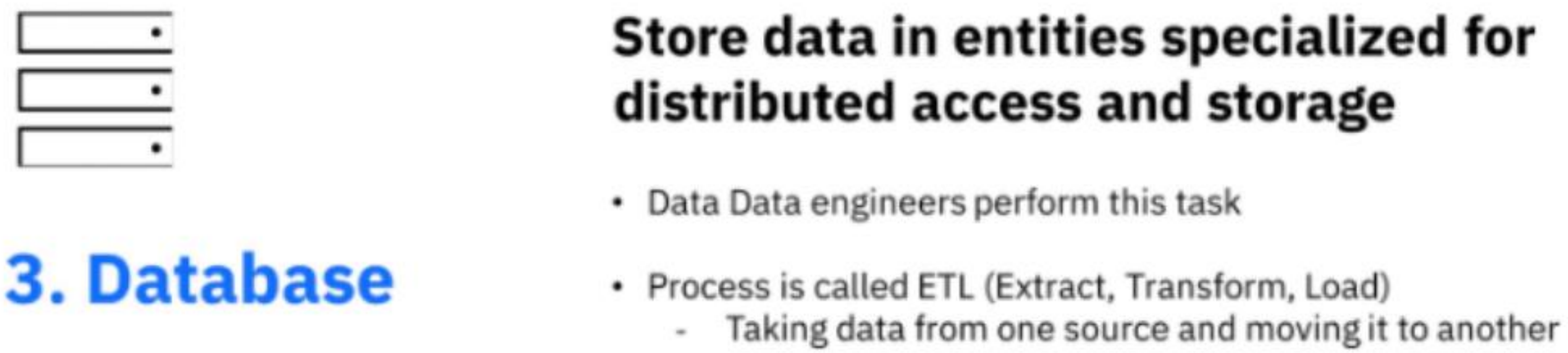
Web APIs

- Companies may expose their data via standardized services

Web scraping

- If no service, sometimes you get the data yourself

Ways Data Scientists Retrieve Data: Database



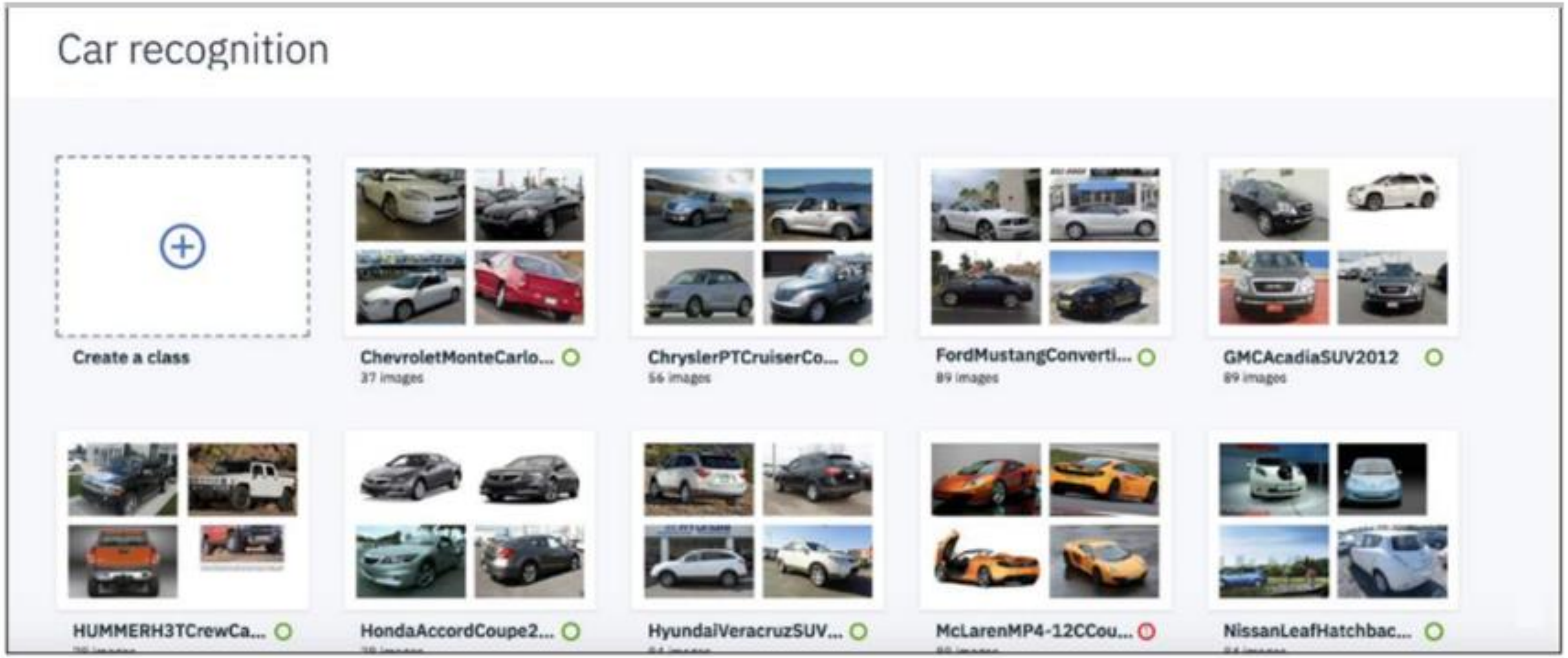
3. Database

Store data in entities specialized for distributed access and storage

- Data Data engineers perform this task
- Process is called ETL (Extract, Transform, Load)
  - Taking data from one source and moving it to another

Ways Data Scientists Retrieve Data: **Unstructured Data**

4. Unstructured Data



Car recognition

Create a class

ChevroletMonteCarlo... ○
37 images

ChryslerPTCruiserCo... ○
56 images

FordMustangConverti... ○
89 images

GMCAcadiaSUV2012 ○
89 images

HUMMERH3TCrewCa... ○

HondaAccordCoupe2... ○

HyundaiVeracruzSUV... ○

McLarenMP4-12CCou... ○

NissanLeafHatchbac... ○

3. Prepare Data

3.1. Process Details for Data Preparation

Unfortunately, we cannot assume that the data (even structured data) is ready to use.

We might have:

- incomplete data
- corrupted data
- un-friendly formats
- “noise” in the data
- irrelevant data
- extra work for cleaning up your data is referred to as “Data Wrangling”.

3.2. Data Cleansing

Data **Analysts** can spend up to 80% of the time cleaning data.

Common Tasks:

- Importing data
- joining multiple datasets
- detecting missing values
- detecting anomalies
- imputing for missing values
- data quality assurance

Concept of Tidy Data

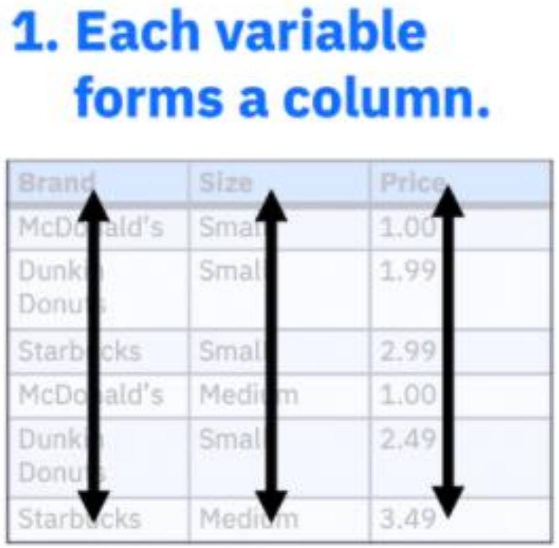
Tidy data lends itself to efficient data analysis and processing.

Key points:

- Transforming your data into standard format, or tidy data, makes analysis and storage easier down the road.
- Additionally, one must make sure the data are in its appropriate types.

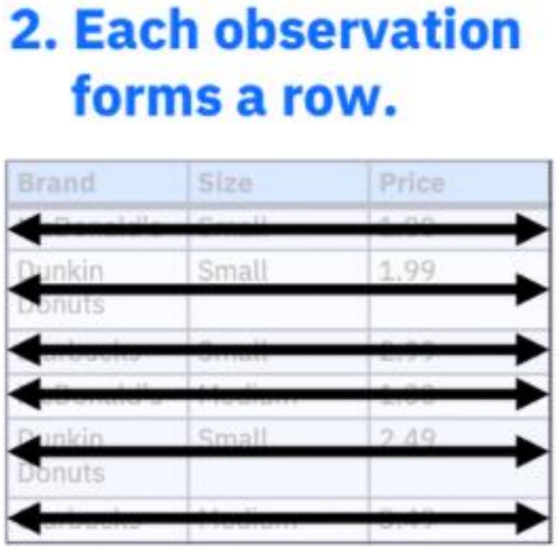
What is Tidy Data?

Tidy data satisfies 3 components:



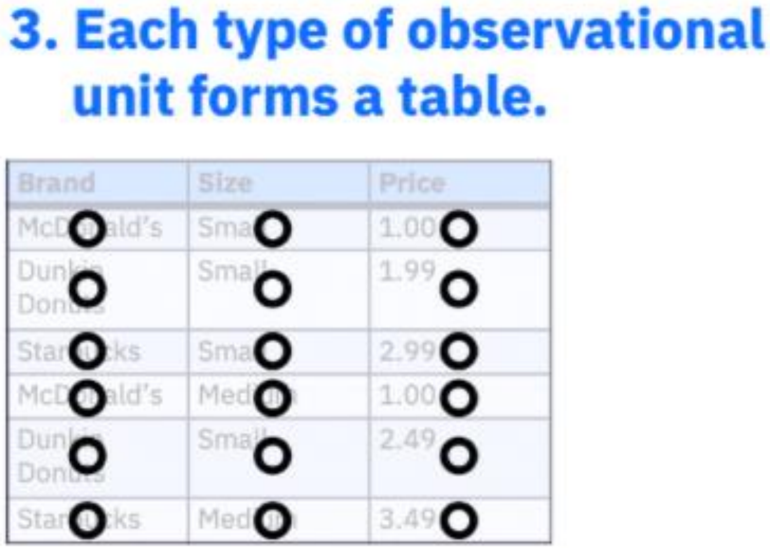
1. Each variable forms a column.

Brand	Size	Price
McDonald's	Small	1.00
Dunkin Donuts	Small	1.99
Starbucks	Small	2.99
McDonald's	Medium	1.00
Dunkin Donuts	Small	2.49
Starbucks	Medium	3.49



2. Each observation forms a row.

Brand	Size	Price
McDonald's	Small	1.00
Dunkin Donuts	Small	1.99
Starbucks	Small	2.99
McDonald's	Medium	1.00
Dunkin Donuts	Small	2.49
Starbucks	Medium	3.49



3. Each type of observational unit forms a table.

Brand	Size	Price
McDonald's	Small	1.00
Dunkin Donuts	Small	1.99
Starbucks	Small	2.99
McDonald's	Medium	1.00
Dunkin Donuts	Small	2.49
Starbucks	Medium	3.49

3.3. Missing Data

Explicit

- marked with Null or NA
- use summary functions to list each variable's count of missing values

Implicit

- not there at all
- not recognizable
- you must explore and visualize your data to notice things that appear "off"

Methods for Handling Missing Data:

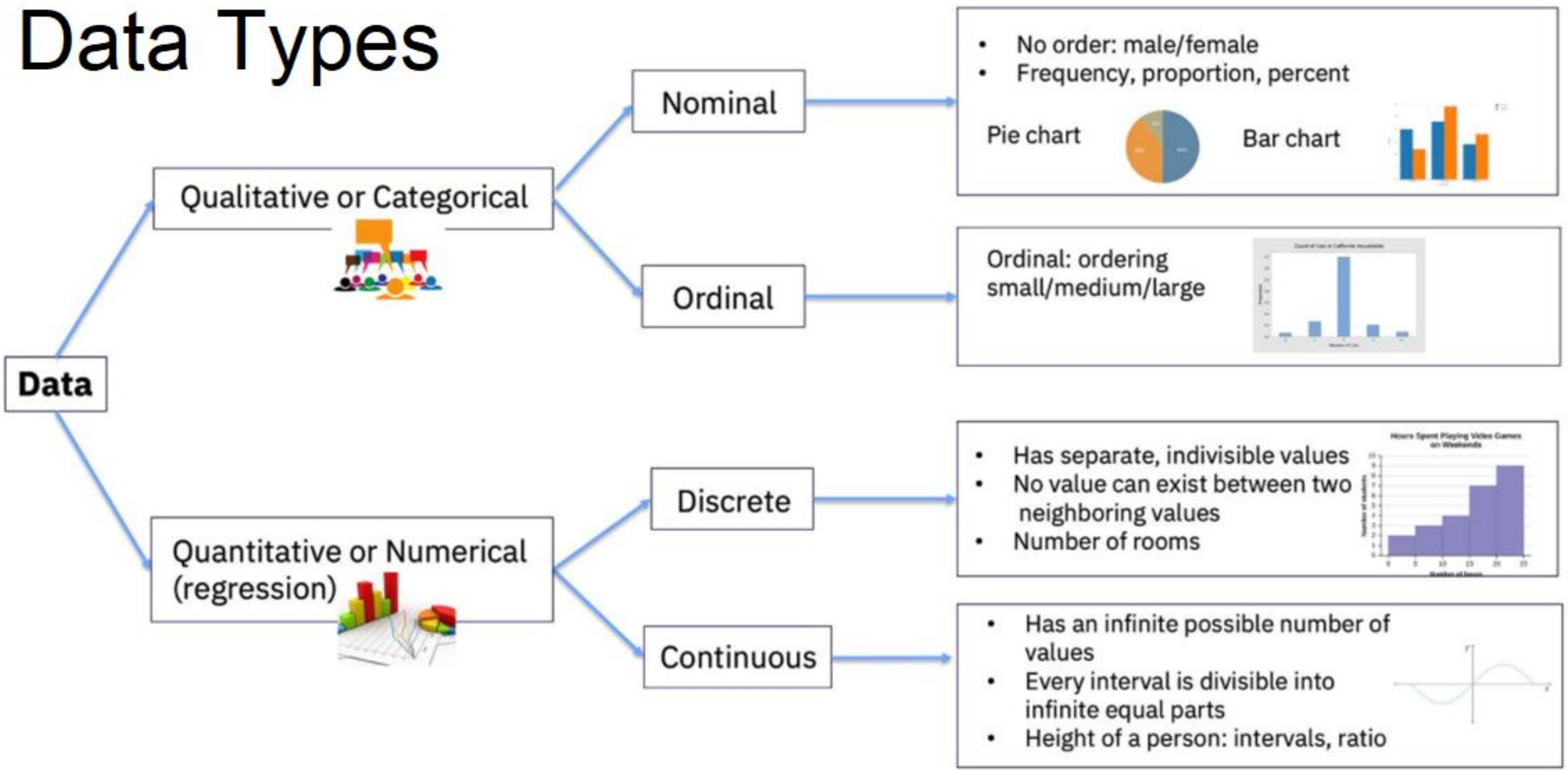
- remove the observation completely
- **impute the observation** (assign another value to it)

Imputation Methods:

- replace with summary statistics such as mean, median, or mode.
- create a new variable that flags a missing column.
- replace NA with an outlier (models will understand that these outliers are associated with missing values)

4. Understand Data

4.1. About Data Types



Data Types

Data

Qualitative or Categorical

Nominal

- No order: male/female
- Frequency, proportion, percent

Pie chart

Bar chart

Ordinal

Ordinal: ordering small/medium/large

Quantitative or Numerical (regression)

Discrete

- Has separate, indivisible values
- No value can exist between two neighboring values
- Number of rooms

Continuous

- Has an infinite possible number of values
- Every interval is divisible into infinite equal parts
- Height of a person: intervals, ratio

4.2. Finding Hidden Insights within Raw Data

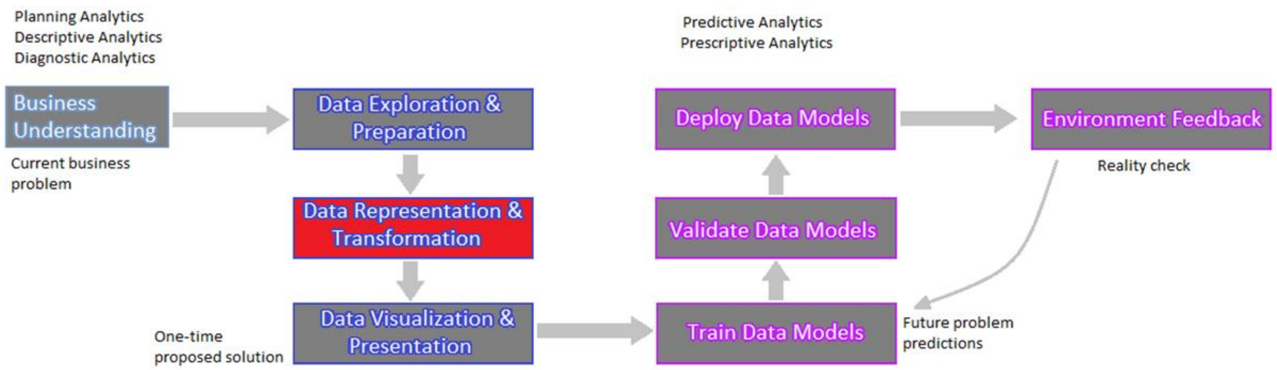
How to find hidden patterns in the data?

