Lecture 1

Module Logistics and Introduction

Ashish Dandekar

Module Logistics

The Teaching Team

Ashish Dandekar
dcsashi@nus.edu.sg

Vivy Suhendra
vivy@nus.edu.sg

Biswadeep Sen
e0989386@u.nus.edu

Lecture

Every Friday 6.30pm to 8.30pm (SR1)

Tutorial

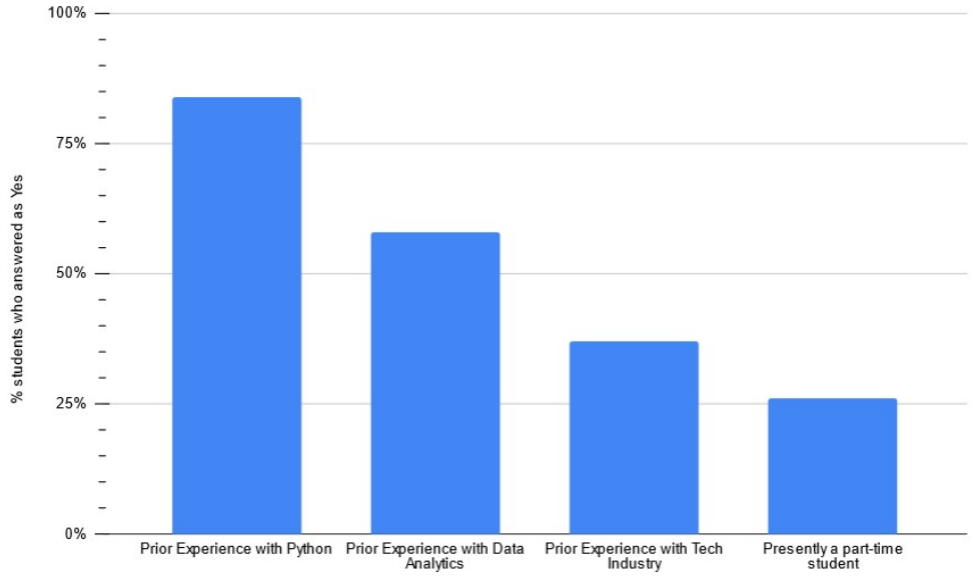
Every Thursday 6.30pm to 7.30pm (LT15) *starts in Week 2*

Class Demographics

- People
- Learning Outcomes
- Assessments
- Term Project
- Lesson Plan

What was your major during the undergraduate degree?

Engineering	18 respondents	**47 %** ✓
Math and Statistics	6 respondents	16 %
Arts and Humanities		0 %
Science	7 respondents	18 %
Law and Public Policy		0 %
Business	7 respondents	18 %

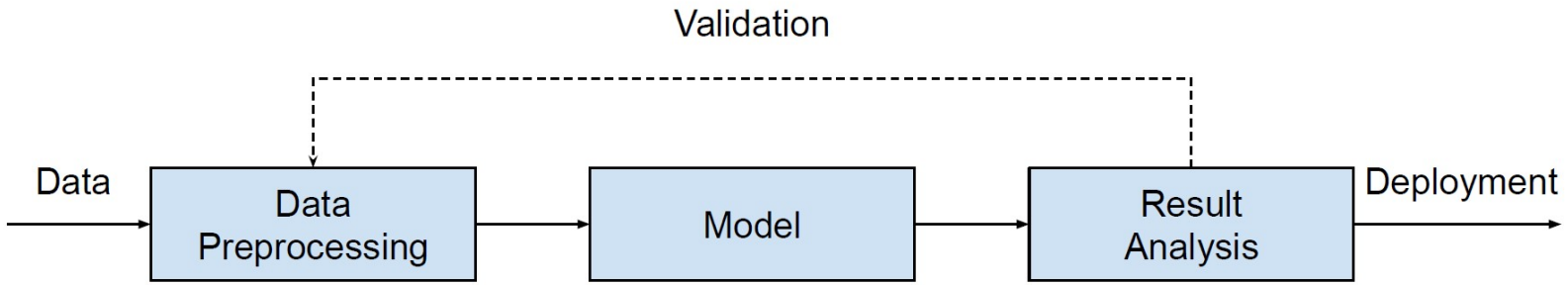


Expectation

What is this module *not* about?

- Big data *(CS5425)*
- Deep learning *(CS5242)*
- AI *(IT5005, CS5446)*
- Fintech *(IS5009)*
- Blockchain *(IS4302)*

What is this module about?



Data preprocessing

- Data cleaning
- Data transformation
- Feature selection

Modeling

- Regression
- Classification
- Clustering

Result analysis

- Qualitative
- Quantitative
- Comparative

Learning Outcomes

What will you learn after the module?

- Design an end-to-end data pipeline.
- Write small scripts in Python to
  - Generate insightful visualizations.
  - Utilize statistical and machine learning toolbox to implement analytics solution.

- Formulate mathematical representations of practical data analytics problems.
- Interpret and analyze the results obtained from the mathematical models.

What do you need to know?

- Python
- Some knowledge of Statistics and Linear Algebra

Assessments

Assignments (20%)

Two take-home assignments in the form of **Jupyter Notebook**.

Term Project (30%)

Interim Exam (20%).

Open-book, open-internet, BYOD exam. Canvas quiz and some coding!

Final Exam (30%).

Open-book, pen-paper based exam. MCQs and numerical problems!

Project Logistics

The Dataset

The combined datasets from Kaggle's *Annual Machine Learning and Data Science Survey* from the past three years (2020-2022)

Project Topic

- Developing a prototypical recommendation engine.
- Developing a webapplication using Streamlit.

Project Team

- Three students per team
- A balanced team

The detailed information can be found HERE.

You can post your questions HERE.

Project Logistics

Milestones

Miestone	Deliverable	Due Date
0: Team Formation	Form a team of three persons each	27th Aug 23:59
1: Exploratory DA	WebApp with EDA	3rd Oct 23:59
2: Recommendation System	WebApp with recommendation system	19th Nov 23:59
3: Presentation	Project Presentation	17th Nov Class time
4: Final Report	Project report	Reading Week

Lesson Plan

Week	Lecture	Tutorial	
1	Introduction and Data Visualisation		
2	Descriptive statistics and Hypothesis testing	Data visualisation with `seaborn`	Milestone 0
3	Linear algebra and Dimensionality reduction	Statistics with `scipy`	
4	Introduction to Machine Learning	Dimensionality reduction	A1 released
5	Regression analysis	ML with `sklearn`	
6	Classification I	Regresion with `statsmodels`	A1 Due
Recess Week (Milestone 1)			

