DATA SCIENCE IN MANUFACTURING
WEEK 4

ANDREW SHERLOCK, JONATHAN CORNEY, DANAI KORRE

LECTURE: WEEK 4

Exploratory Data Analysis
And Data Visualization

BY THE END OF THIS LECTURE YOU SHOULD:

- Understand exploratory data analysis and techniques
- Learn how data visualisation is used for manufacturing data
- Understand why data visualisation is important
- Become familiar with common types of data visualisation

EXPLORATORY DATA ANALYSIS (EDA)

refers to the critical process of performing initial investigations on data so as to discover patterns, to spot anomalies, to test hypothesis and to check assumptions with the help of summary statistics and graphical representations [1].

EXPLORATORY DATA ANALYSIS (EDA)

EDA is usually more important for observational or found data, rather than for data produced by experiments that have been specifically designed to test a predetermined hypothesis.

The term exploratory data analysis was developed by its founder John W. Tukey in the 70s who argued that data analysis ought to be seen as a science in its own right.

EDA prioritises visualisation as best way to generate insight about data, because it can combine detail (every data point might be plotted) with a summary of the relationship of any observation points.

[2]

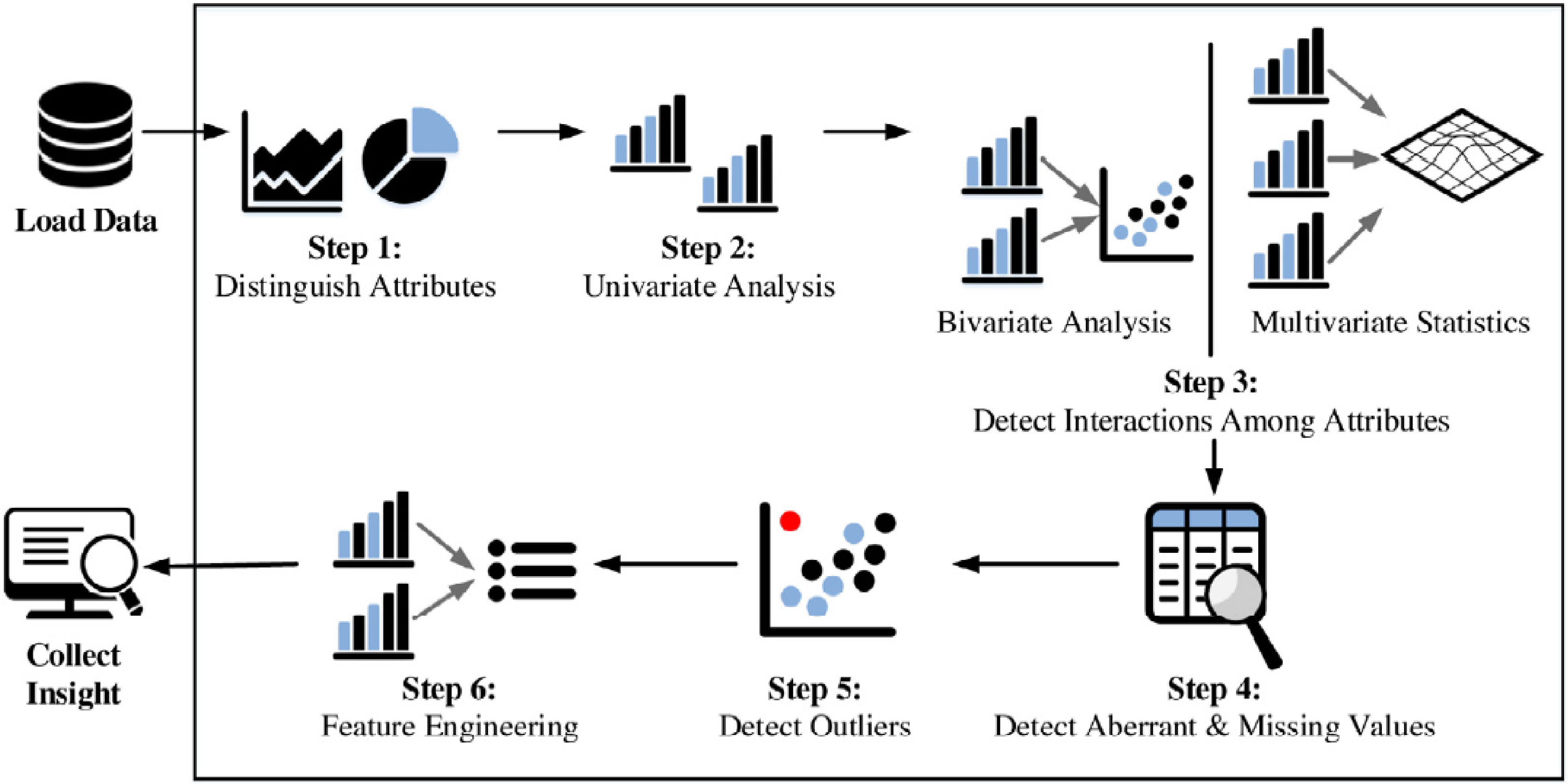
THE GOALS OF THE EDA PROCESS

A proper EDA hopes to accomplish several goals:

- To question the data and determine if there are problems inherent in the dataset;
- To determine if the data on hand is sufficient to answer a particular research question or whether additional feature engineering is required;
- To develop a framework for answering the research question;
- To refine the questions and/or research problem based on what you have learned about the data.