The following operations are common:

- Handling messy, inconsistent, o missing data
- Trying to combine data from multiple sources
- Reporting on data that was entered manually
- Performing summary statistics (can include mean, median, mode, extreme values, range, standard deviation)

5. Statistics and Representation Techniques

Data Science Method:



Data Representation and Transformation phase involves the following four steps:

- Statistical Analysis
- Exploratory Visualization
- Data Formatting
- Algorithm Alignment

Data Analysts typically use: **descriptive statistics** and **data visualization techniques** to:

- Understand the data content
- Assess data quality
- Discover initial insights about the data

Additional data collection may be necessary to fill gaps.

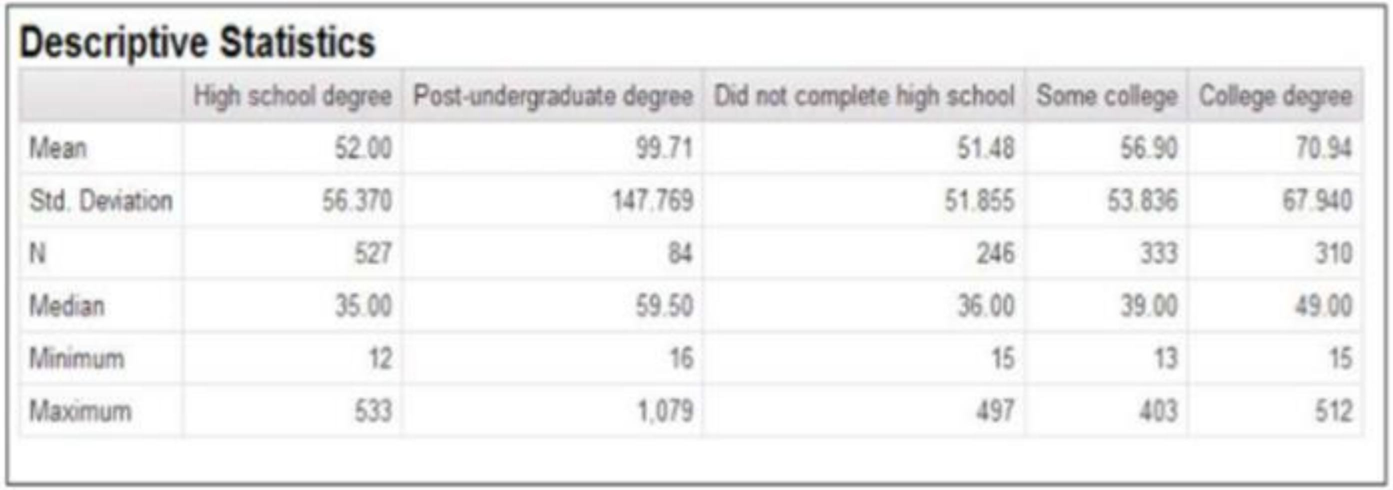
Statistical Analysis

The data representation phase should use mathematical tools such as:

- **statistics**
- **correlations**
- **chi-square tests**

Descriptive Statistics allow you to describe a vast, complex data set using just a few key numbers.

- You can also create a table that displays summarized statistics for cases grouped by categorical data based on a single measure.
- The table below shows the mean household income for customers grouped by education level.



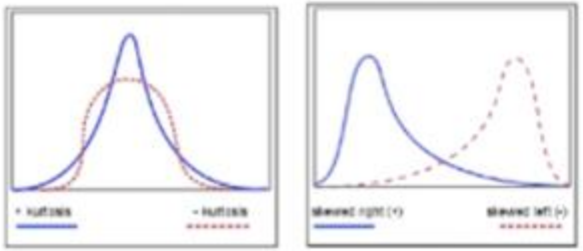
Descriptive Statistics

	High school degree	Post-undergraduate degree	Did not complete high school	Some college	College degree
Mean	52.00	99.71	51.48	56.90	70.94
Std. Deviation	56.370	147.769	51.855	53.836	67.940
N	527	84	246	333	310
Median	35.00	59.50	36.00	39.00	49.00
Minimum	12	16	15	13	15
Maximum	533	1,079	497	403	512

Descriptive Statistics Quantitatively Summarize a Data Set

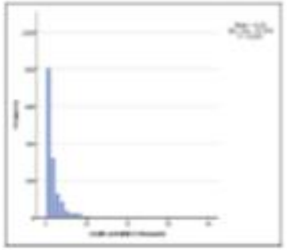
You can use **descriptive statistics** to:

- Look at averages, such as **mean** or **median**
- Obtain information, such as the **mean for groups of interest**, that you might need to interpret other statistical tests
- Provide graphical representations of data, such as **histograms** and **boxplots**

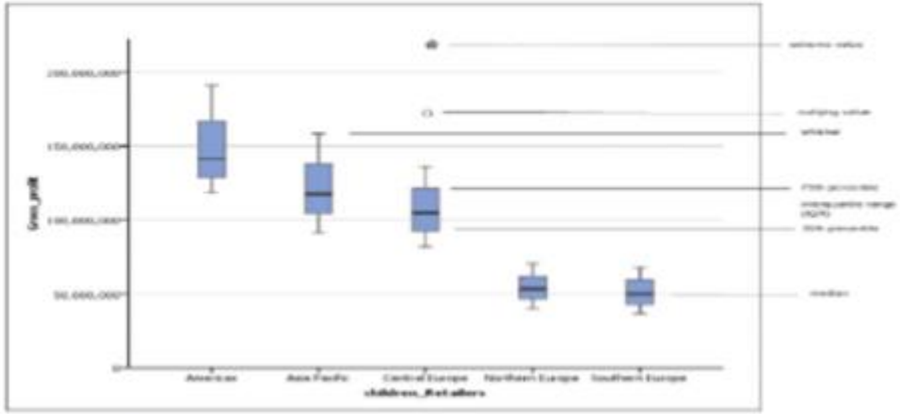


Descriptive Tables

- Measures of Central Tendency
- Measures of Dispersion
- Measures of Distribution



Histograms



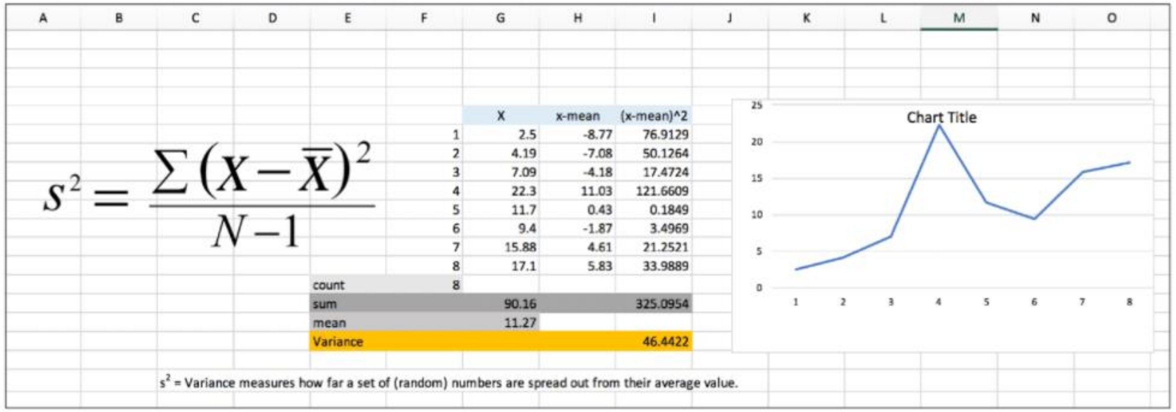
Boxplots

Variance vs. Standard Deviation

Variance measures the average degree to which each point differs from the mean.

The greater the variance, the larger the overall data range. It is a good way to spot the outliers and gives you an idea of the overall spread.

Standard Deviation is the square root of the variance.
The calculation of the variance uses squares because it weights outliers more heavily than data very near the mean.



$$s^2 = \frac{\sum(X-\overline{X})^2}{N-1}$$

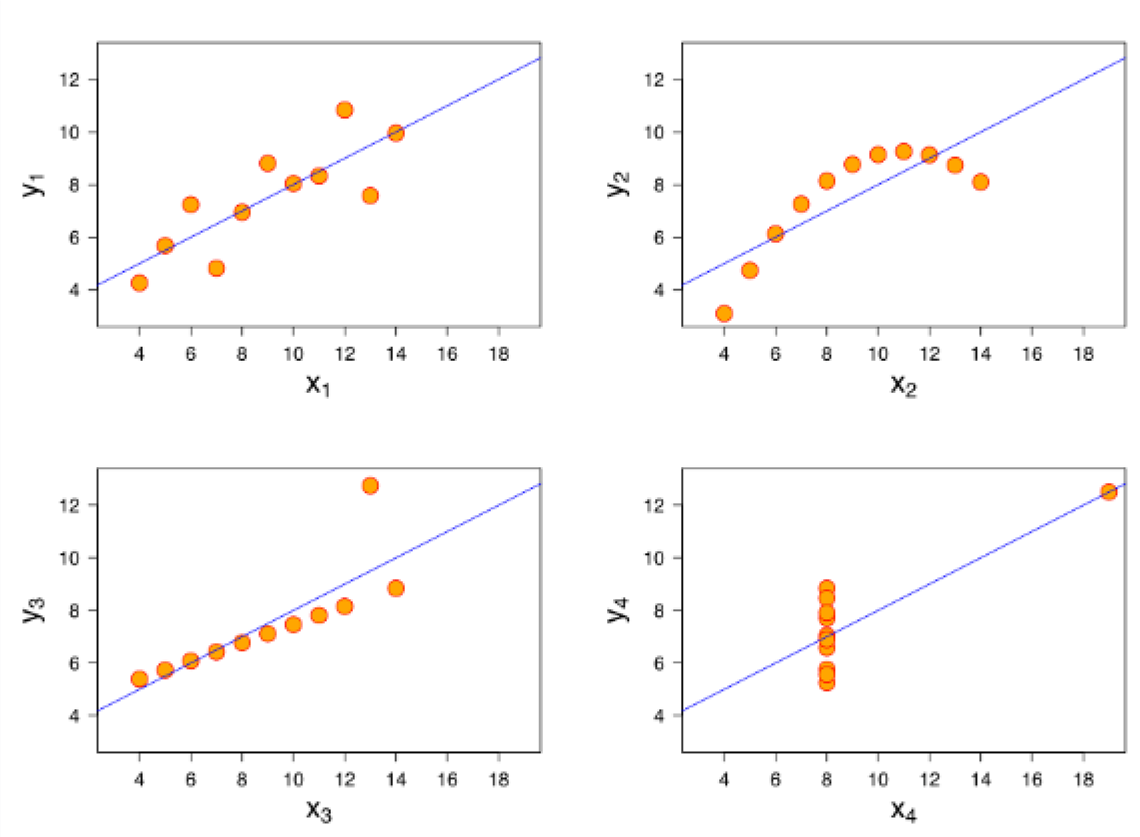
		X	x-mean	(x-mean)^2
	1	2.5	-8.77	76.9129
	2	4.19	-7.08	50.1264
	3	7.09	-4.18	17.4724
	4	22.3	11.03	121.6609
	5	11.7	0.43	0.1849
	6	9.4	-1.87	3.4969
	7	15.88	4.61	21.2521
	8	17.1	5.83	33.9889
count	8			
sum		90.16		325.0954
mean		11.27		
Variance				46.4422

s^2 = Variance measures how far a set of (random) numbers are spread out from their average value.

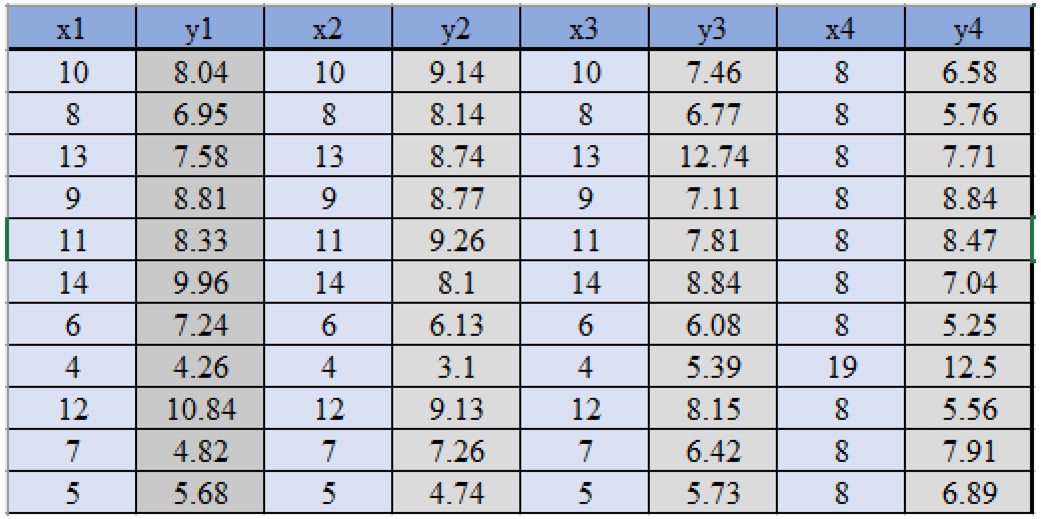
Risks of Using Descriptive Statistics

There is a danger in relying only on descriptive statistics and ignoring the overall distribution.

Anscombe’s quartet offers a classic example of this risk.