Lesson Plan

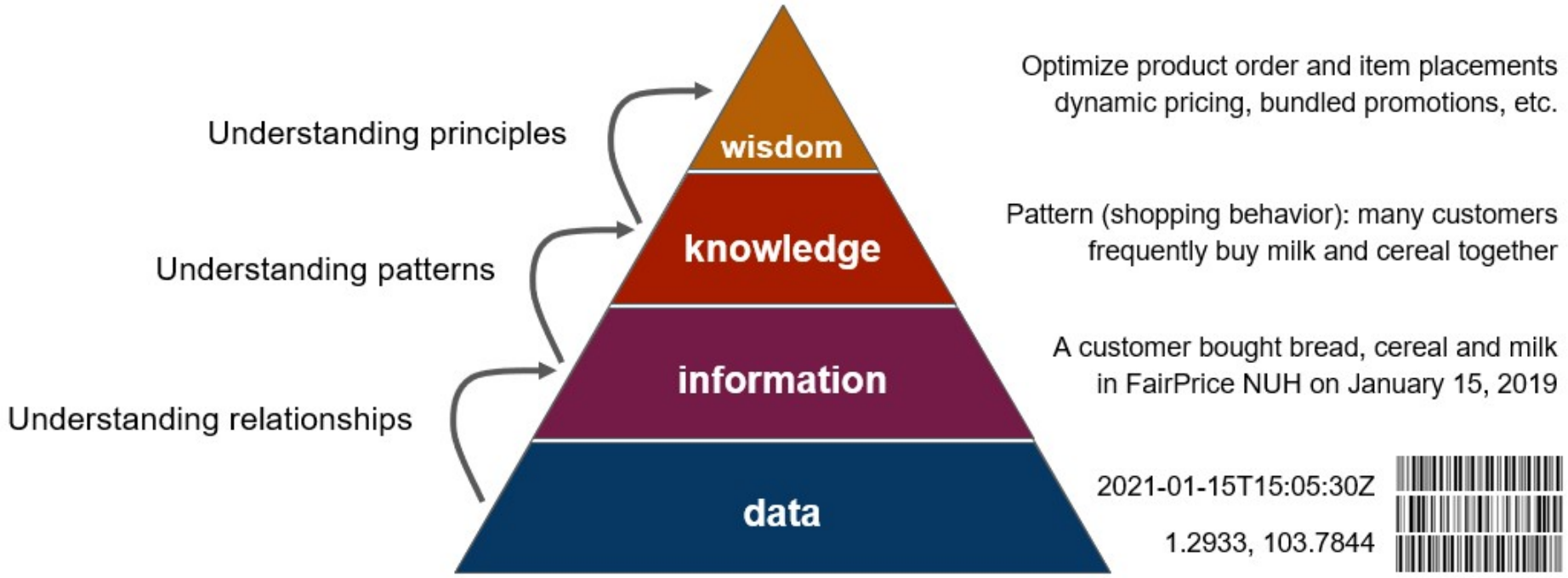
Week	Lecture	Tutorial	
7	Interim Exam		
8	Classification - 2	Decision Trees	
9	Unsupervised Learning		A2 released
10	Topics in ML	Clustering	
11	Linear Programming		AS2 Due
12	NUS Well-Being Day		
12	Final lecture		Milestone 2

Introduction to Data Analytics

Introduction

What is Data Analytics?

*The **non-trivial** extraction of implicit, previously unknown, and potentially **useful** information from data.* *(Frawley, Piatetsky-Shapiro, Matheus; 1991)*



What is *not* Data Analytics?

- Trivial patterns
  - Looking up a phone number in the directory.
  - Finding train stations on EWL from the station repository.
  - Calculating income tax based on the income.
- Not-useful patterns
  - Collecting click stream data from an e-commerce platform.
  - Querying a web search engine.

Spurious Analytics

- Patterns that do not generalise over the unseen data.
- Spurious correlations.
  - *Vegetarians miss fewer flights!*
  - *Smart people like curly fries!*
  - *Shark attacks increase when ice cream sales increase!*
- Finding patterns in the `dirty` data.

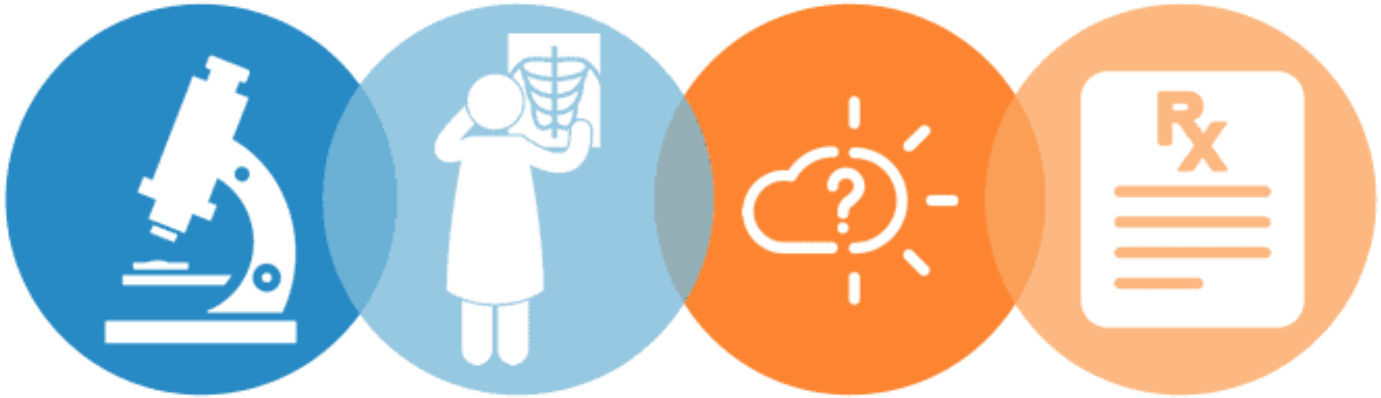
Introduction

Why to learn?

	Overall	Automotive, Aerospace, Supply Chain & Transport	Aviation, Travel & Tourism	Chemistry, Advanced Materials & Biotechnology	Consumer	Energy Utilities & Technologies	Financial Services & Investors	Global Health & Healthcare	Information & Communication Technologies	Infrastructure	Mining & Metals	Oil & Gas	Professional Services
User and entity big data analytics	85	84	89	79	85	85	86	87	93	65	62	87	85
App- and web-enabled markets	75	76	95	71	88	65	89	80	93	53	50	61	74
Internet of things	75	82	95	58	73	85	65	67	86	76	50	83	74
Machine learning	73	87	79	58	82	77	73	80	91	53	69	70	74
Cloud computing	72	76	79	67	67	73	65	73	91	71	62	78	76
Digital trade	59	68	68	62	82	58	70	53	70	47	50	57	59
Augmented and virtual reality	58	71	68	50	48	65	59	67	72	59	62	65	53
Encryption	54	58	53	25	42	38	73	67	67	41	25	57	53
New materials	52	71	32	79	79	65	22	60	30	82	62	83	41
Wearable electronics	46	61	53	46	45	42	49	73	49	24	25	70	35
Distributed ledger (blockchain)	45	32	37	29	39	54	73	67	67	18	38	48	50
3D printing	41	61	21	58	42	54	19	53	35	41	50	57	29
Autonomous transport	40	74	58	54	39	46	16	20	44	41	50	30	41
Stationary robots	37	53	37	50	42	35	27	47	35	35	38	52	29
Quantum computing	36	29	32	25	33	46	43	33	44	24	19	43	41
Non-humanoid land robots	33	42	26	21	36	27	32	40	37	29	25	30	24
Biotechnology	28	18	0	42	52	42	11	87	23	12	44	39	24
Humanoid robots	23	29	26	17	18	8	35	13	33	12	25	13	24
Aerial and underwater robots	19	18	16	17	12	35	5	0	19	29	25	52	21

Source: Future of Jobs Survey 2018, World Economic Forum.