[3]

FUNDAMENTAL STEPS OF EDA PROCESS



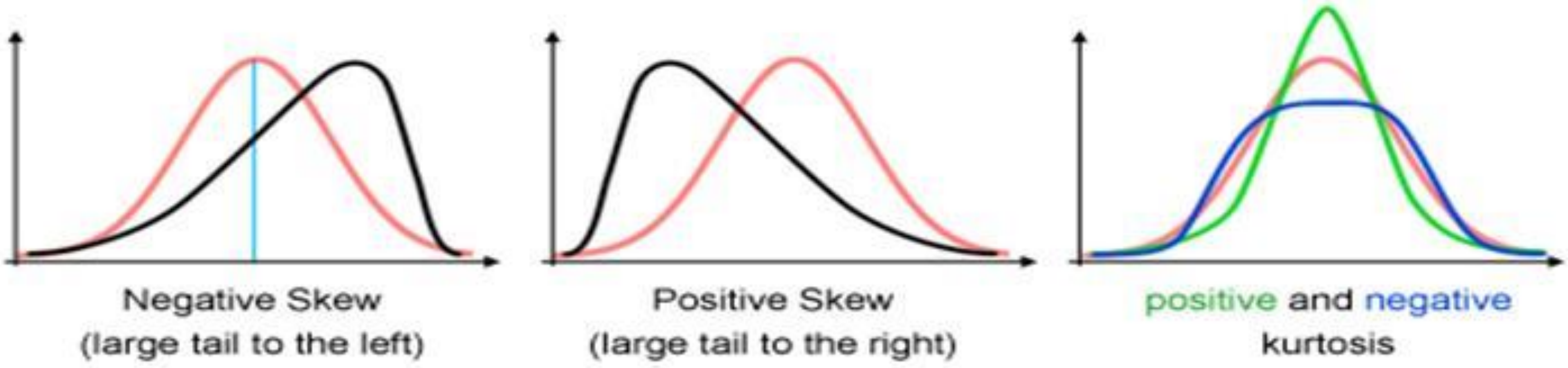
Source: Ghosh et al. 2018

	Univariate	Multivariate

Graphical	• **Quantitative Variable**:
	  • Histogram
	  • Boxplots
	  • Normal QQ-plot
	• **Categorical Variable**: Bar Charts
	• Time data – Line Plot
		• **One Categorical and One Quantitative Variable:**
		  • Side by side Boxplots
		• **Two or More Categorical Variables:**
		  • Grouped Bar Chart
		• **Two or More Quantitative Variables:**
		  • Scatterplot
		  • Correlation Heatmap
		  • Pairplot
		• Missing Data Detection

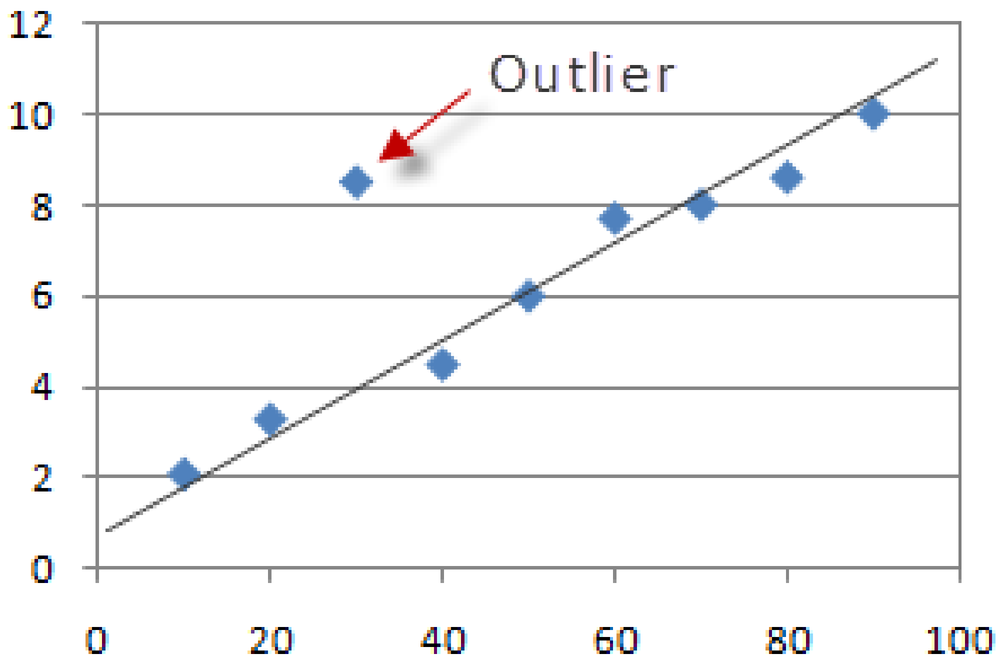
Non-Graphical	• **Categorical Variable**: tabular representation of frequency (or relative frequency)
	• **Quantitative Variable**:
	  • Location (mean, median)
	  • Spread (IQR. Std dev, range)
	  • Modality (mode)
	  • Shape (skewness, kurtosis)
	  • Outliers
	• Missing Data Detection
	