- They have very different distributions and appear differently when plotted on scatter plots.
- nearly identical in simple descriptive statistics
- Can fools the regression model



x1	y1	x2	y2	x3	y3	x4	y4
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

Four Data-sets

Average Value of x = 9

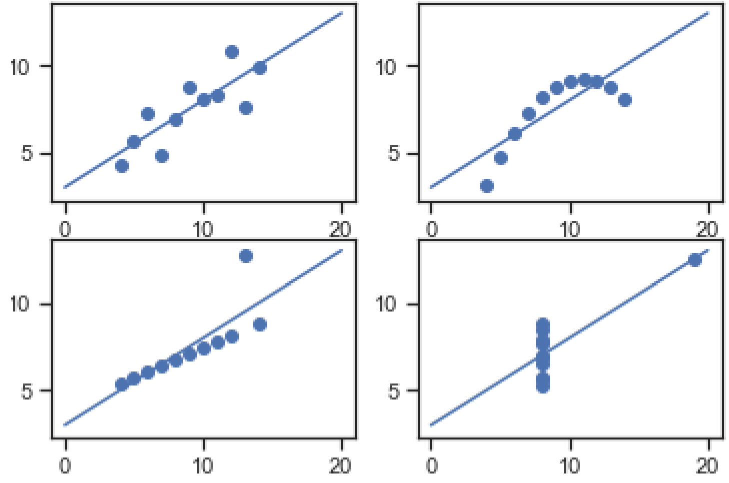
Average Value of y = 7.50

Variance of x = 11

Variance of y = 4.12

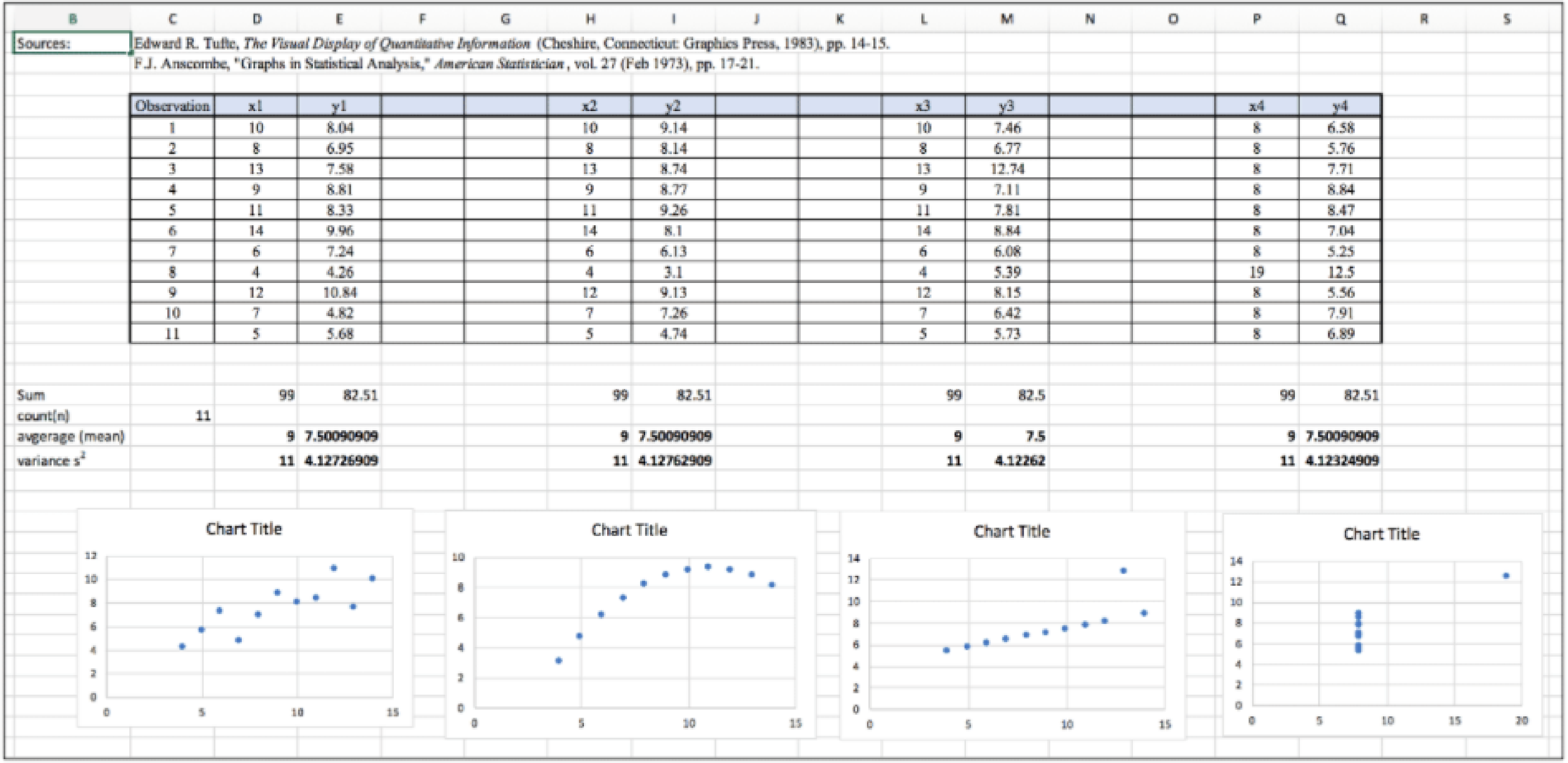
Correlation Coefficient = 0.816

Linear Regression Equation : y = 0.5 x + 3



Graphical Representation of Anscombe's Quartet

Anscombe’s Quartet



Sources: Edward R. Tufte, *The Visual Display of Quantitative Information* (Cheshire, Connecticut: Graphics Press, 1983), pp. 14-15.
F.J. Anscombe, "Graphs in Statistical Analysis," *American Statistician*, vol. 27 (Feb 1973), pp. 17-21.

Observation	x1	y1			x2	y2			x3	y3			x4	y4
1	10	8.04			10	9.14			10	7.46			8	6.58
2	8	6.95			8	8.14			8	6.77			8	5.76
3	13	7.58			13	8.74			13	12.74			8	7.71
4	9	8.81			9	8.77			9	7.11			8	8.84
5	11	8.33			11	9.26			11	7.81			8	8.47
6	14	9.96			14	8.1			14	8.84			8	7.04
7	6	7.24			6	6.13			6	6.08			8	5.25
8	4	4.26			4	3.1			4	5.39			19	12.5
9	12	10.84			12	9.13			12	8.15			8	5.56
10	7	4.82			7	7.26			7	6.42			8	7.91
11	5	5.68			5	4.74			5	5.73			8	6.89

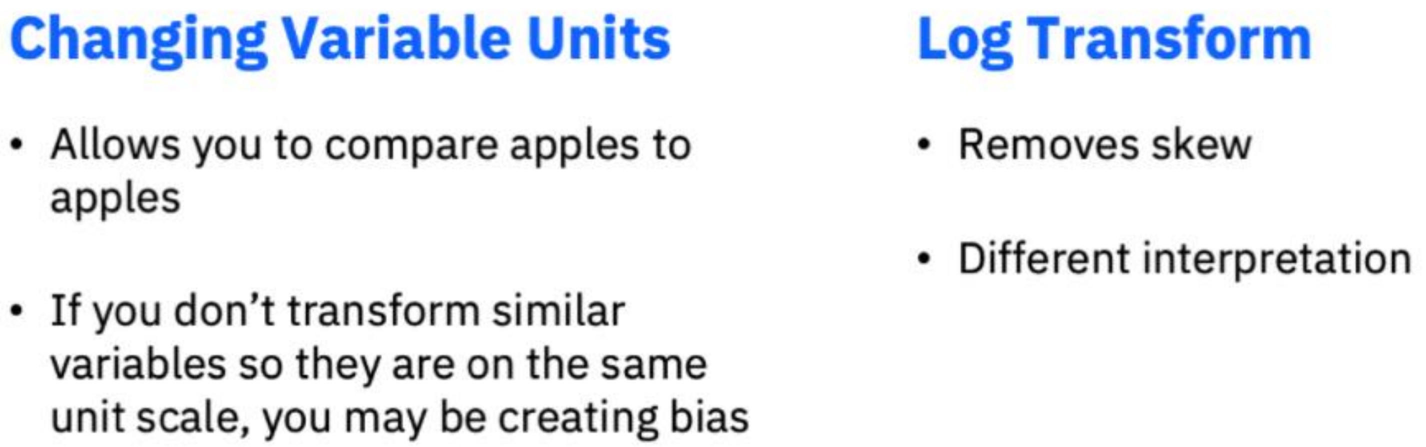
Sum		99	82.51	99	82.51	99	82.5	99	82.51
count(n)	11								
avgerage (mean)		**9**	**7.50090909**	**9**	**7.50090909**	**9**	**7.5**	**9**	**7.50090909**
variance s^2		**11**	**4.12726909**	**11**	**4.12762909**	**11**	**4.12262**	**11**	**4.12324909**

6. Data Transformation

data representation is often followed by data transformation.
transformations are done depending on the problem at hand.

Data Scaling

Data Transformation is an important step for machine learning



Changing Variable Units

- Allows you to compare apples to apples
- If you don’t transform similar variables so they are on the same unit scale, you may be creating bias

Log Transform

- Removes skew
- Different interpretation

Data Normalization

What is data normalization?

The process of rescaling the data into a specific range, usually [0, 1] or [-1, 1].

Why data normalization?

When you want to disregard the magnitude of features and focus on relative importance.

Convenient for training

Helps preventing under- and overflow of weights during the gradient descent algorithm.

Recommended transformation in machine learning: $z = (x - \min)/(\max - \min)$

Data Standardization

What is data standardization?

The process of rescaling the data such that the mean becomes zero and the standard deviation becomes 1.

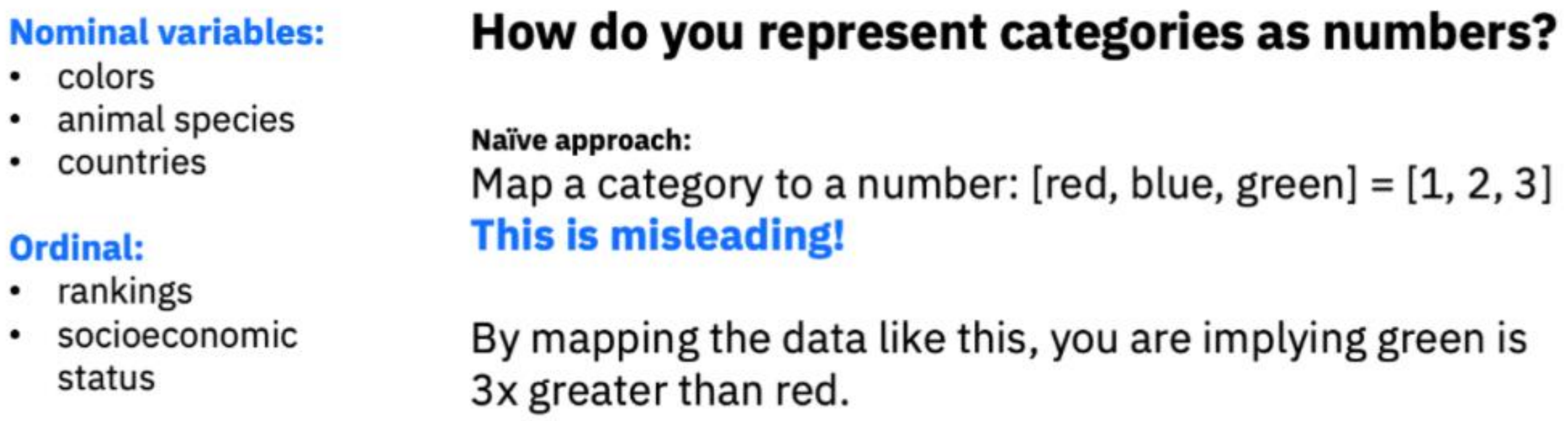
Why data standardization?

Can be useful in certain classification tasks where the shape of the histogram differs from class to class.

Recommended transformation in machine learning: z = (x – mean)/std

7. Representing and Transforming Unstructured Data

Categorical Variables must be mapped to a number in order to be used by a machine learning model



Nominal variables:

- colors
- animal species
- countries

Ordinal:

- rankings
- socioeconomic status

How do you represent categories as numbers?

Naïve approach:

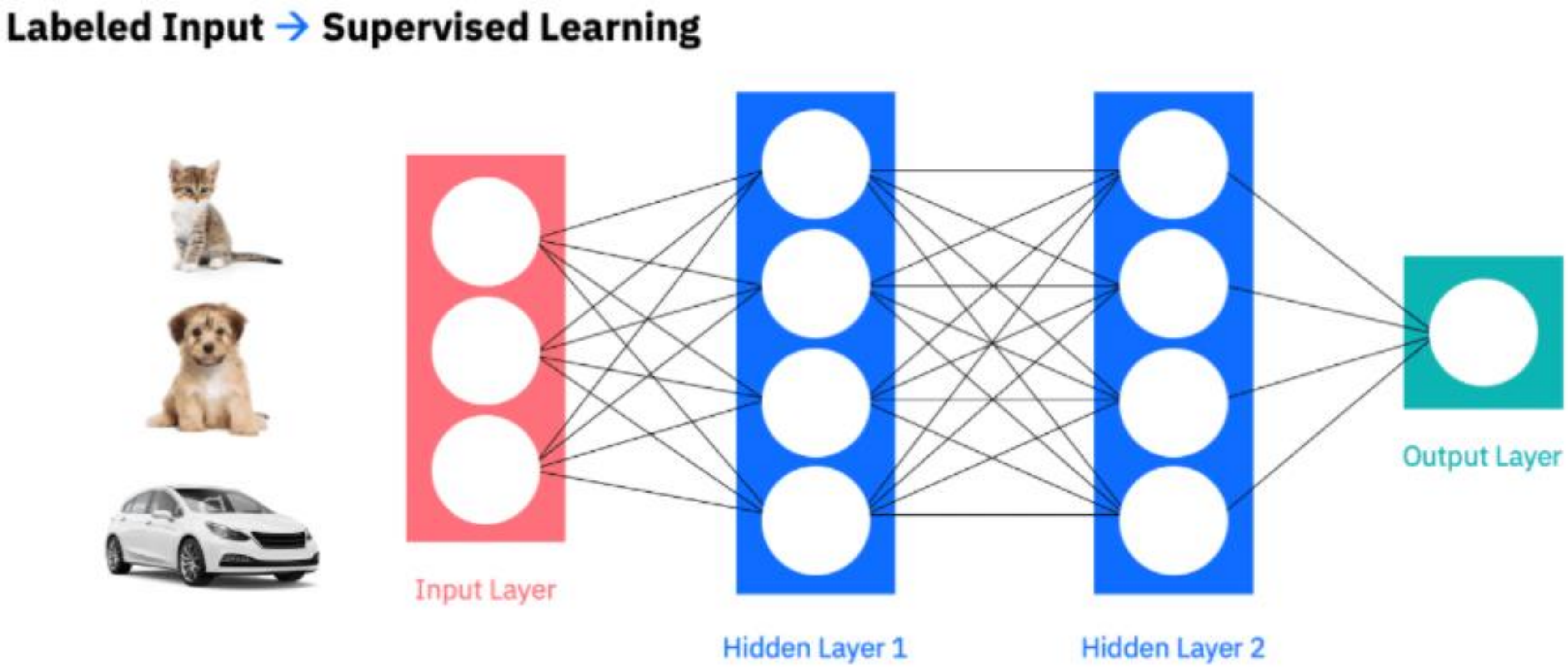
Map a category to a number: [red, blue, green] = [1, 2, 3]

This is misleading!

By mapping the data like this, you are implying green is 3x greater than red.

Representing Multicategory Data with 1-hot Encoding

1-hot Encoding allows you to encode categorical data into numbers.



What if your data comprises images, tweets, or videos?

[x,x,x]

[1,0,0]

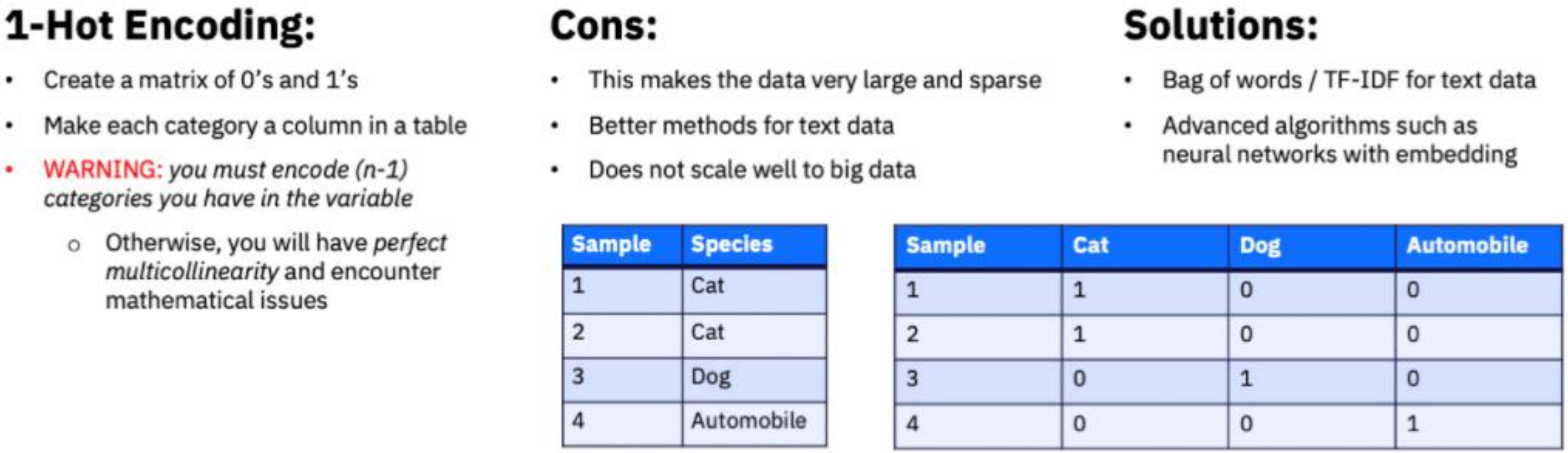
[x,x,x]

[0,1,0]

Index or element in a vector

[x,x,x]

[0,0,1]



1-Hot Encoding:

- Create a matrix of 0's and 1's
- Make each category a column in a table
- WARNING: *you must encode (n-1) categories you have in the variable*
  - Otherwise, you will have *perfect multicollinearity* and encounter mathematical issues

Cons:

- This makes the data very large and sparse
- Better methods for text data
- Does not scale well to big data

Solutions:

- Bag of words / TF-IDF for text data
- Advanced algorithms such as neural networks with embedding

Sample	Species
1	Cat
2	Cat
3	Dog
4	Automobile

Sample	Cat	Dog	Automobile
1	1	0	0
2	1	0	0
3	0	1	0
4	0	0	1

8. Data Transformation Tools

Tools that can assist in data transformation are:

- Excel
- SQL
- Python
- R
- Tableau
- PowerPoint

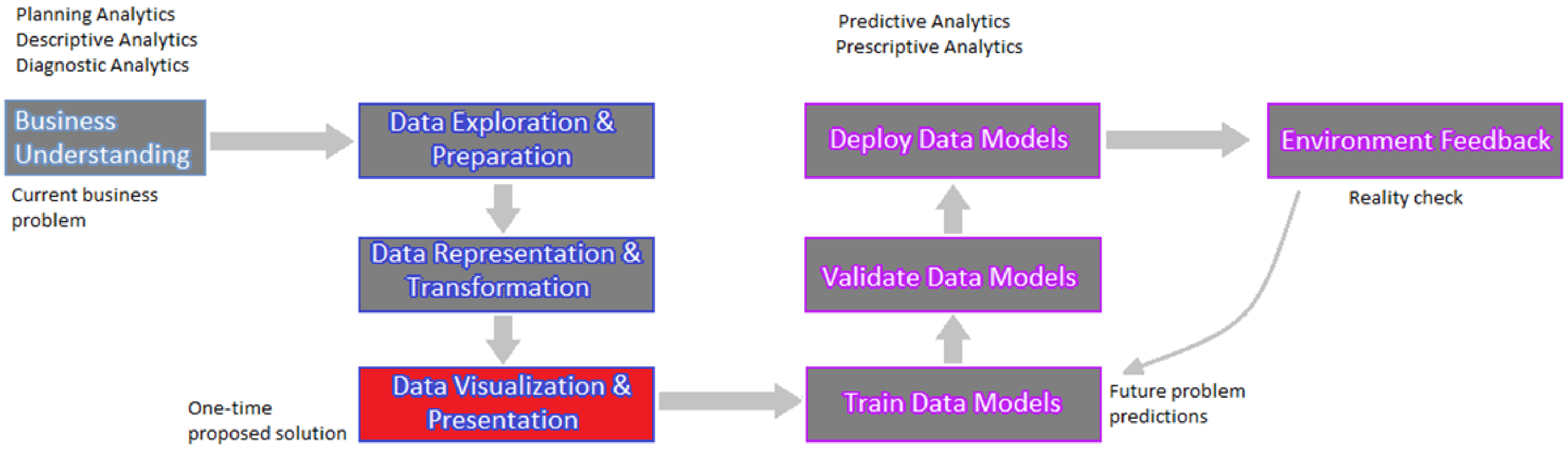
Technologies:

- Apache Spark
- Jupyter Notebooks

Open-source libraries:

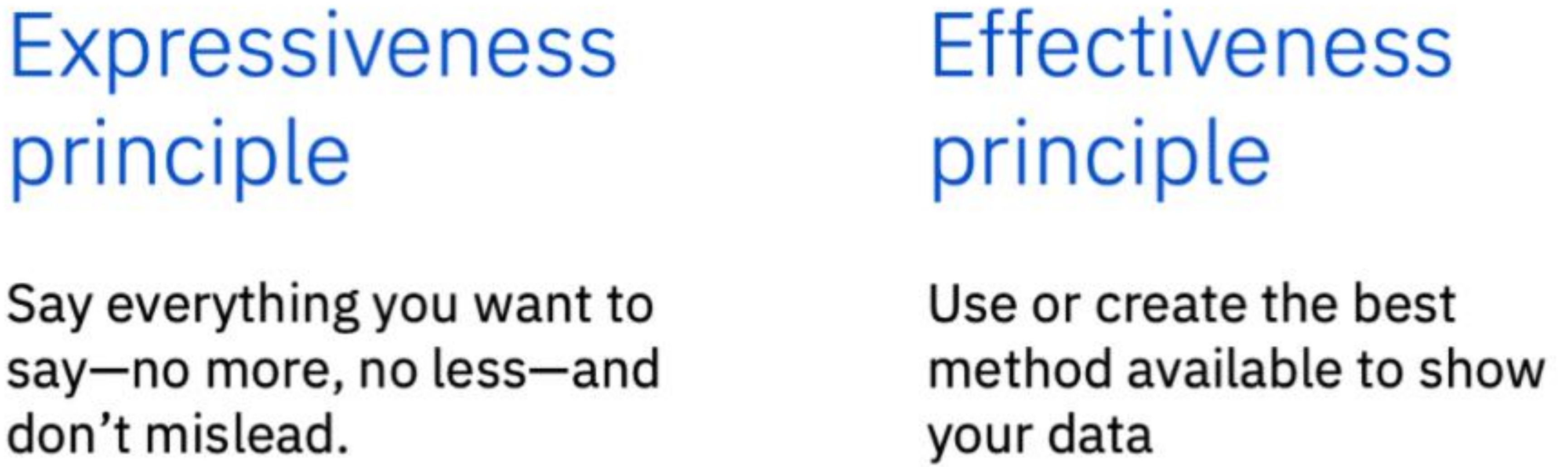
- SciPy
- NumPy
- Scikit-learn
- Pandas
- Natural Language Toolkit

9. Decision-Centered Visualization



Structure and Style

Consider the following two key principles when visualizing:



Expressiveness principle

Say everything you want to say—no more, no less—and don’t mislead.

Effectiveness principle

Use or create the best method available to show your data

Human-centered Reflection

Designing a data visualization goes beyond an aesthetic exercise.

Purpose

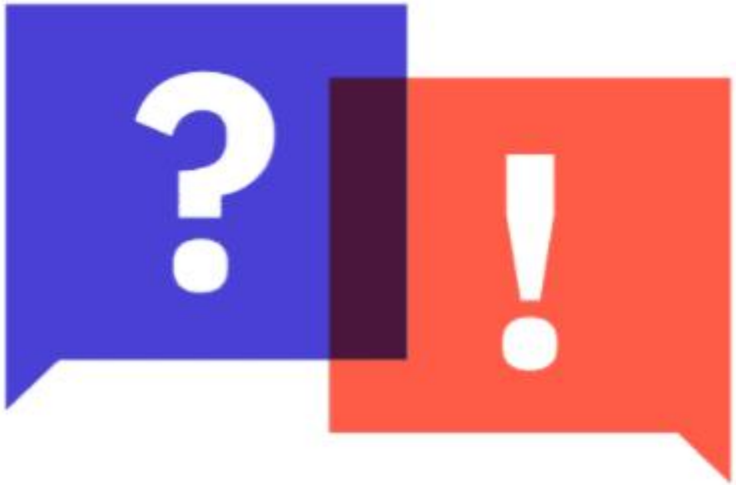
Audience

Data

Context

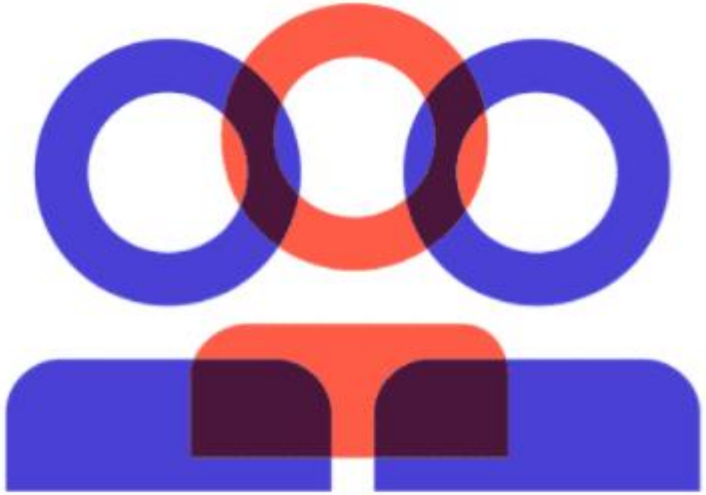
Purpose

- **Where are you starting from**—a user need, a data set, a request from a manager or exec?
- **What problem are you trying to address** and why will data visualization help to solve it?
- **What goals** do you hope to accomplish with the vis?
- **What is the nature of your intention**—to make a point, tell a story, provide deep exploration?



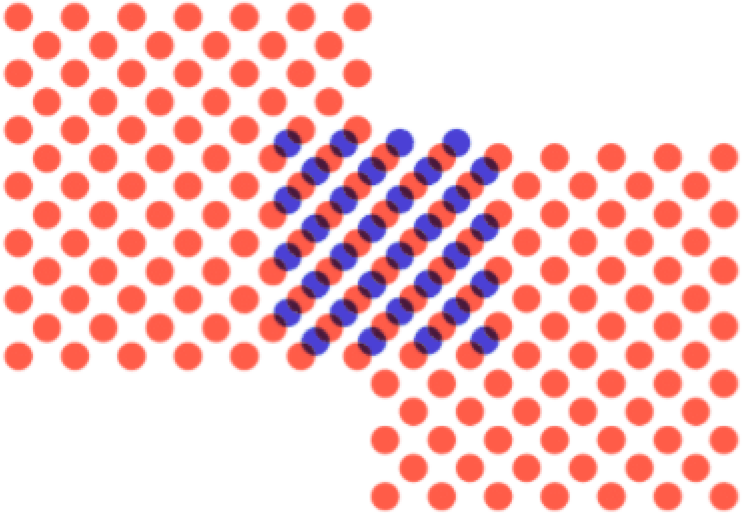
Audience

- **Who is the target** user for your data vis?
- **What does your user want** to do with their data?
- **What cultural, domain, or industry-specific** needs does your user have for the visualization?
- **What user outcomes** will indicate you've been successful?



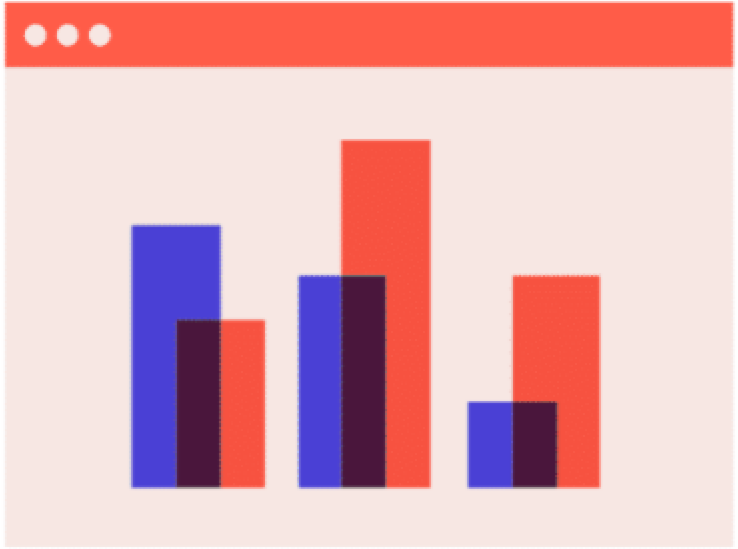
Data

- **Do you have a usable** data set?
- **Are you designing mock-ups** with real data?
- **Will the visualization need** to get periodically updated?
- **What is your plan** to make the visualization accessible?
- **What is your strategy** for language support?



Context

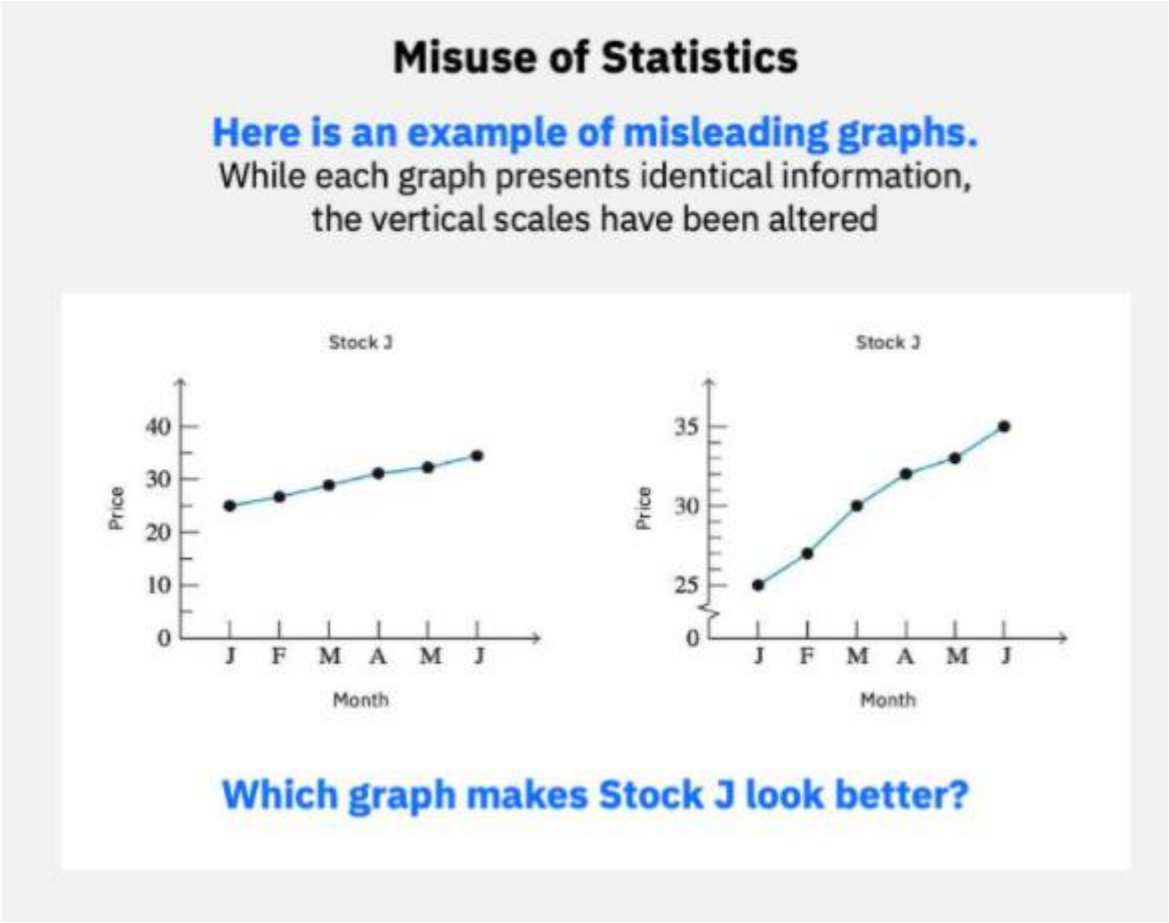
- **Where will the data vis live** — in software or a website, a report or presentation, an article or blog post?
- **Where will your user be** when viewing or exploring the data vis?
- **Is it going to be static or dynamic**, passively consumed or interactive?



Exploratory Data Analysis (EDA)

Refers to the critical process of performing initial investigations on data so as to discover:

- Relationships and trends without a specific goal in mind
- Whether it structured or unstructured data
- Spot anomalies
- Test hypothesis and check assumptions
- Note summary statistics that can be misleading!



Misuse of Statistics

Here is an example of misleading graphs. While each graph presents identical information, the vertical scales have been altered

Which graph makes Stock J look better?

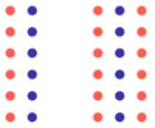
10. Fundamentals of Visualization

10.1. Perception

Different aspects are considered when doing visualization:

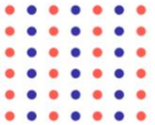
- grouping
- categorization
- sorting

Grouping



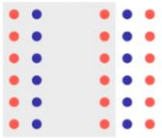
Similarity

Elements are perceived as groups depending on the visual Characteristics they share, like color or value.



Proximity

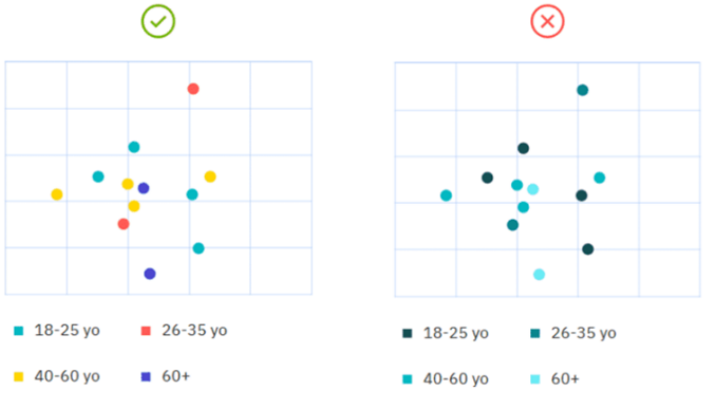
Stronger than similarity, the human eye perceives elements to be related based on how close they are to one another.