Introduction



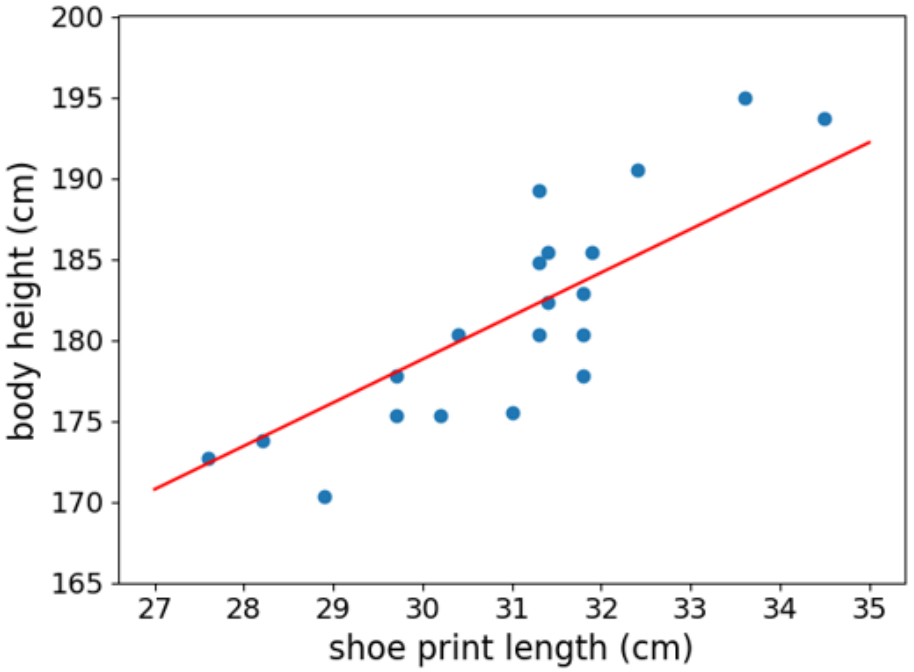
Retail Analytics

- Distribution of sales of various items in a store.
- Why was the sale of a particular item lower in Week 2?
- Estimate turnover on loyal customer discount campaign.
- Which campaign would yield the maximum profit?

Data analytics models

Regression

- **Input.** Labeled dataset with multiple features (or attributes).
- **Pattern.** `Numerical` value as a function of other features.



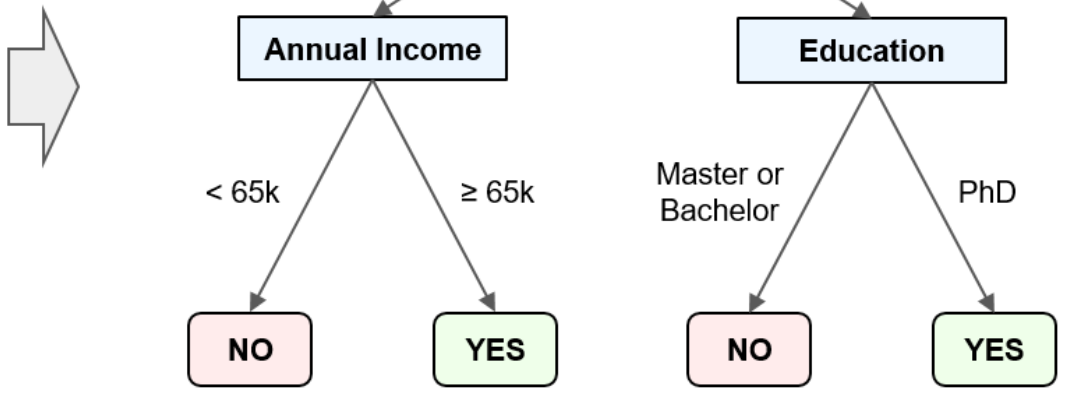
What is the expected height of a person that leaves a shoe print of size 32cm?

Data analytics models

Classification

- **Input.** Labeled dataset with multiple features (or attributes).
- **Pattern.** `Categorical` value as a function of other features.

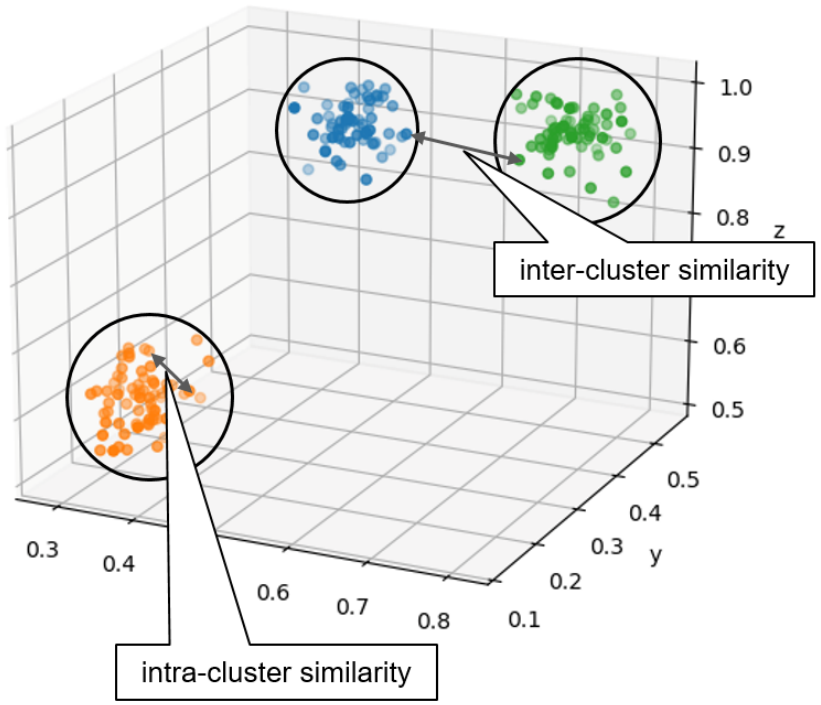
Age	Edu-cation	Marital Status	Annual Income	Credit Default
23	Masters	Single	75k	Yes
35	Bachelor	Married	50k	No
26	Masters	Single	70k	Yes
41	PhD	Single	95k	Yes
18	Bachelor	Single	40k	No
55	Master	Married	85k	No
30	Bachelor	Single	60k	No
35	PhD	Married	60k	Yes
28	PhD	Married	65k	Yes