	[3]
		• **One Categorical and One Quantitative Variable:** *standard univariate nongraphical statistics for the quantitative variables separately for each level of the categorical variable.*
		  • Mean
		  • Median
		  • Range and Spread measures
		• **Two or More Categorical Variables:**
		  • Correlation
		  • Covariance
		  • Descriptive stat per each variable
		• Missing Data Detection

SKEWNESS AND KURTOSIS



Illustrating skewness and kurtosis in a distribution. Source: Sharma 2017.

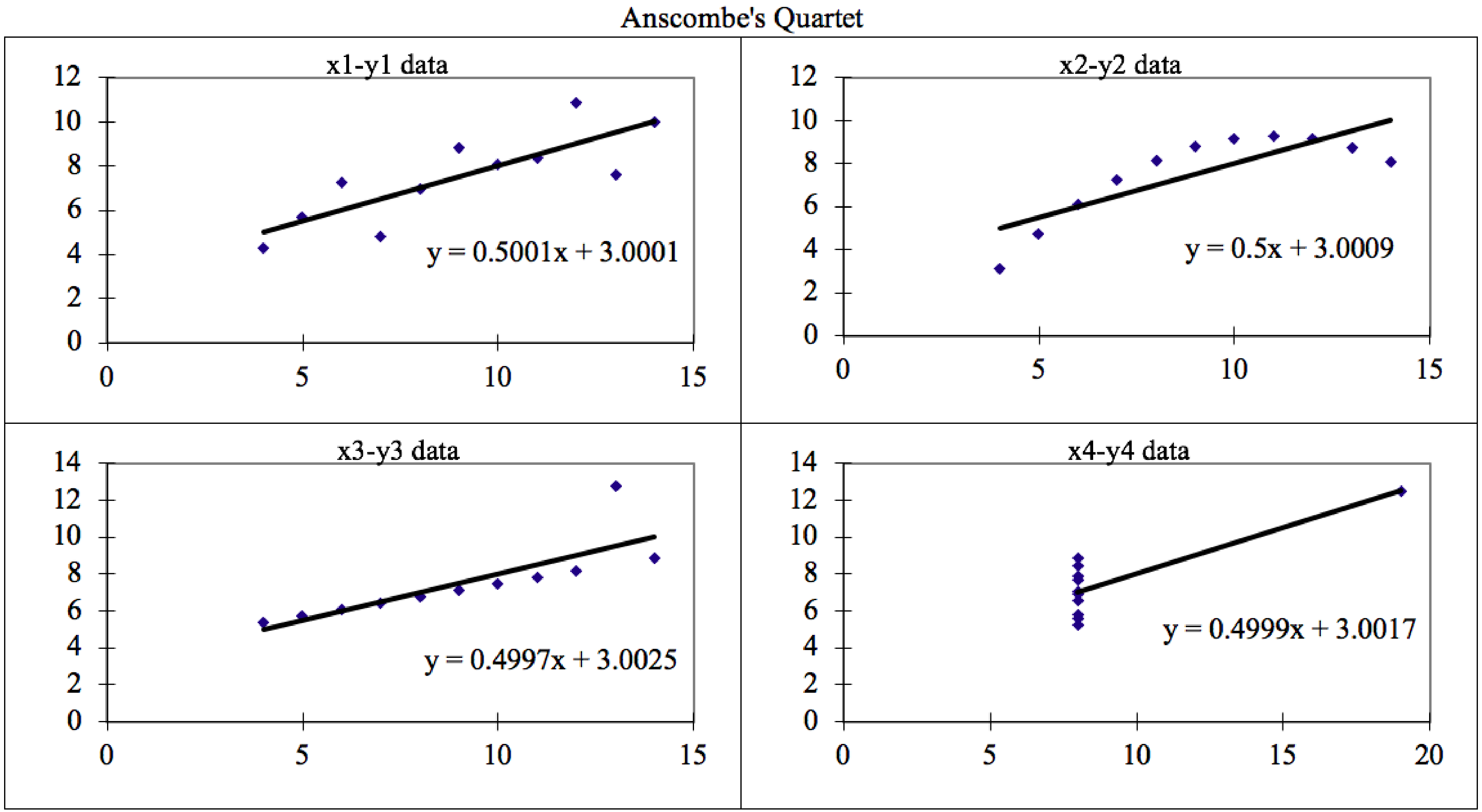
OUTLIERS



Outlier example in linear regression. Source: Math Open Reference 2011

This is often taken as a sign that the data point may actually be an error.