Enclosure

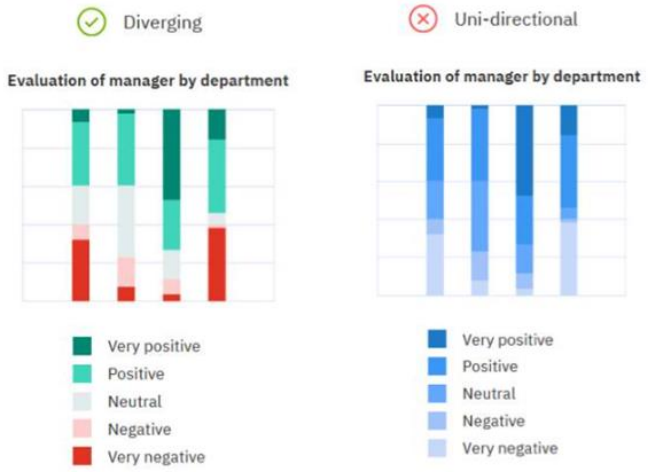
Introduced by Palmer in 1992, the common region principle shows how enclosing elements in other elements helps people see individual items as distinct groups.

Categorization



Color schemes for nominal data typically use different hues to identify discrete categories.

Sorting



Data that progresses from **low to high** can be communicated with a sequential color scheme.

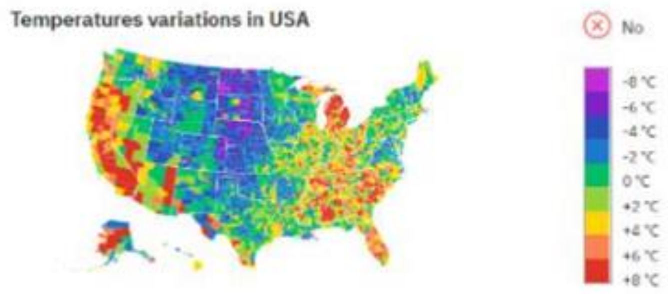
Data that progresses outward from a **middle value** can be represented with a diverging scheme.

Use light colors for the middle data value and dark colors for the end values.

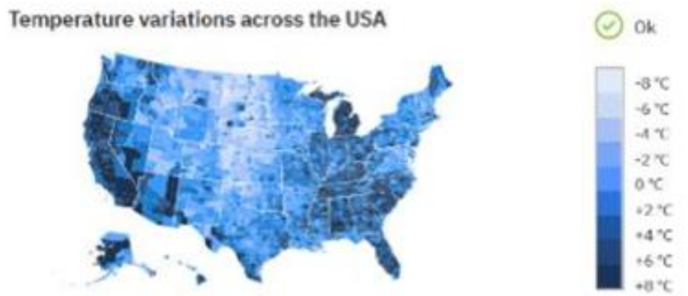
10.2. Intervals and Ratios

Breaking up quantitative data values into discrete classification or bins makes them easier to read than using a continuous gradient scale.

Rainbow

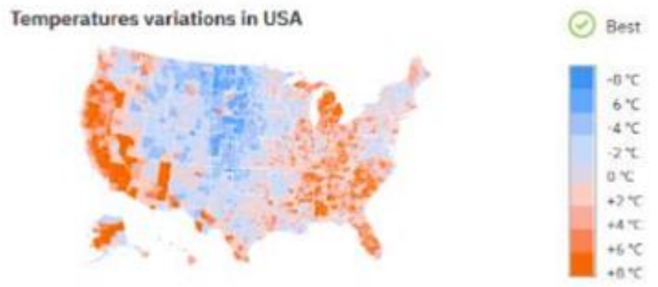


Uni-directional



While using a uni-directional palette when displaying positive/negative temperatures is actually correct, but is it the most effective way to communicate the data? Consider the context of the data when attempting to show its meaning.

Divergent

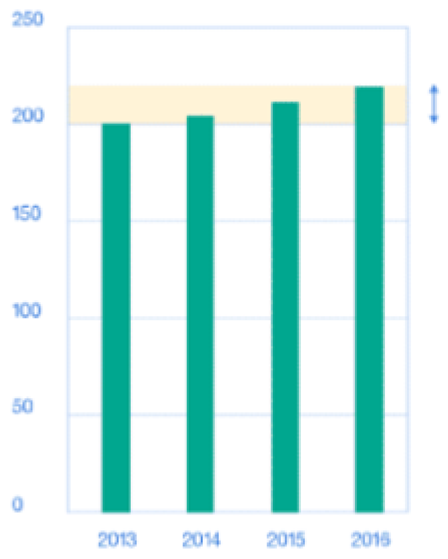


10.3. Manage Tricky Situations

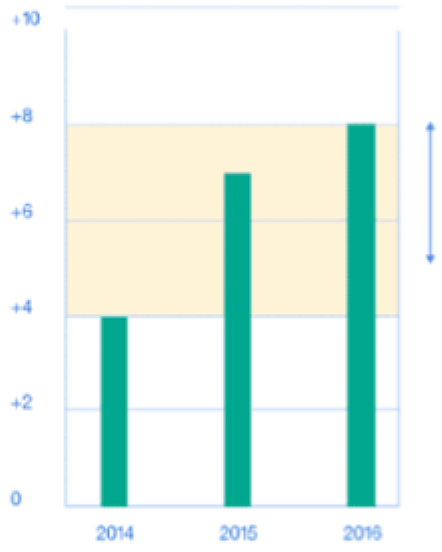
Size variations (bar height, bubbles size, or line segments) can sometimes be undistinguishable.

Possibly display only the **changes** (instead of absolute values).

Revenues



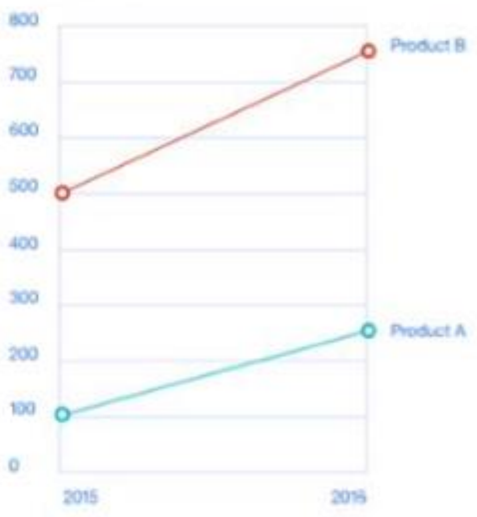
Revenues Variation



10.4. Comparable Scales

Sometimes, we need to compare entities that have strong different scales.
In these cases, percentages can tell much more than absolute values.

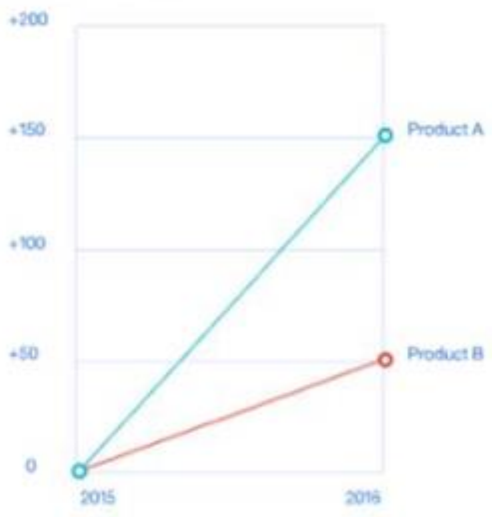
Revenues Trend



Focus on Actual Values

The chart is correct if we want to show the different weight of the products, but it is less effective in showing their performance.

Revenues Trend (%)



Highlight Performance

Using percentage in the chart, the reader immediately understands the Product A outperformed Product B

10.5. Use Reference Points

Make the main idea pop out, calibrating what is around it.

fig. A

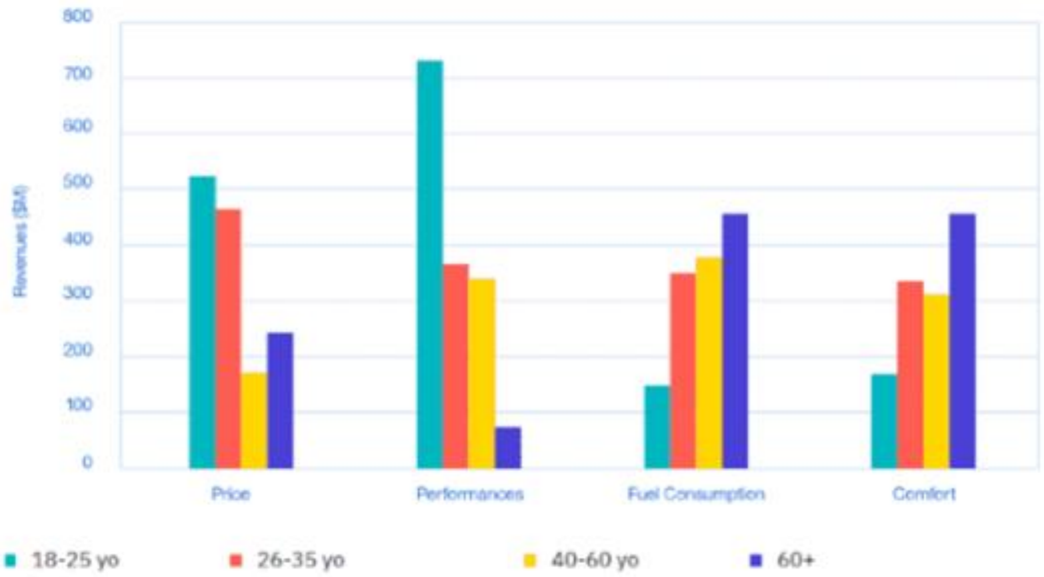
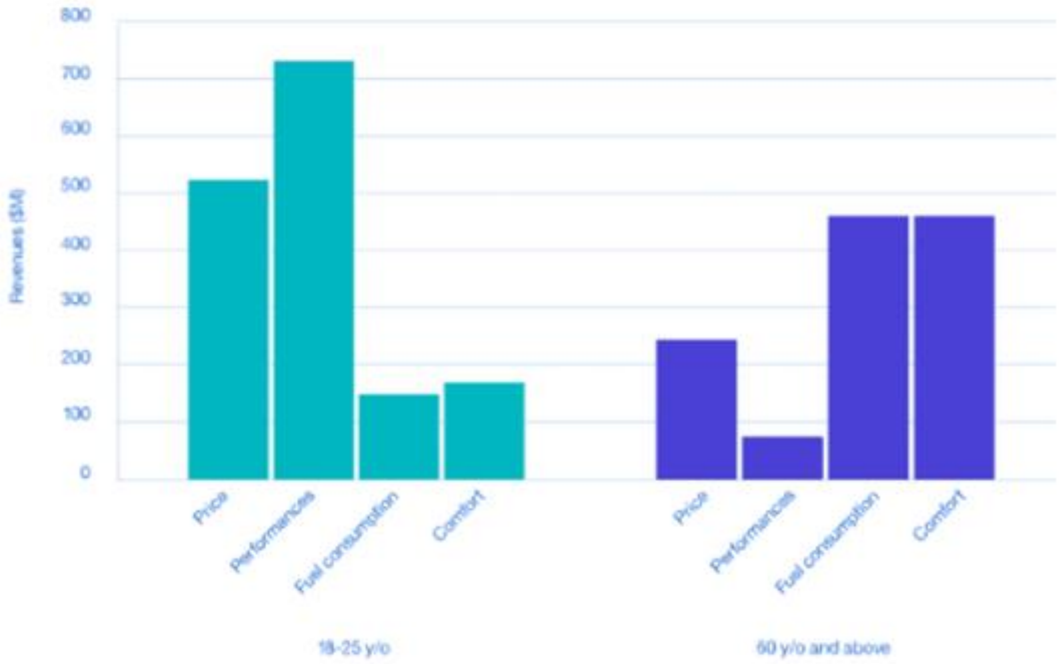


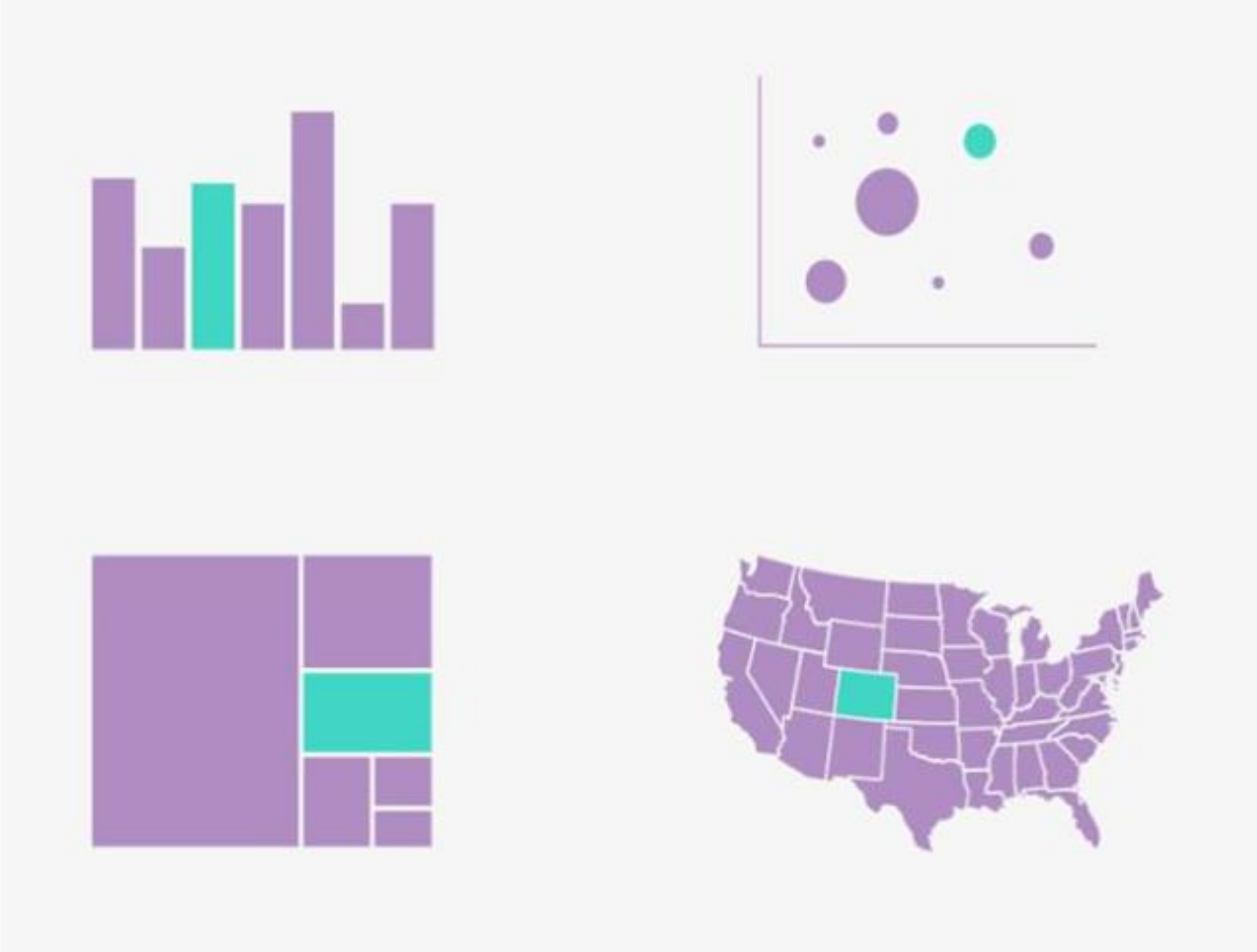
fig. B



10.6. Create Interactions Between Multiple Visualizations

Create dialog between multiple visualizations

When visualizations are using the same data on a single screen, give the user a way to identify patterns among views.



10.7. Foster Iterative Data Interpretation

Annotate

Provide users with tools for recording, organizing, and communication insights gained during exploration.

Record & Archive

After conducting analysis, users need to review, summarize and communicate their findings, often in the form of reports or presentations.