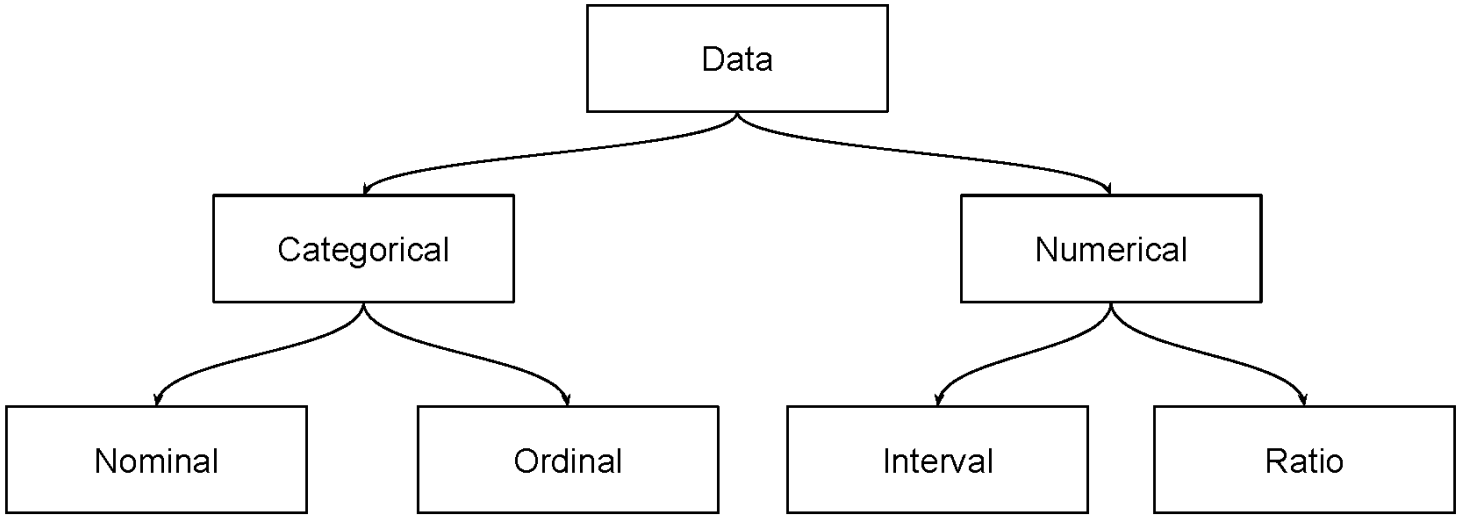
Data analytics models

Clustering

- **Input.** Dataset with multiple features (or attributes).
- **Pattern.** *Clusters* of data based on specified `similarity` notion.



Types of Data



- Values are labels
- Operations: =, ≠
- Examples: gender, colors

- Values are labels with order
- Operations: =, ≠, <, >
- Examples: grades, degree

- Values are measurements with meaningful distance
- Operations: =, ≠, <, >, +, -
- Examples: Temperature, latitude

- Values are measurements with meaningful ratio
- Operations: =, ≠, <, >, +, -, /
- Examples: age, salary

Types of Data

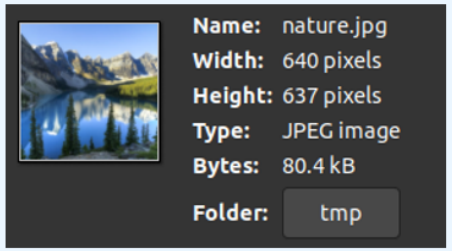
(Well-)Structured Data

- Highly organized: adheres to predefined data model
- Each object has the same fixed set of attributes
- Easy to search, aggregate, manipulate, analyze data
- Examples: Relational databases, spreadsheets

Age	Edu-cation	Marital Status	Income Level	Credit Approval
23	Masters	Single	Mid	No
35	College	Married	High	Yes
26	Masters	Single	High	No
41	PhD	Single	Mid	Yes
18	Poly	Single	Low	No
55	Poly	Married	High	Yes
....	...	...	...	...

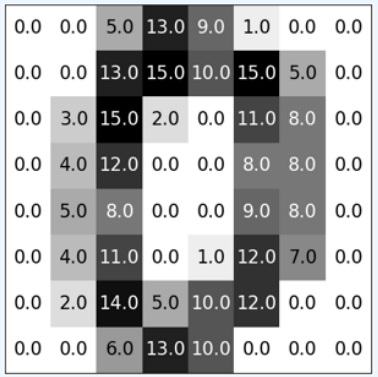
Semi-Structured Data

- No rigid data model: mix of structured & unstructured data
- Data exchange formats: XML, JSON, CSV
- Tagged unstructured data (e.g., photo + date/time, location, exposure, resolution, flash, etc)



Unstructured Data

- No fixed data model
- Requires more advanced data analysis techniques
- Examples: images, videos, audio, text, social media



Types of Data

Collection of records

Age	Education	Married	Income	Credit Approval
23	Masters	No	75k	Yes
35	Bachelors	Yes	50k	No
26	Masters	No	70k	Yes
41	PhD	No	95k	Yes
55	Masters	Yes	80k	No

Types of Data

Collection of transactions

ID	Items
1	covid-19, anosmia, cough, fatigue
2	flu, anosmia, headache
3	covid-19, anosmia, headache, fatigue, fever
4	flu, depression, fatigue

Types of Data

Graph data

https://www.lta.gov.sg/

Time series data

https://www.sg.finance.yahoo.com/

Data Visualisation

Motivation

- Anscombe's Quartet
- Is it always necessary?
- The rule of thumb
- Standard visualisations

Set I		Set II		Set III		Set IV	
X	Y	X	Y	X	Y	X	Y
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
12.0	10.84	12.0	9.13	12.0	8.15	8.0	5.56
7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89

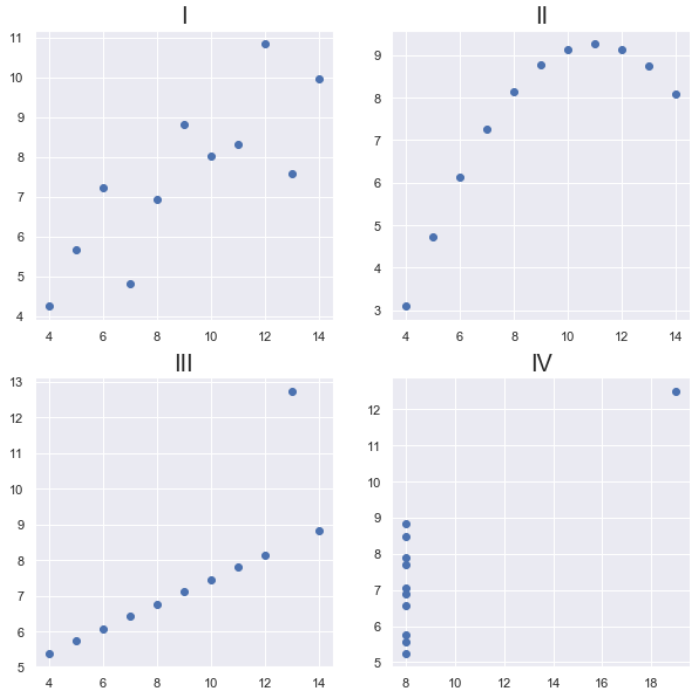
Descriptive statistics

- $\mu_X = 9.0$
- $\mu_Y = 7.5$
- $\rho_{XY} = 0.82$
- $Y = 3 + 0.5X$

Motivation

- **Anscombe's Quartet**
- Is it always necessary?
- The rule of thumb
- Standard visualisations