THE ANSCOMBE'S QUARTET



The Anscombe's quartet. Source: Gupta 2020

FUNDAMENTAL STEPS OF EDA PROCESS

- Structure of the data: number of data points, number of features, feature names, data types, etc.
- Check for consistency across datasets.
- Identify what data signifies (called measures) for each of data points and be mindful while obtaining metrics.
- Calculate key metrics for each data point (summary analysis):
  - Measures of central tendency (Mean, Median, Mode);
  - Measures of dispersion (Range, Quartile Deviation, Mean Deviation, Standard Deviation);
  - Measures of skewness and kurtosis.

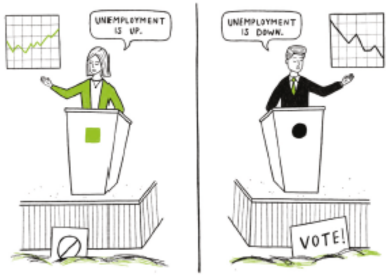
[4]

FUNDAMENTAL STEPS OF EDA PROCESS

- Investigate visuals:
  - Histogram for each variable;
  - Scatterplot to correlate variables.
- Calculate metrics and visuals per category for categorical variables (nominal, ordinal).
- Identify outliers and mark them. Based on context, either discard outliers or analyse them separately.
- Estimate missing points using *data imputation techniques (for industrial databases see Lakshminarayan, Harp and Samad, 1999)*.

[4]

STATISTICAL FALLACIES

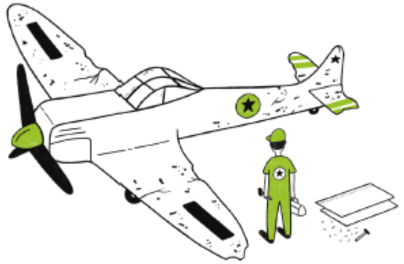


CHERRY PICKING

Selecting results that fit your claim and excluding those that don't.

DATA DREDGING

Repeatedly testing new hypotheses against the same set of data, failing to acknowledge that most correlations will be the result of chance.

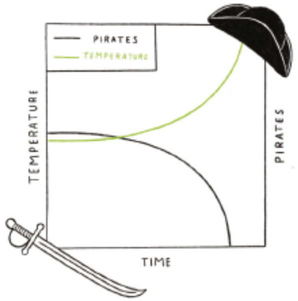


SURVIVORSHIP BIAS

Drawing conclusions from an incomplete set of data, because that data has 'survived' some selection criteria.

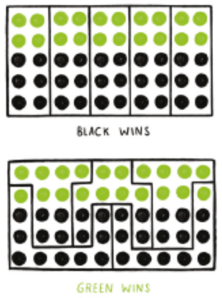
COBRA EFFECT

Setting an incentive that accidentally produces the opposite result to the one intended. Also known as a Perverse Incentive.



FALSE CAUSALITY

Falsely assuming when two events appear related that one must have caused the other.



GERRYMANDERING

Manipulating the geographical boundaries used to group data in order to change the result.

THE UNIVERSITY of EDINBURGH
School of Engineering

Source: geckoboard.com

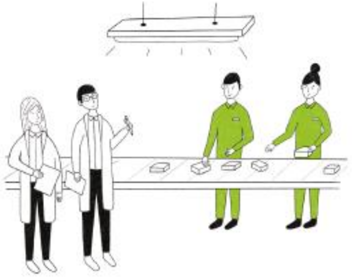
STATISTICAL FALLACIES

SAMPLING BIAS

Drawing conclusions from a set of data that isn't representative of the population you're trying to understand.

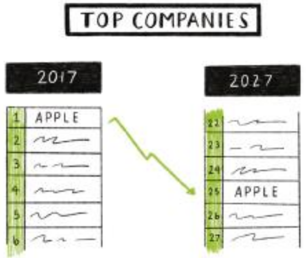
GAMBLER'S FALLACY

Mistakenly believing that because something has happened more frequently than usual, it's now less likely to happen in future (and vice versa).