Share

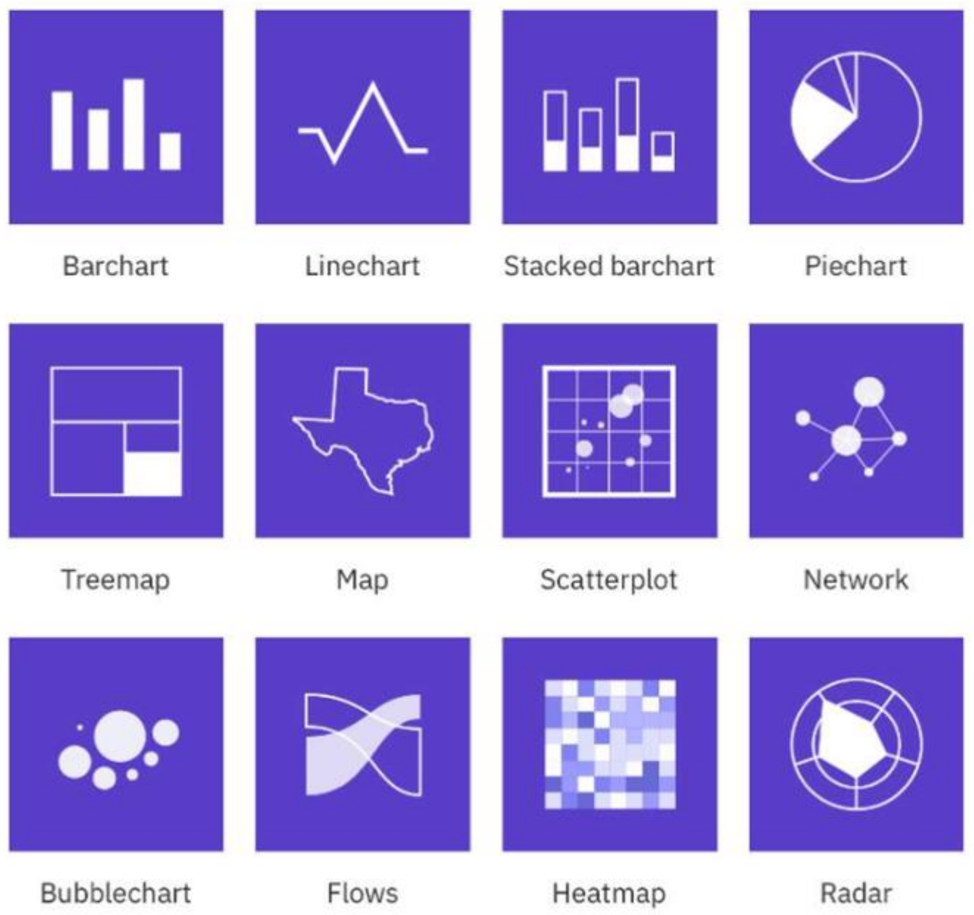
Collaboration, with social interaction and multiple interpretations, is fundamental to the analysis process.

11. Common Graphs

11.1. Introduction

Explore visualizations based on your intent.

These charts are a curated set of visualizations for a range of common needs.



11.2. Bar, Line, and Pie Charts

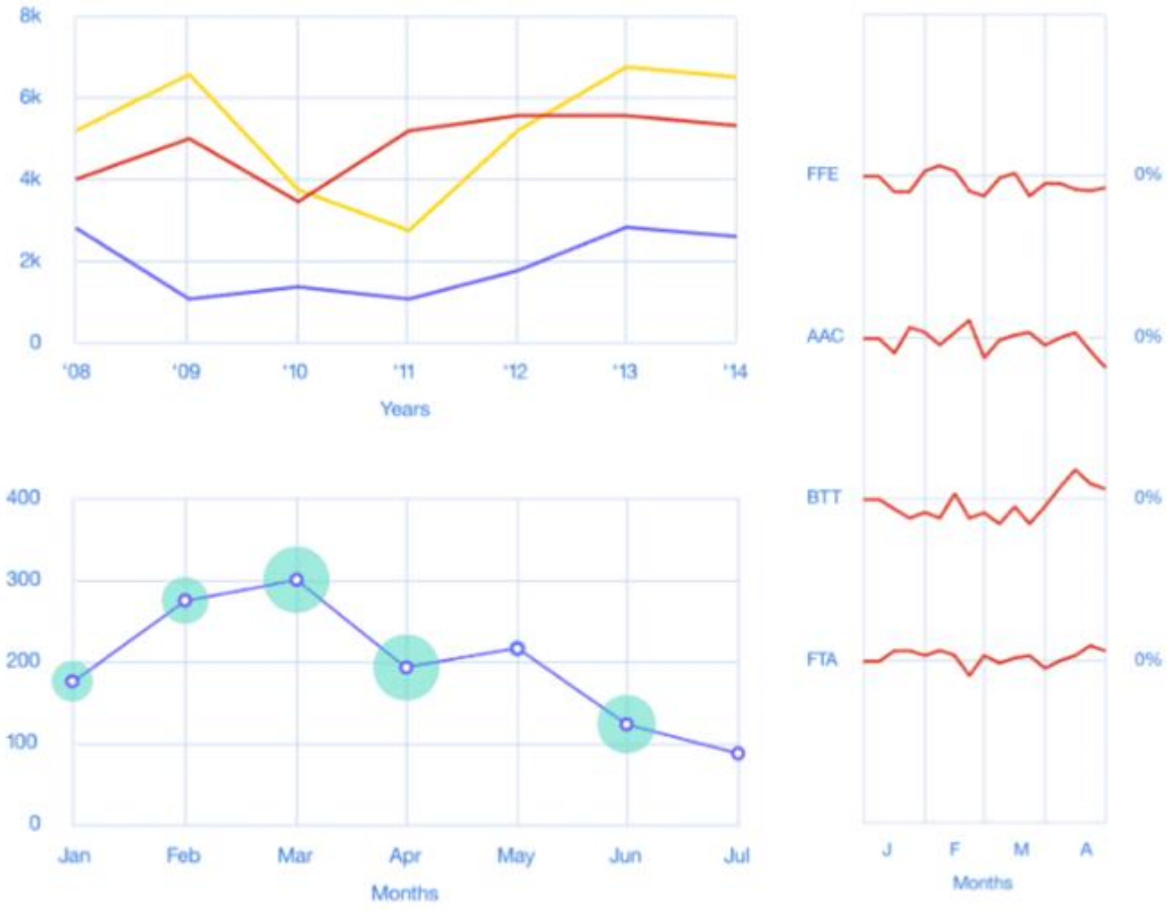
Line chart

This graph model displays information as a series of data points connected by straight line segments.

I'm going to use this model when I want to:
explore in time | compare | show correlations

I'm going to use this model when I have this kind of data set:
time-based data

Not recommended for:
Avoid if not comparing values over time, as it might create confusion. Select a bar graph in this case



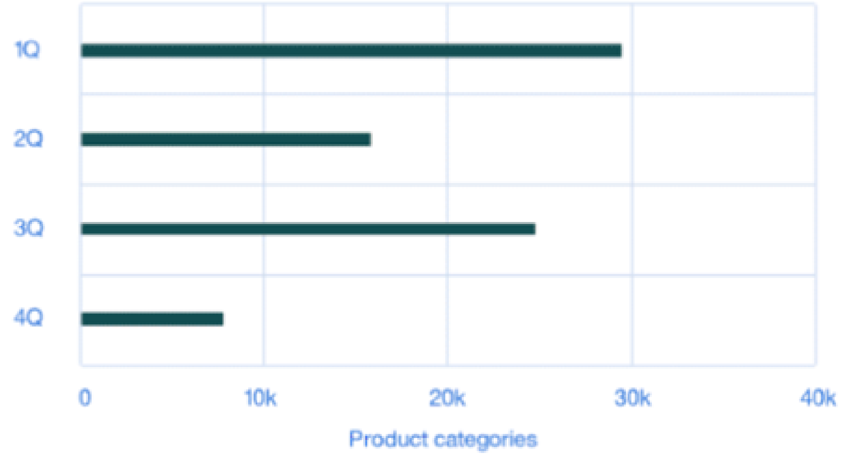
Bar chart

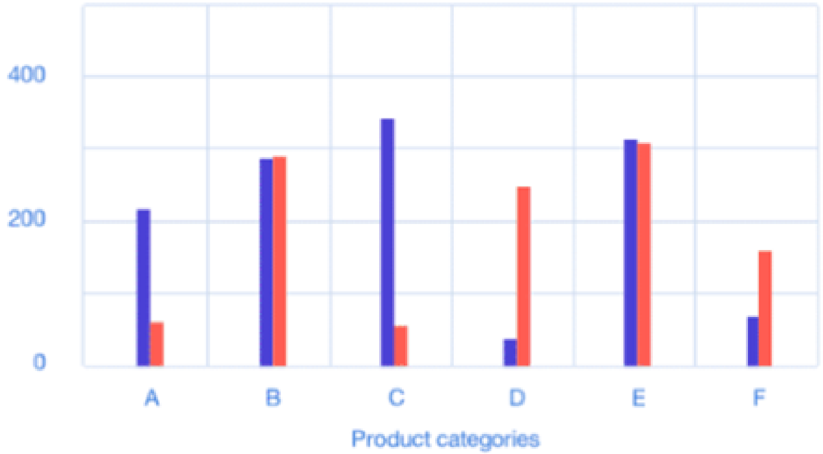
Rectangular bars with lengths proportional to the values they represent.

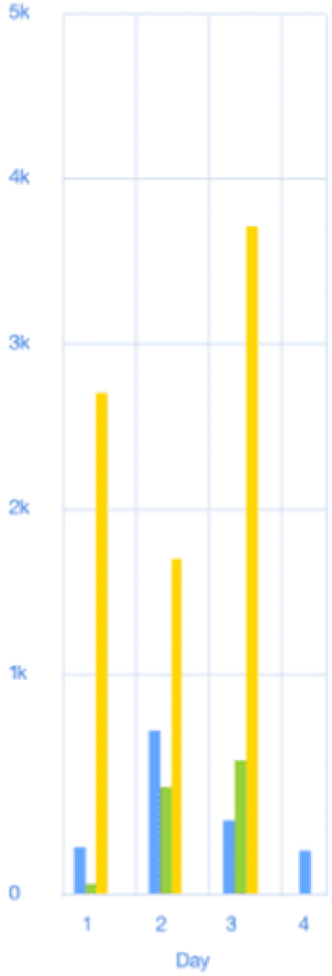
I'm going to use this model when I want to:
explore in time | compare | show correlations

I'm going to use this model when I have this kind of data set:
time-based data | categorized data

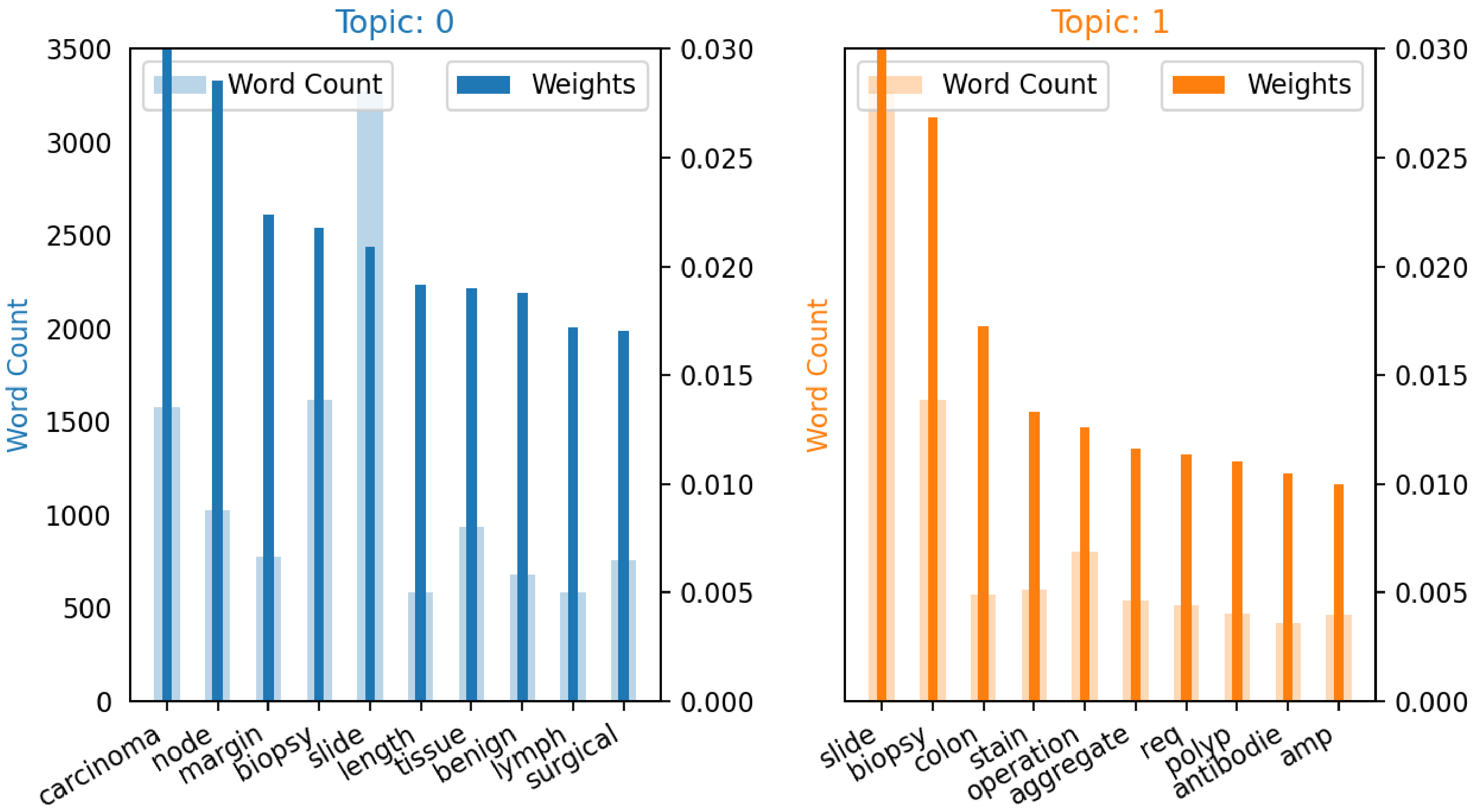
Not recommended for:
Never use to compare values with different units or hierarchy.