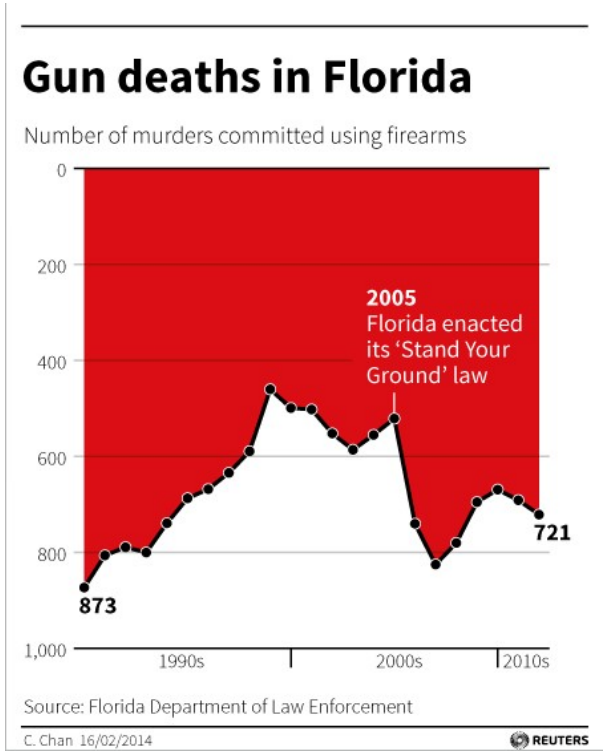
Set I		Set II		Set III		Set IV	
X	Y	X	Y	X	Y	X	Y
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
12.0	10.84	12.0	9.13	12.0	8.15	8.0	5.56
7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89

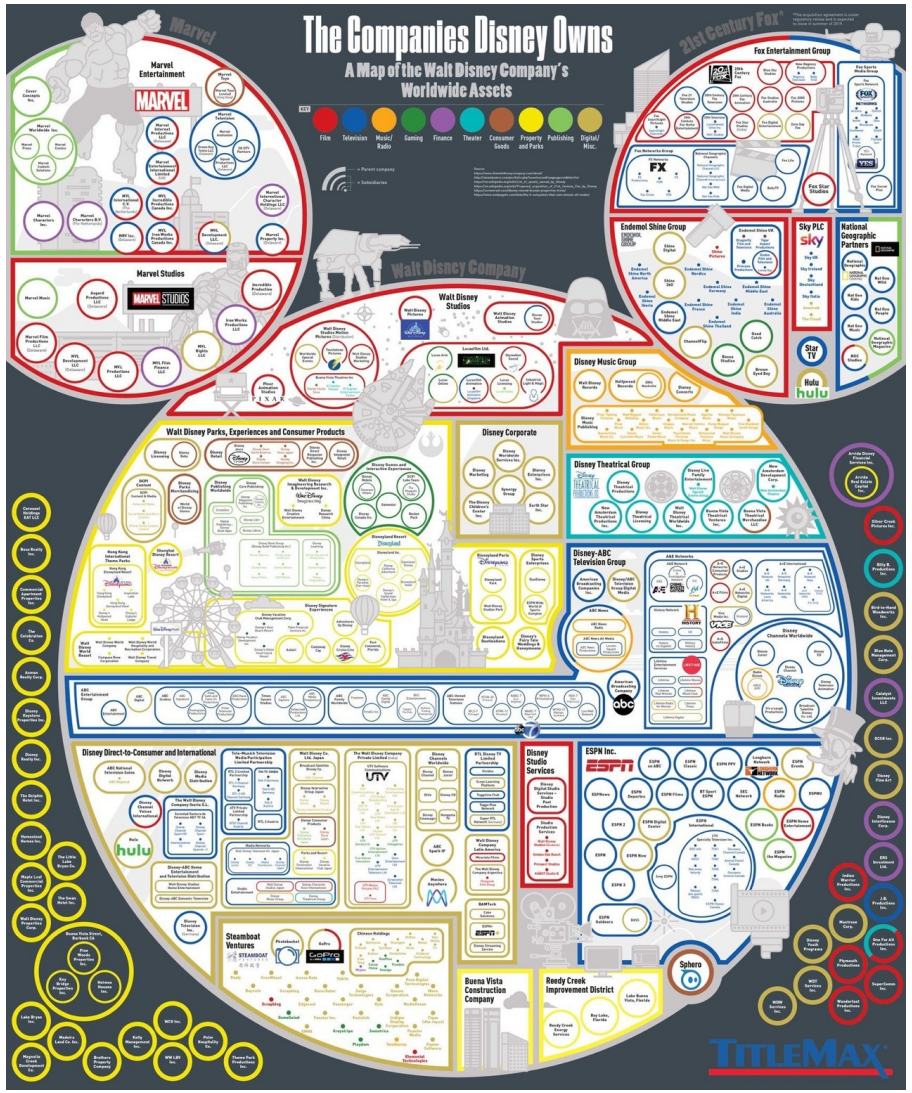


Motivation

Anscombe's Quartet
› Is it always necessary?
The rule of thumb
Standard visualisations

Is picture always *worth* a thousand words?