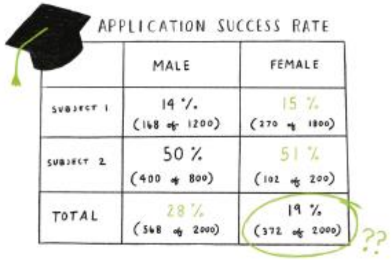
HAWTHORNE EFFECT

The act of monitoring someone can affect their behaviour, leading to spurious findings. Also known as the Observer Effect.



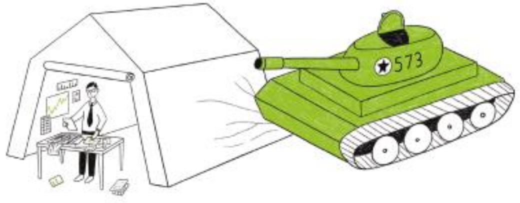
REGRESSION TOWARDS THE MEAN

When something happens that's unusually good or bad, it will revert back towards the average over time.



SIMPSON'S PARADOX

When a trend appears in different subsets of data but disappears or reverses when the groups are combined.



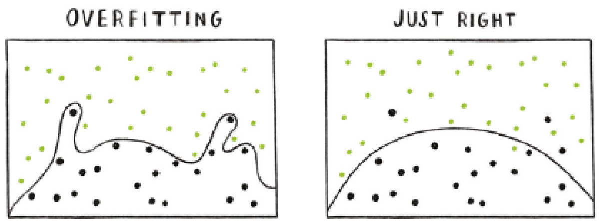
MCNAMARA FALLACY

Relying solely on metrics in complex situations and losing sight of the bigger picture.

THE UNIVERSITY of EDINBURGH
School of Engineering

Source: geckoboard.com

STATISTICAL FALLACIES

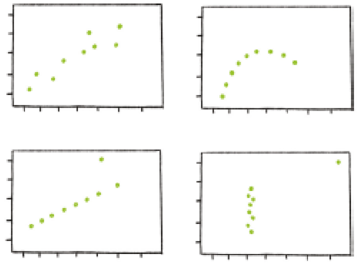


OVERFITTING

Creating a model that's overly tailored to the data you have and not representative of the general trend.

PUBLICATION BIAS

Interesting research findings are more likely to be published, distorting our impression of reality.



DANGER OF SUMMARY METRICS

Only looking at summary metrics and missing big differences in the raw data.