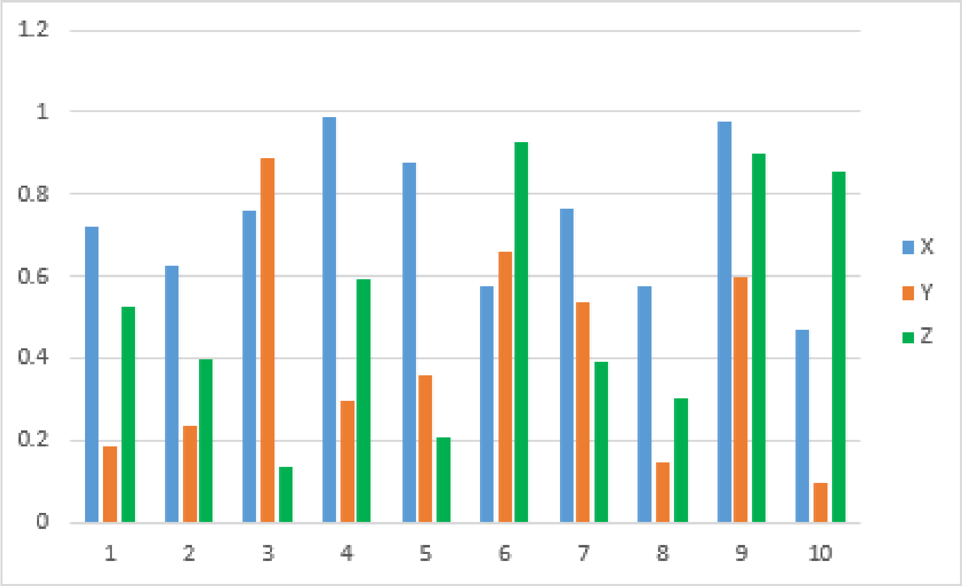


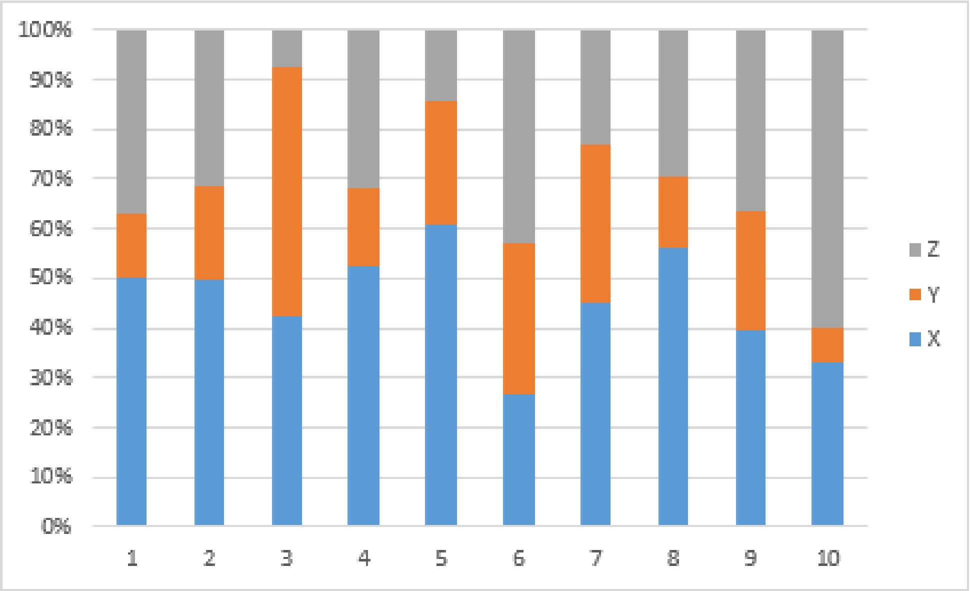


Example of Bar Charts for Categorical Data



Converting a Categorical Bar Chart into a Stacked Bar Chart





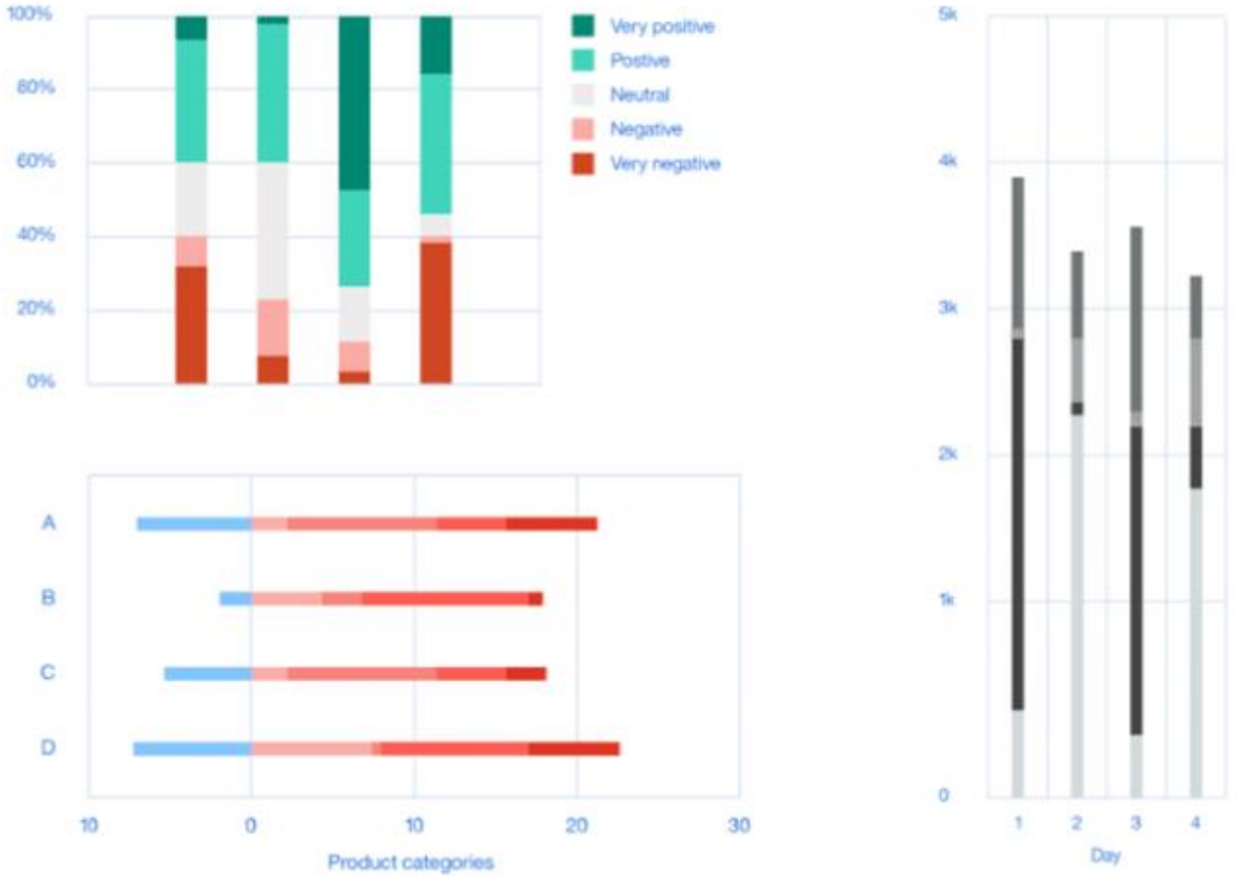
Stacked bar chart

A variant of the bar graph, where each rectangle is divided in multiple parts.

I'm going to use this model when I want to:
explore in time | compare | show correlations | show subdivisions

I'm going to use this model when I have this kind of data set:
time-based data | categorized data

Not recommended for:
Never use when the focus is on comparing the sizes of the individual categories or when the total sum of the elements in the bar is not relevant.



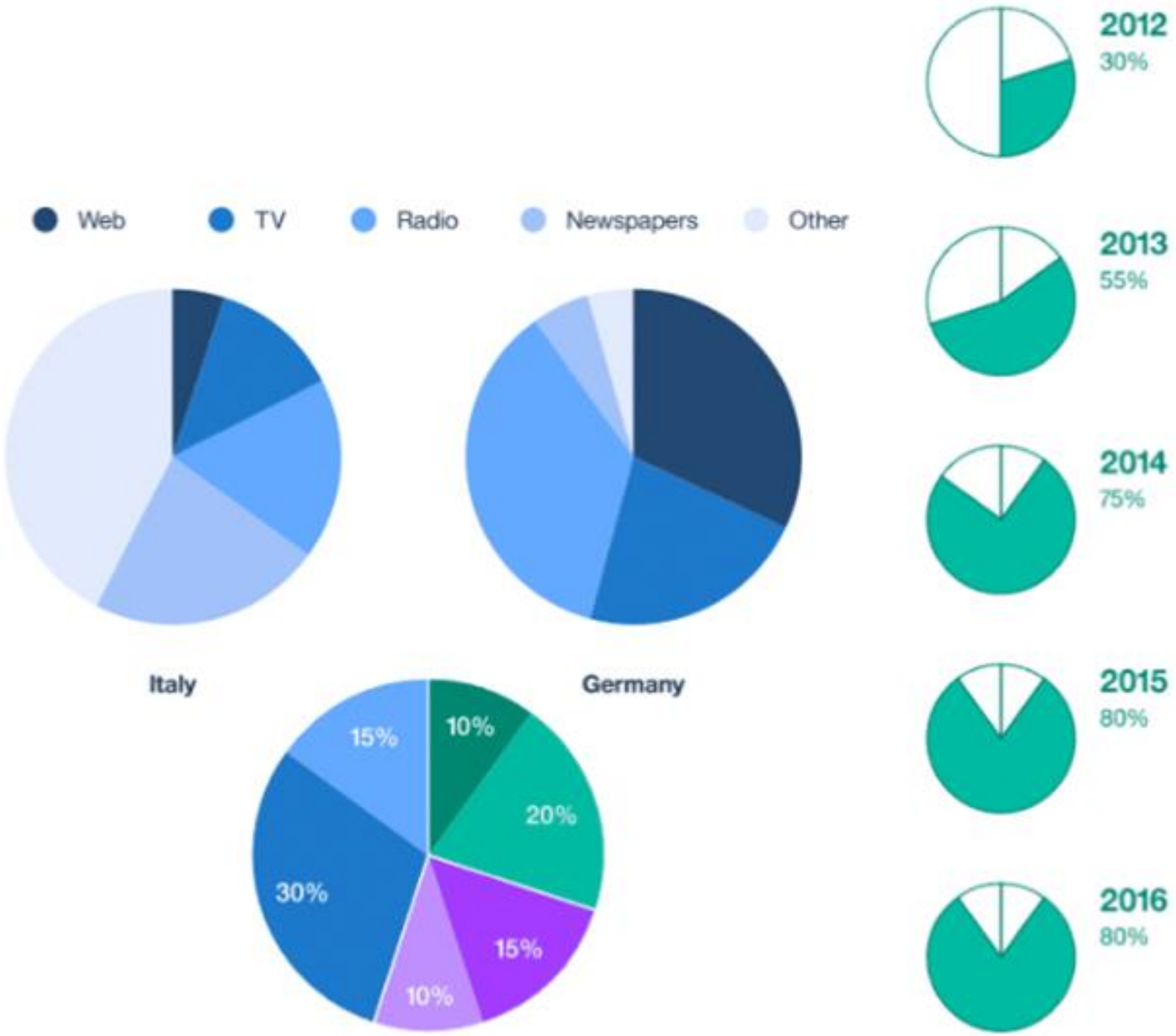
Pie chart

Circular graph model divided into sectors, illustrating proportions.

I’m going to use this model when I want to:
compare | show subdivisions

I’m going to use this model when I have this kind of data set:
categorized data

Not recommended for:
Don’t use when you have more than six categories.



11.3. Maps

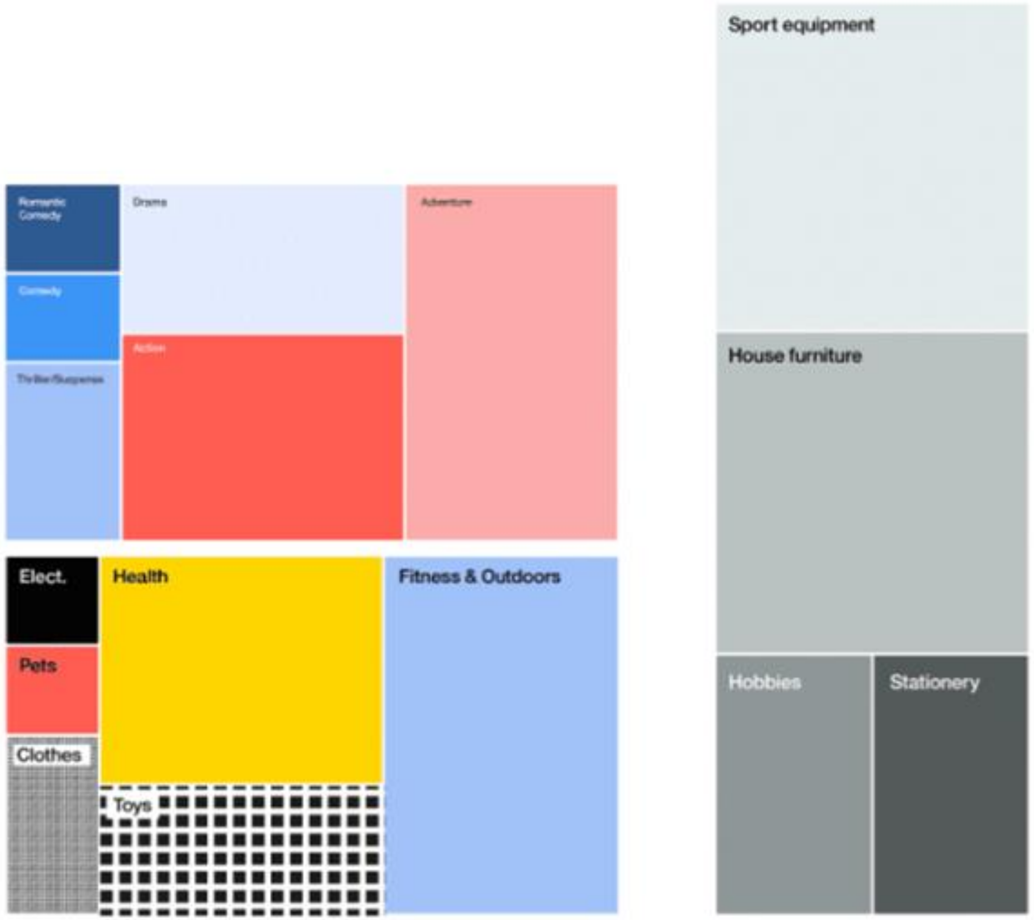
Treemap

Displays hierarchical data as a set of nested rectangles, which parts combined, make a larger rectangle.

I'm going to use this model when I want to:
compare | show subdivisions

I'm going to use this model when I have this kind of data set:
categorized data | geographic distribution

Not recommended for:
Don't use a tree map for data grouped in more than 25 different categories.



Map

Cartography is used to display geographical data.

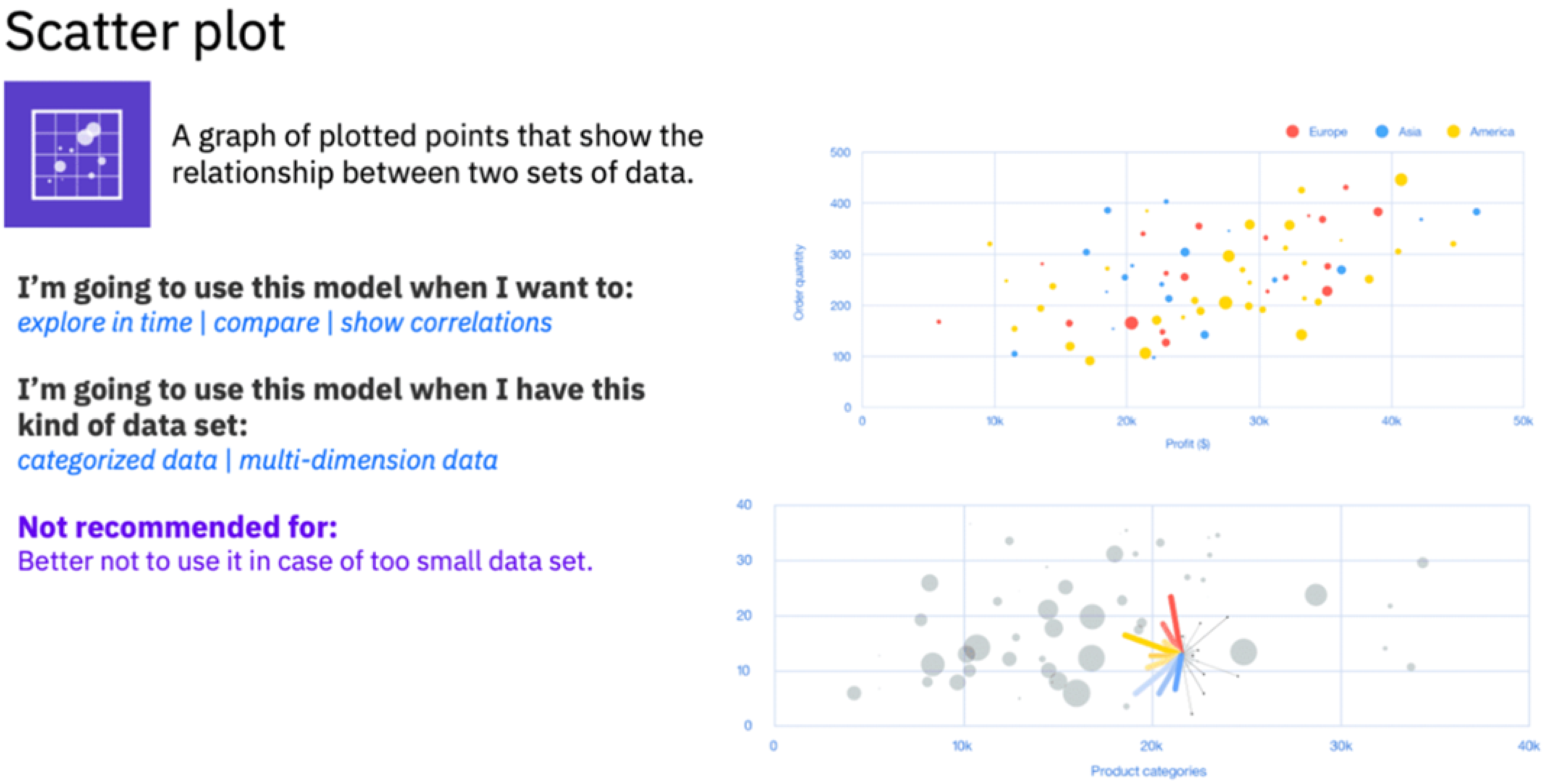
I’m going to use this model when I want to:
explore in time | compare | distribute geographically

I’m going to use this model when I have this kind of data set:
geographic distribution

Not recommended for:
Don’t use it if the data set has geographical data that’s not relevant to your use case.

-8 °C
-6 °C
-4 °C
-2 °C
0 °C
+2 °C
+4 °C
+6 °C
+8 °C

11.4. Scatter Plots



Scatter plot

A graph of plotted points that show the relationship between two sets of data.