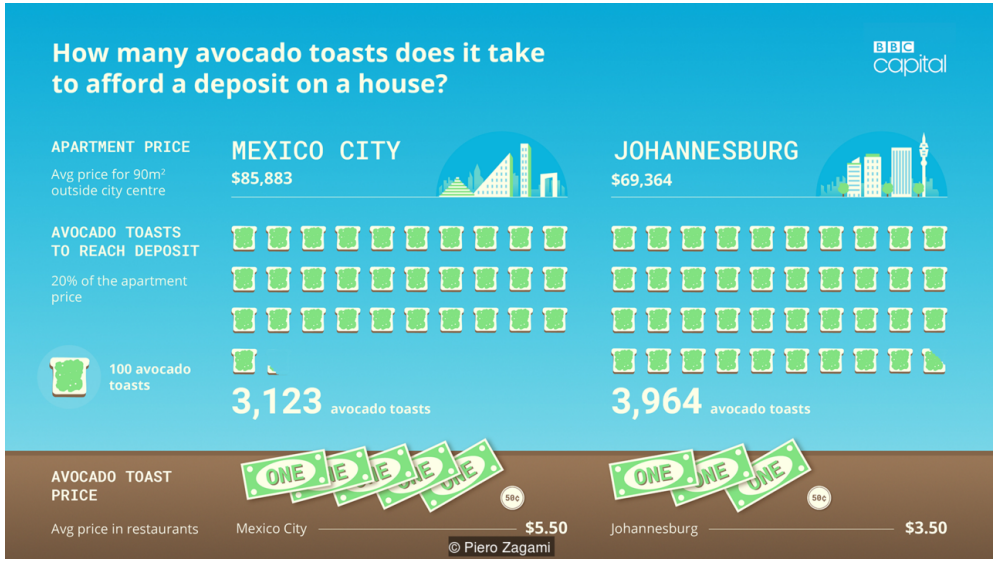


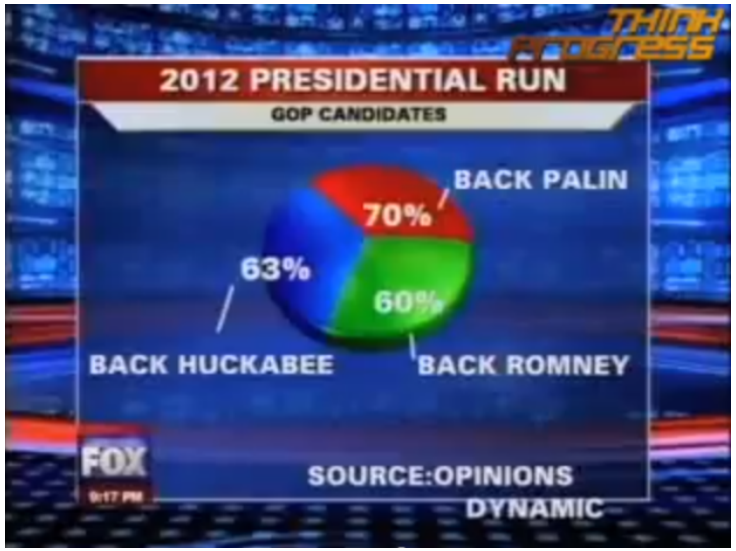


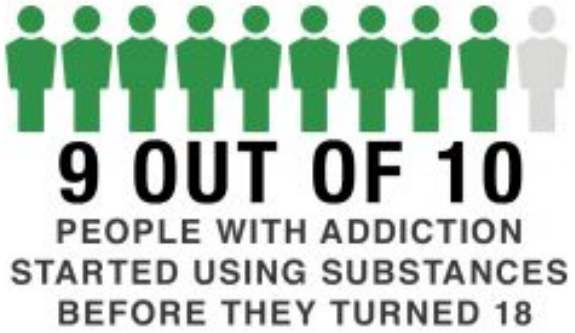
Motivation

Anscombe's Quartet
› Is it always necessary?
The rule of thumb
Standard visualisations

Is picture always *worth* a thousand words?







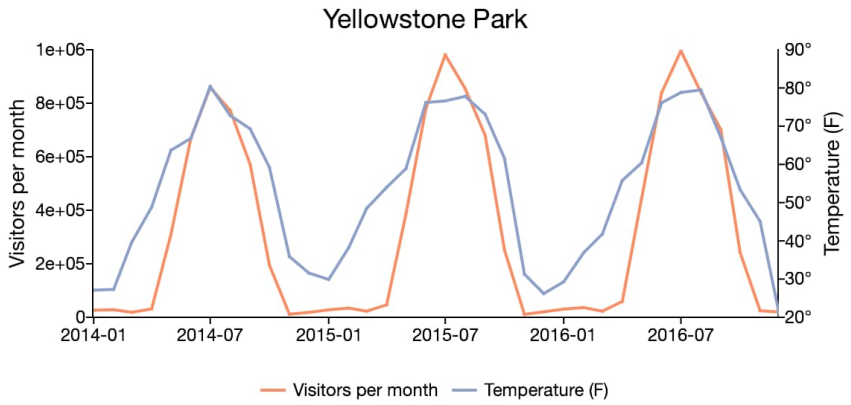
The rule of thumb

Table versus Plot

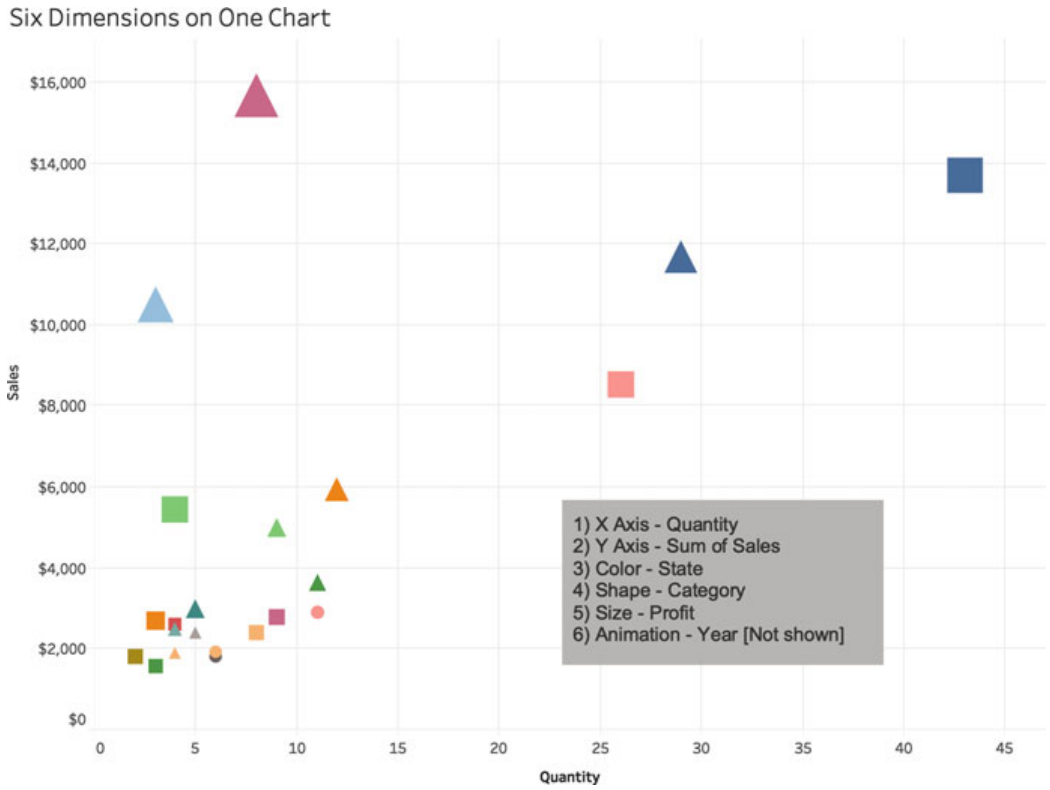
Prefer a table over a chart if

- You are interested in the **values**.
- You want to compare **individuals rather than a trend**.
- You require a **precision**.
- Your data comprises of **multiple units of measurement**.

Class	Range
A	[85, 100]
B	[72, 85]
C	[60, 72]