Source: geckoboard.com

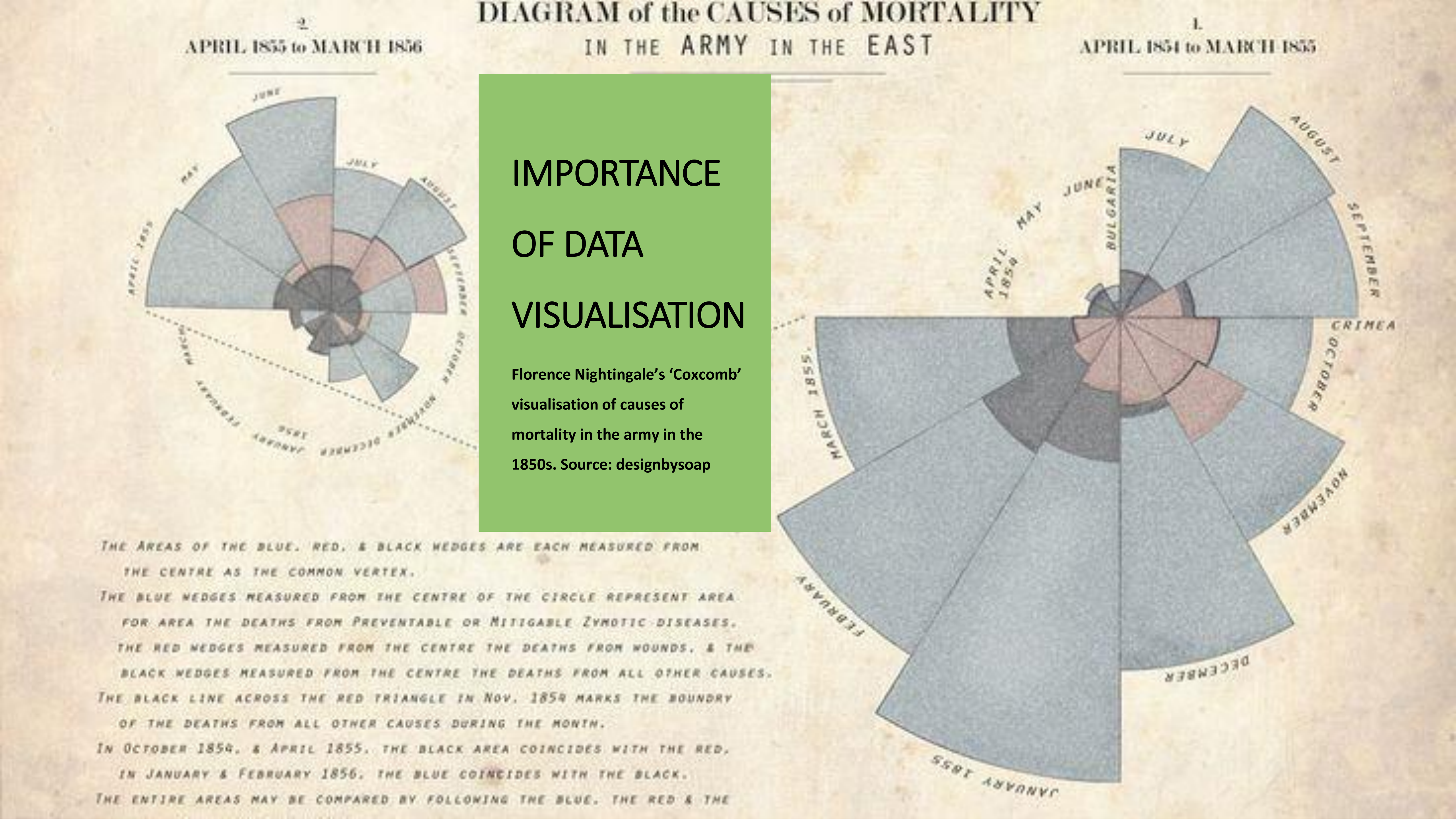
WHAT IS DATA VISUALISATION

Data visualization is the process of translating large data sets and metrics into charts, graphs and other visuals. The resulting visual representation of data makes it easier to identify and share real-time trends, outliers, and new insights about the information represented in the data (IBM).

IMPORTANCE OF DATA VISUALISATION

John Snow's dot map showing locations of cholera cases. Source: Friendly and Denis 2001, 1850+: Dot map of disease.

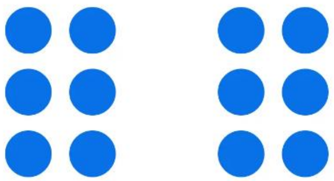
THE UNIVERSITY of EDINBURGH
School of Engineering



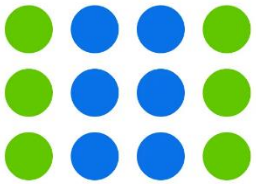
IMPORTANCE OF DATA VISUALISATION

Florence Nightingale's 'Coxcomb' visualisation of causes of mortality in the army in the 1850s. Source: designbysoap

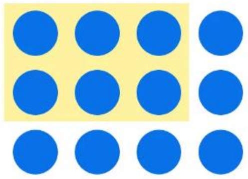
GESTALT PSYCHOLOGY



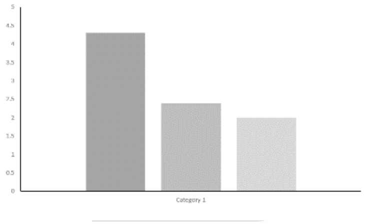
PROXIMITY



SIMILARITY



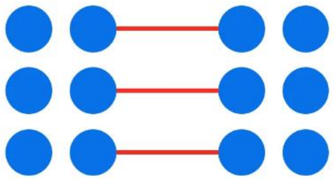
ENCLOSURE



CLOSURE

CONTINUITY

Gestalt theory emphasizes that the whole of anything is greater than its parts. That is, the attributes of the whole are not deducible from analysis of the parts in isolation. [5]