I’m going to use this model when I want to:
explore in time | compare | show correlations

I’m going to use this model when I have this kind of data set:
categorized data | multi-dimension data

Not recommended for:
Better not to use it in case of too small data set.

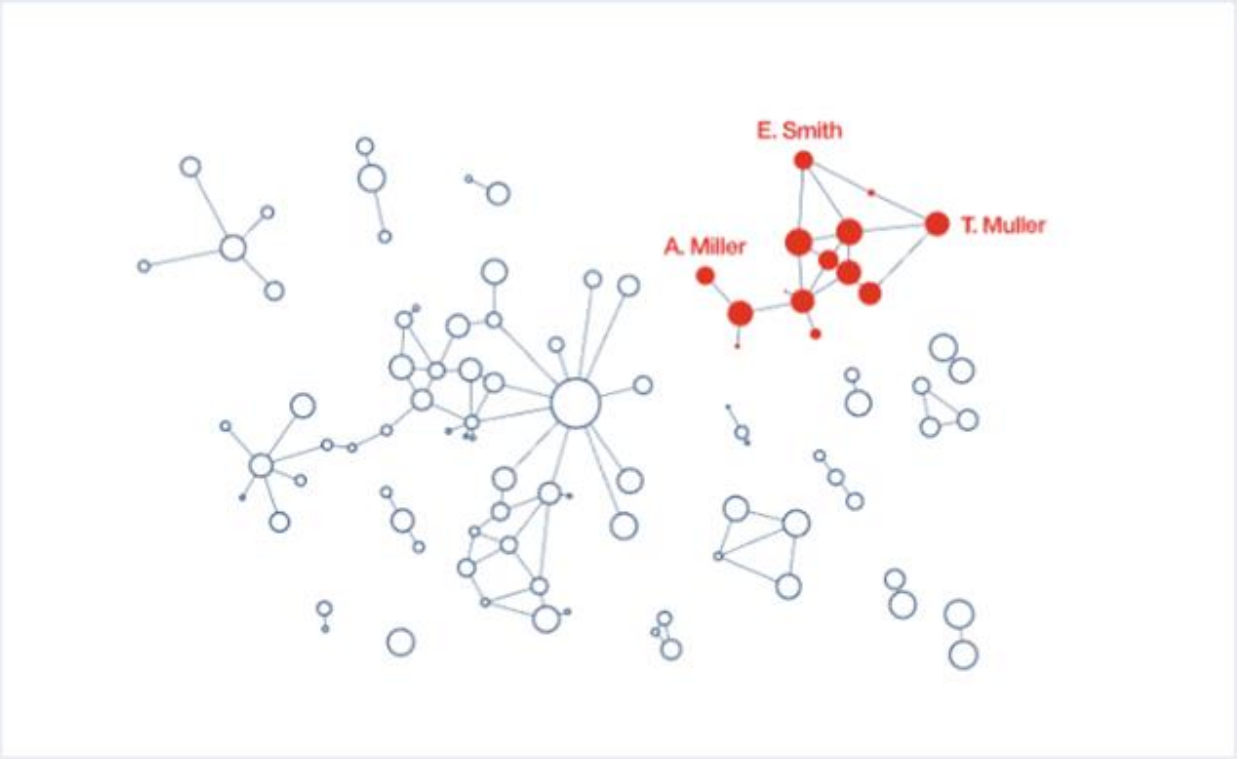
Network

A graph where nodes are connected and positioned depending on their mutual relationship.

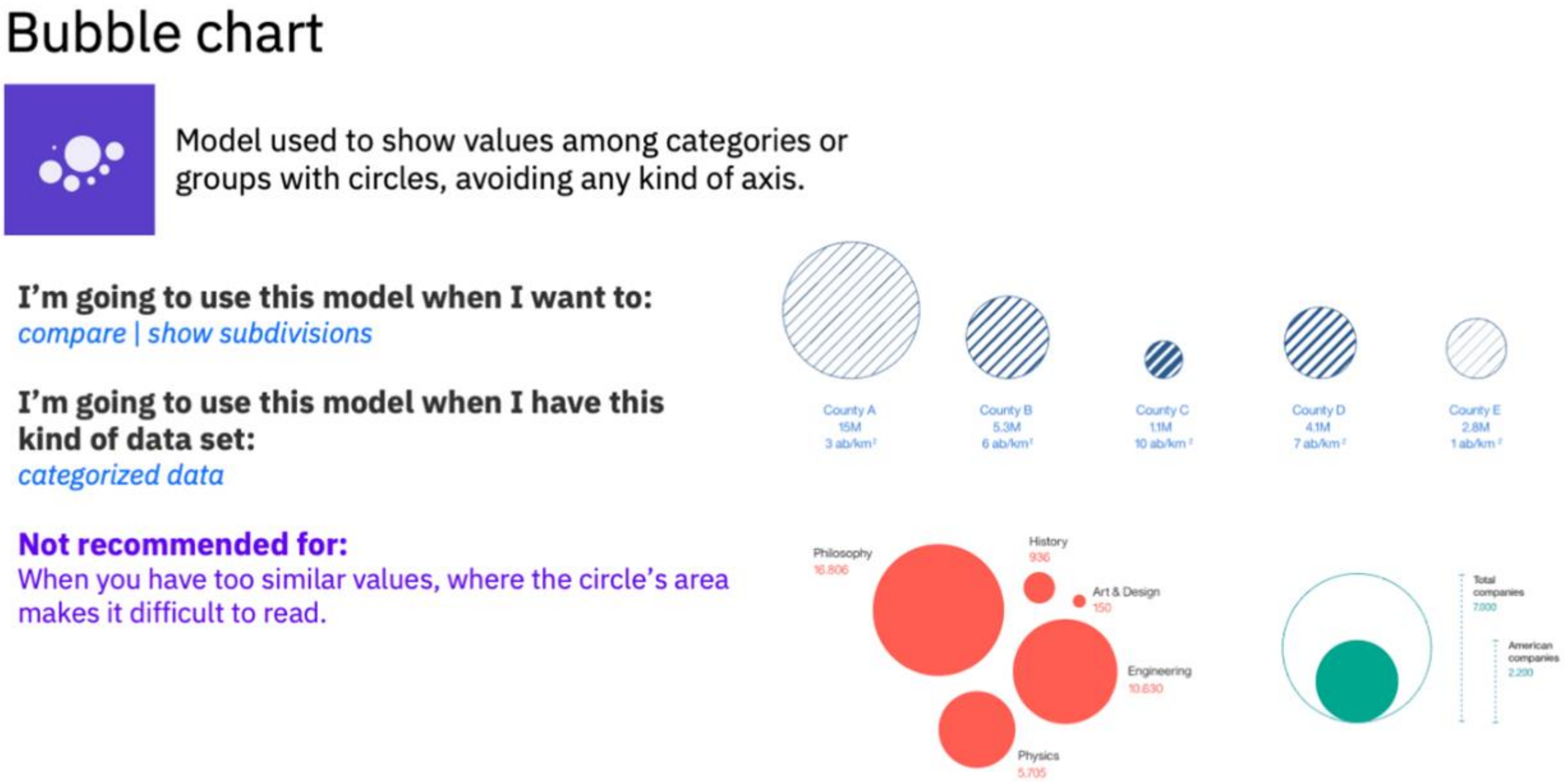
I'm going to use this model when I want to:
show relationships

I'm going to use this model when I have this kind of data set:
multi-dimension data

Not recommended for:
Hard for beginners and common users to understand, better for experts.



11.5. Bubble and Flow Charts



Bubble chart

Model used to show values among categories or groups with circles, avoiding any kind of axis.

I'm going to use this model when I want to:
compare | show subdivisions

I'm going to use this model when I have this kind of data set:
categorized data

Not recommended for:
When you have too similar values, where the circle's area makes it difficult to read.

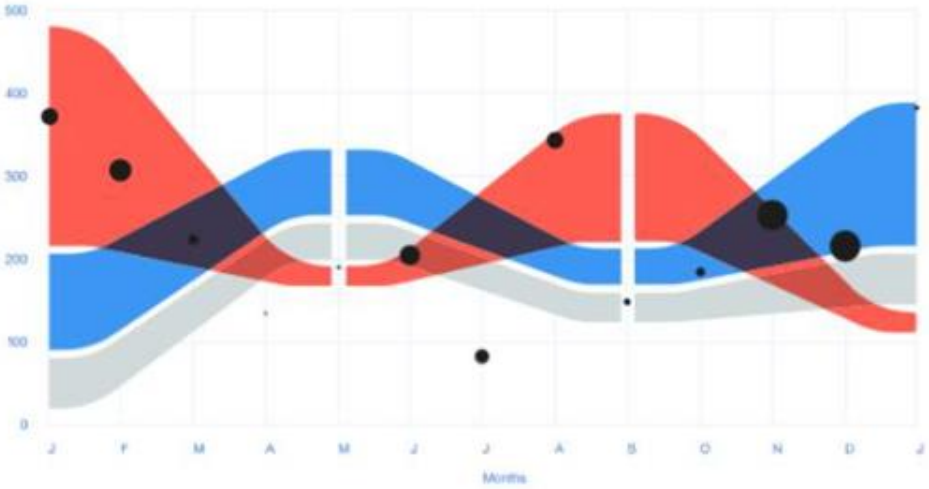
Flows

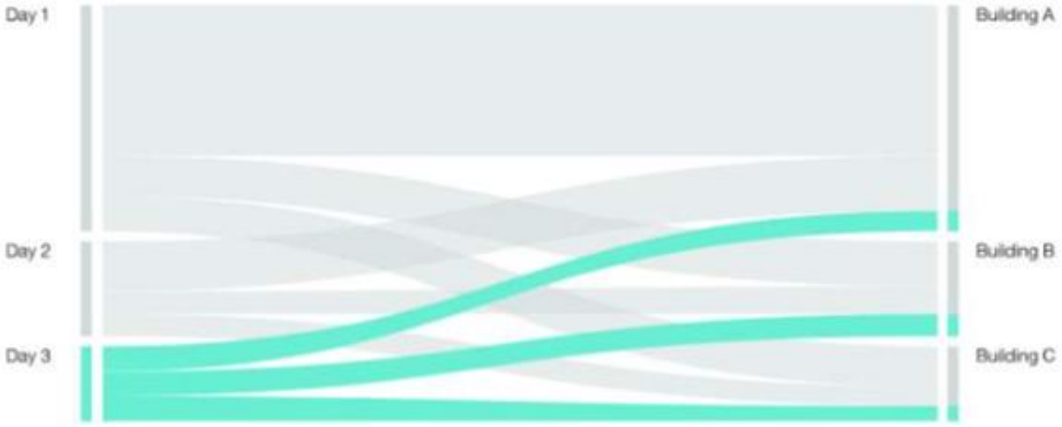
Chart used to show different behaviors among multiple steps and situations.

I'm going to use this model when I want to:
show relationships | show subdivisions

I'm going to use this model when I have this kind of data set:
categorized data

Not recommended for:
A large amount of categories and flows, as it reduces readability.





11.6. Heat and Radar Maps

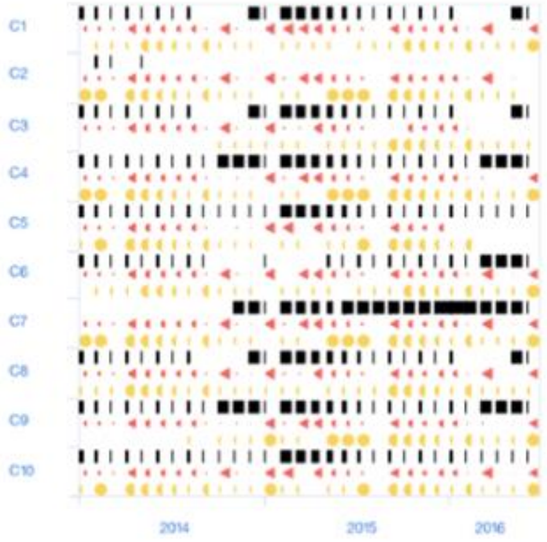
Heat map

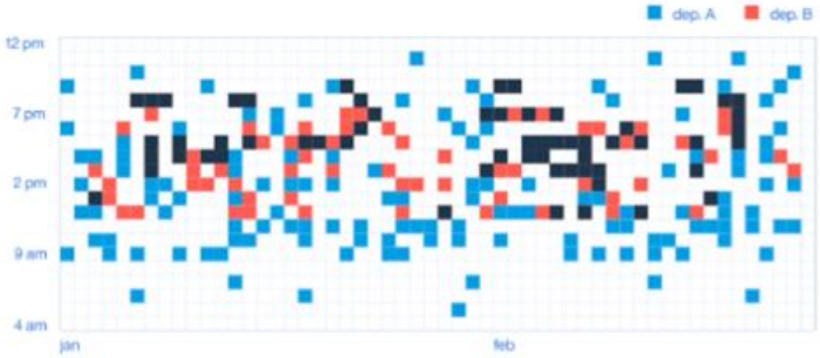
Represents mutual correlations of variables within a data set.

I'm going to use this model when I want to:
show correlations | show relationships

I'm going to use this model when I have this kind of data set:
multi-dimensional data

Not recommended for:
One of the main strengths of a heat map is its ability to highlight patterns. Don't use it when you have only a few indicators.





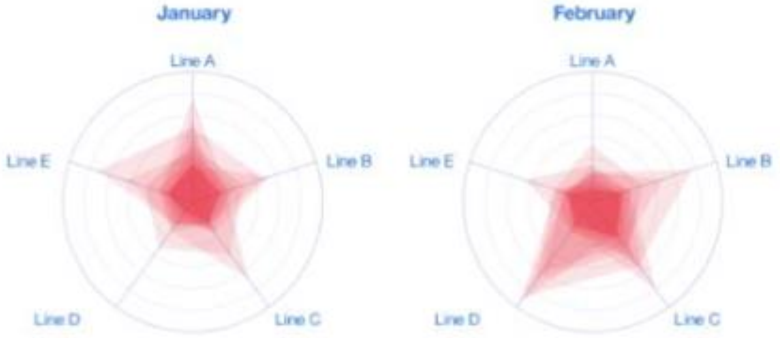
Radar

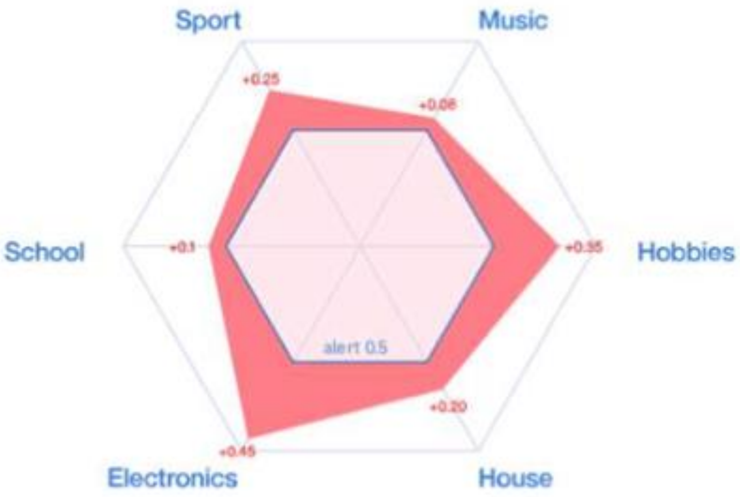
Chart used to represent values of multiple indicators simultaneously.

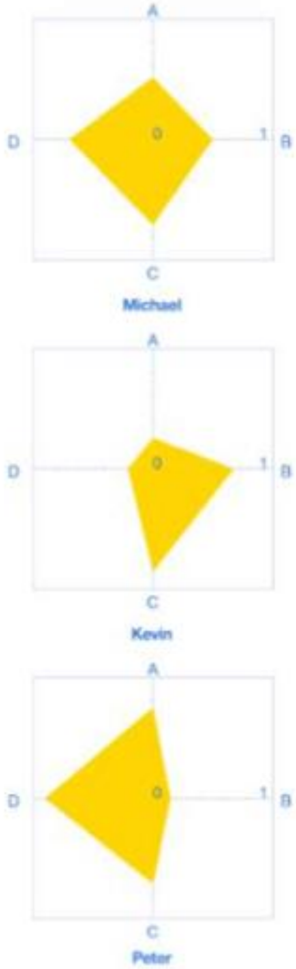
I'm going to use this model when I want to:
show correlations | compare

I'm going to use this model when I have this kind of data set:
categorized data

Not recommended for:
When doing a time comparison, radial representations are not the best to compare lengths.







Thank you!