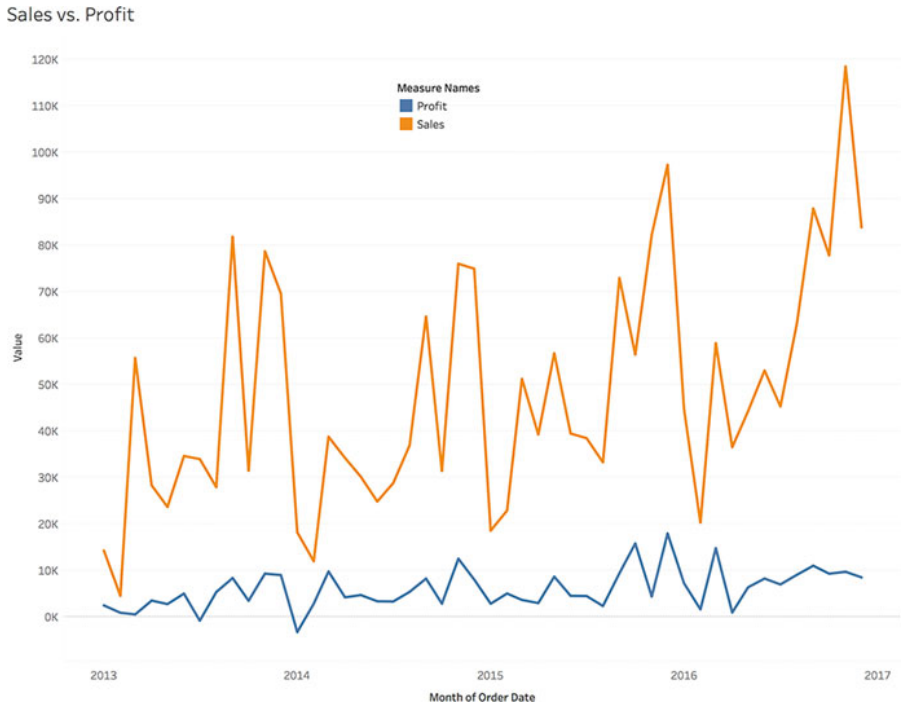
The rule of thumb

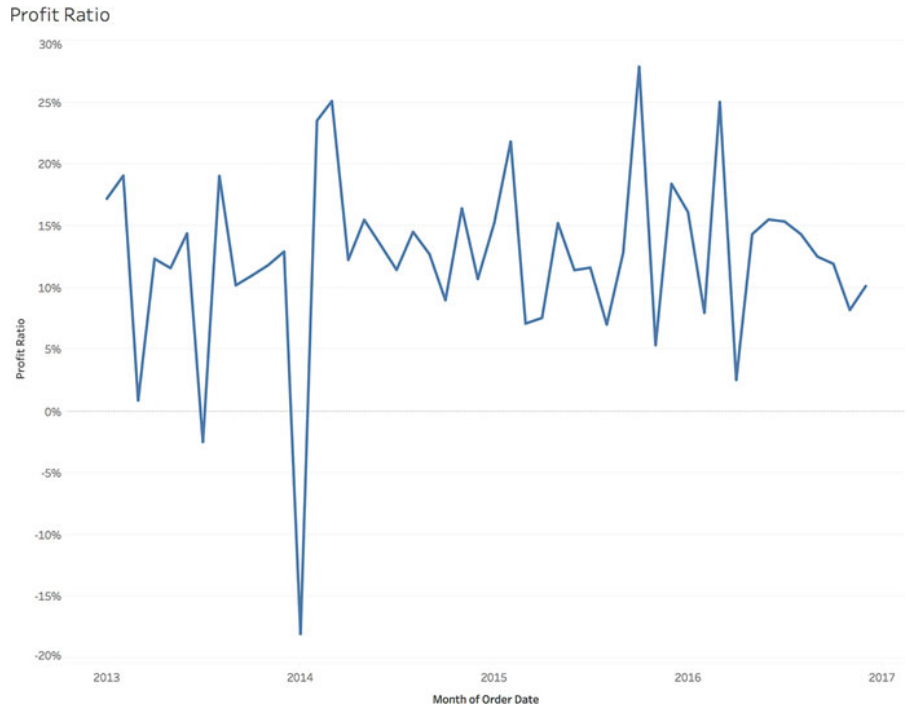


Do not use more than **three** dimensions inside a chart!

The rule of thumb

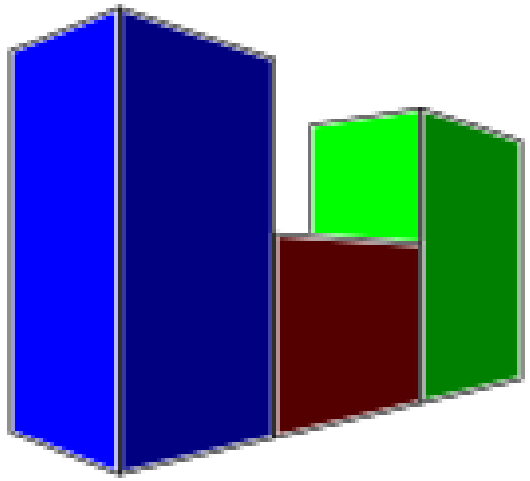
Do not rely on the readers to interpret the intended information that you want to convey!

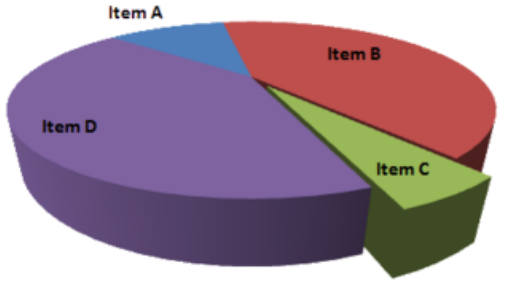


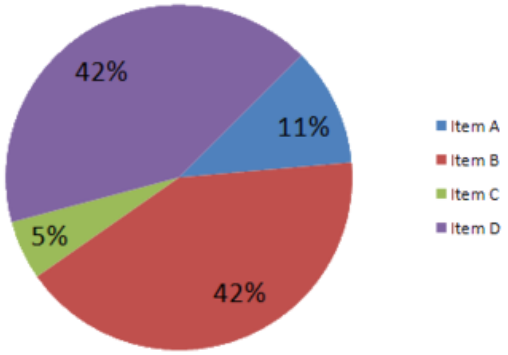


The rule of thumb

A 3D plot on the 2D paper distorts the dimension and creates a false image for the viewers.



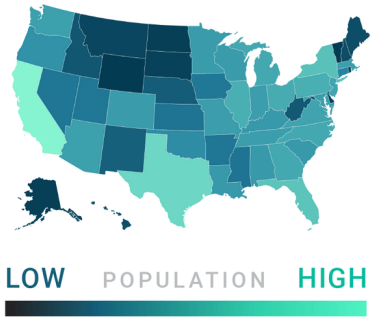




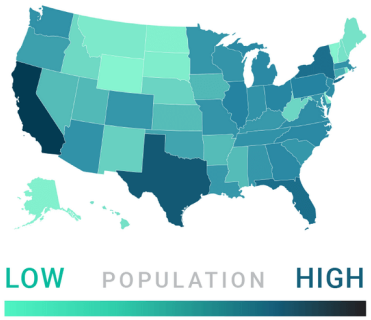
The rule of thumb

Do not play with the intuition.

NOT IDEAL



BETTER



Ensure the use of right contrast.