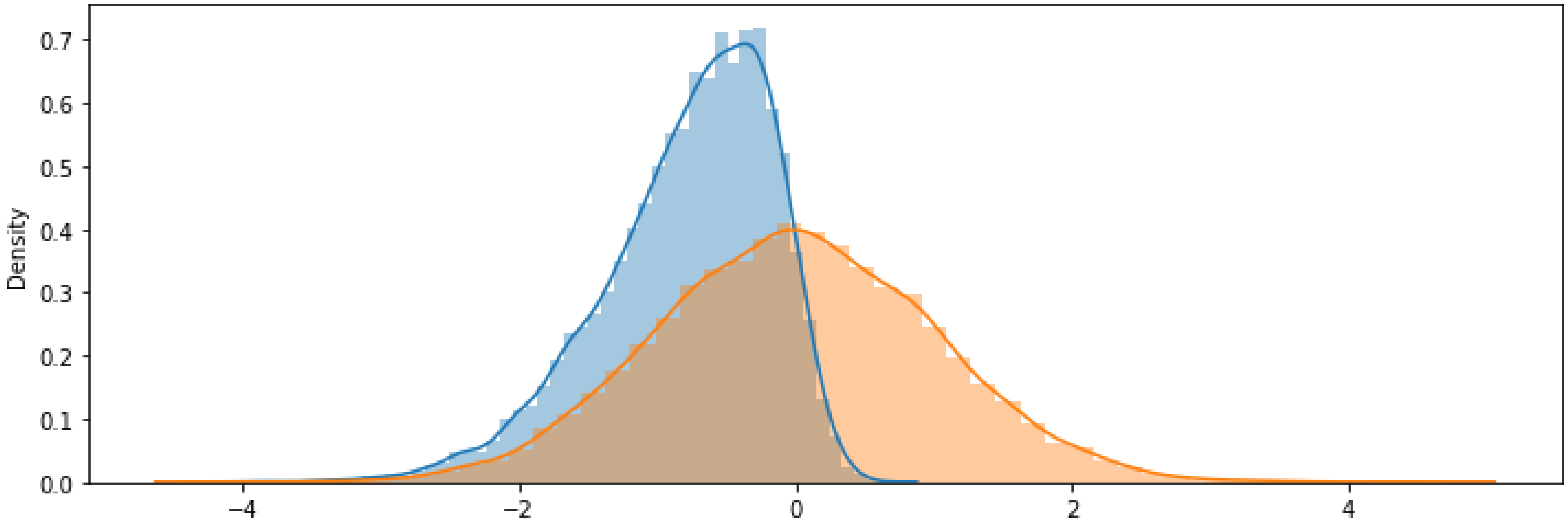


CONNECTION

EXPLORATORY DATA VISUALISATION

Exploratory graphics are used for looking for results. Many may be used, and they should be fast and informative. They are not intended for presentation, so that detailed legends and captions are unnecessary [6]. Often used during the data carpentry stage of data science lifecycle.