CONTRAST

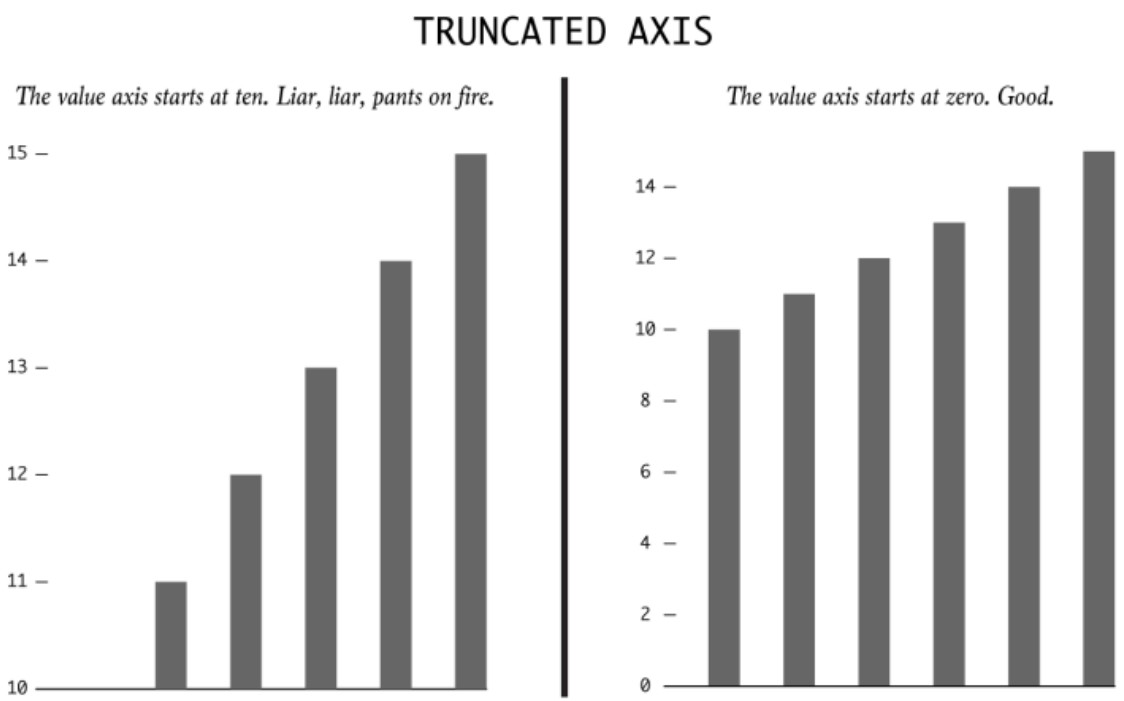
CONTRAST RATIOS

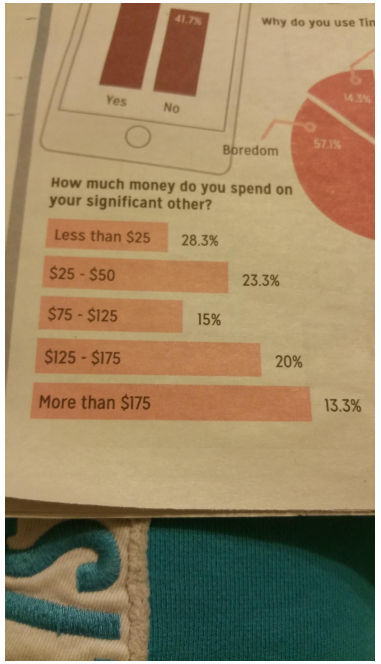
1.0				
1.1	Choose if you dislike readers.	That's bad.	That's bad.	Horrible.
1.5	Ok in 1% of the cases.	Not ideal.	That's bad.	My eyes!
2.5	Can be a good choice.	Ok.	Not ideal.	That's bad.
4.5	Safe choice.	Great.	Ok.	Not ideal.

https://www.dataquest.io/blog/what-to-consider-when-choosing-colors-for-data-visualization/

The rule of thumb

Do not lie!

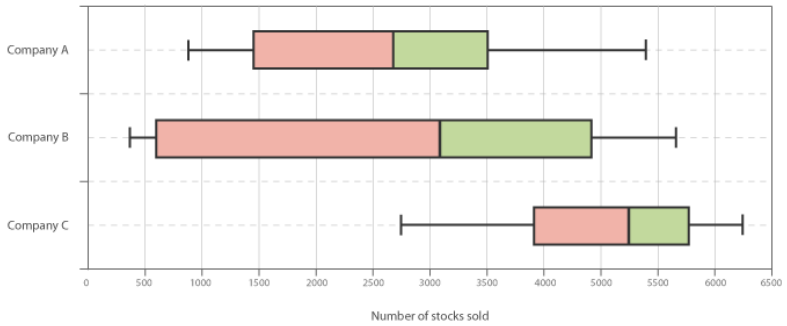




Popular visualisations

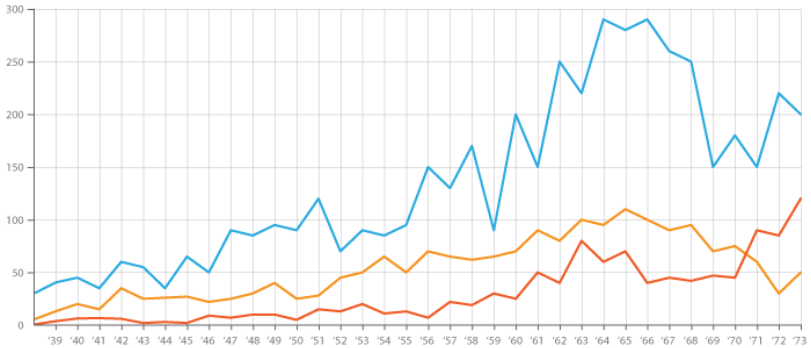
Box plot.

To visualise statistical tendancies



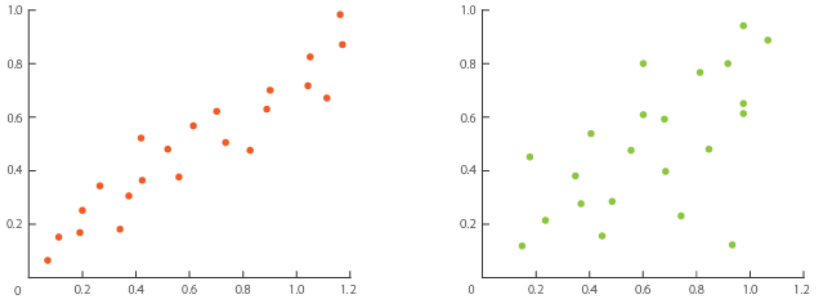
Line plot.

To visualise trends in the ordered data



Scatter plot.

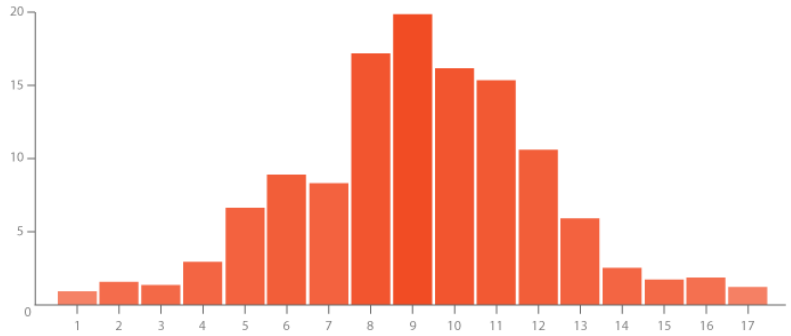
To visualise relationship between two features



Popular visualisations

Bar plot.

To compare different categories



Pie chart.

To compare parts to whole

Word cloud.

To compare statistics of the words in a *bag*

Summary

Summary

Module logistics

- **Lecture.** Every Friday 6.30pm to 8.30pm
- **Tutorial.** Every Thursday 9.30am to 10.30am
- **Interim Exam.** 6^{th} Oct at 7pm
- **Final Exam.** 4^{th} Dec at 5pm

Data analytics

- Defining analytics
- Scope of analytics
- Commonly used models
- Types of data

Data visualisation

Project Teams.

Form the project teams of four members by the end of next week!

Thank you!

Feel free to reach out to me at

dcsashi (at) nus (dot) edu (dot) sg