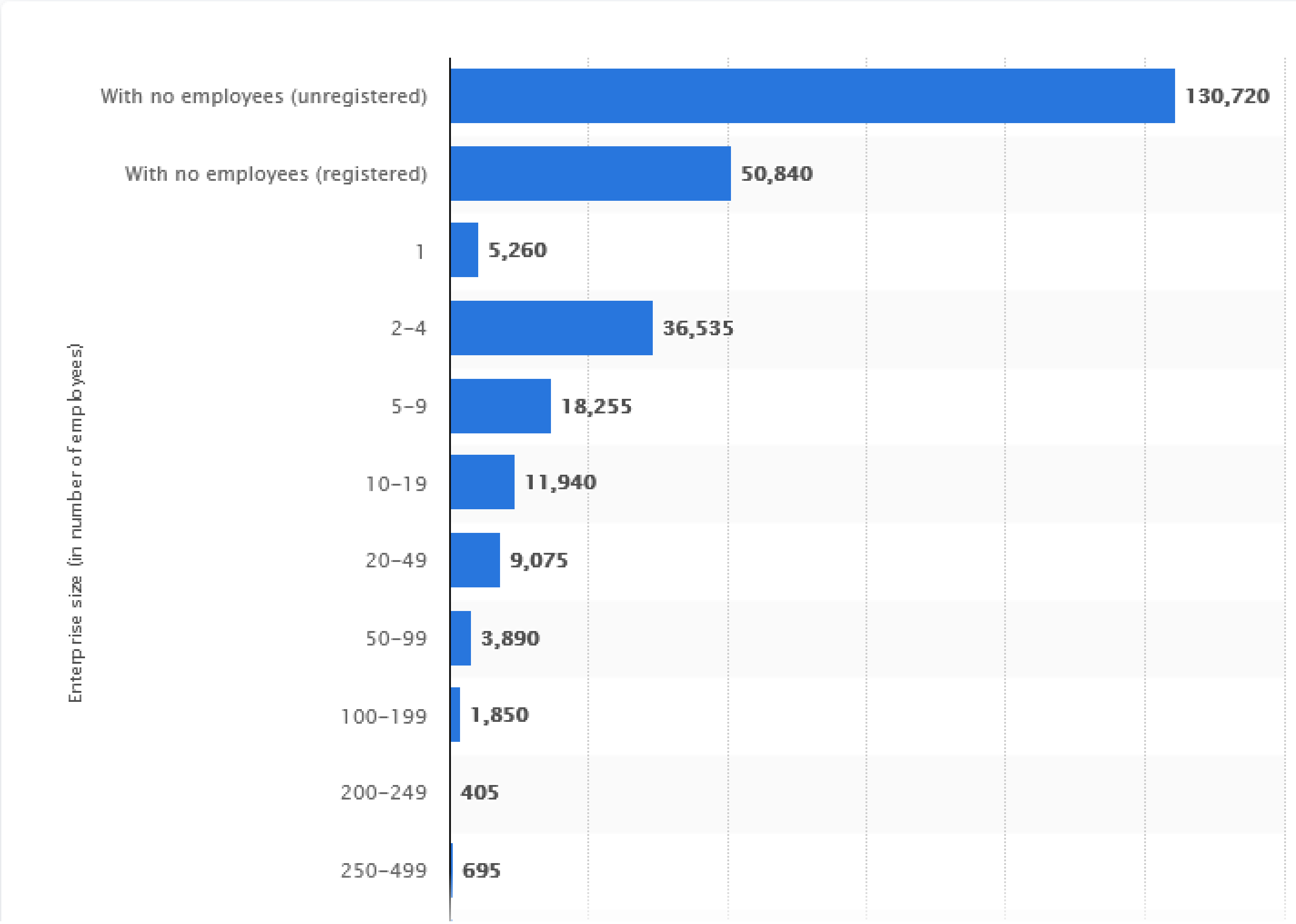
EYEBALLING DATASET DISTRIBUTIONS WITH HISTOGRAMS



EXPLANATORY DATA VISUALISATION

The differences between graphics for presentation and graphics for exploration lie in both form and practice. Explanatory/presentation graphics are generally static, and a single graphic is drawn to summarize the information to be presented. These displays should be of high quality and include complete definitions and explanations of the variables shown and of the form of the graphic. They may give no hint as to how a result was reached, but they should offer convincing support for its conclusion [6].

NUMBER OF BUSINESS ENTERPRISES IN THE MANUFACTURING SECTOR IN THE UNITED KINGDOM IN 2021, BY ENTERPRISE SIZE



THE UNIVERSITY of EDINBURGH
School of Engineering

Source: Statista

COMMON TYPES OF PLOTS

Standard chart graphics

- Error plots
- Histogram
- Heat map
- Scatter plot
- Pie chart
- Line chart
- Area chart
- Gantt chart
- Bar chart

Spatial plots and maps

- Point map
- Choropleth
- Raster surface

Topology structures

- Scatter plot matrix
- Linear topology
- Raster surface
- Tree network topology
- Graph models

Various charts to aid exploratory data analysis. Source: Grosser 2018

CONSIDERATIONS

- Accuracy and consistency
- Bias
  - Terms
  - Semantics
- Context

- Objectivity
- Many visualizations, different angles and perspectives
- Generate hypotheses
- Ethics

Source: Cole Nussbaumer Knaflic

CONSIDERATIONS

- Abbreviated Axes
- Dualling Data
- Confusing Charts
- Choropleth Colouring
- Horrible Histograms

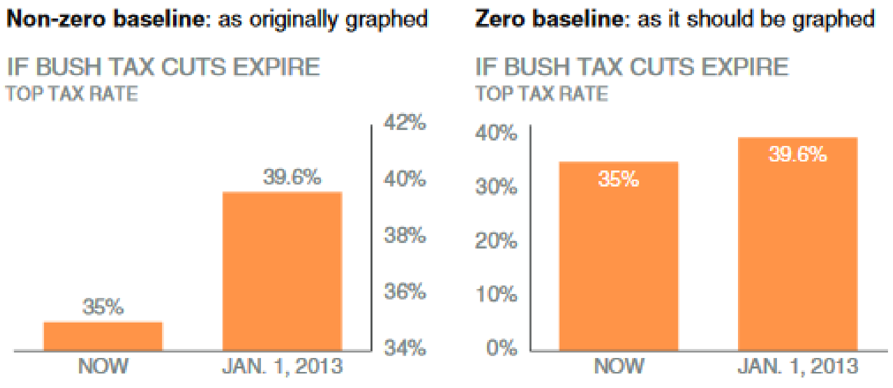
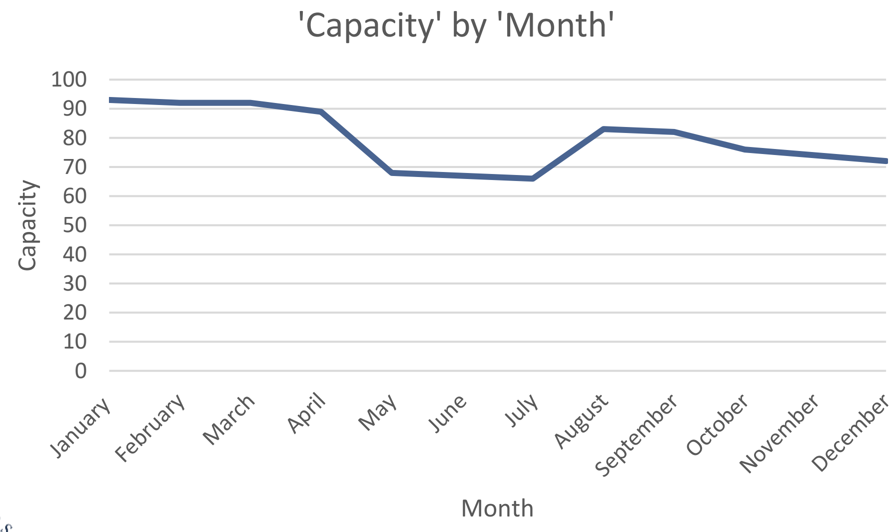
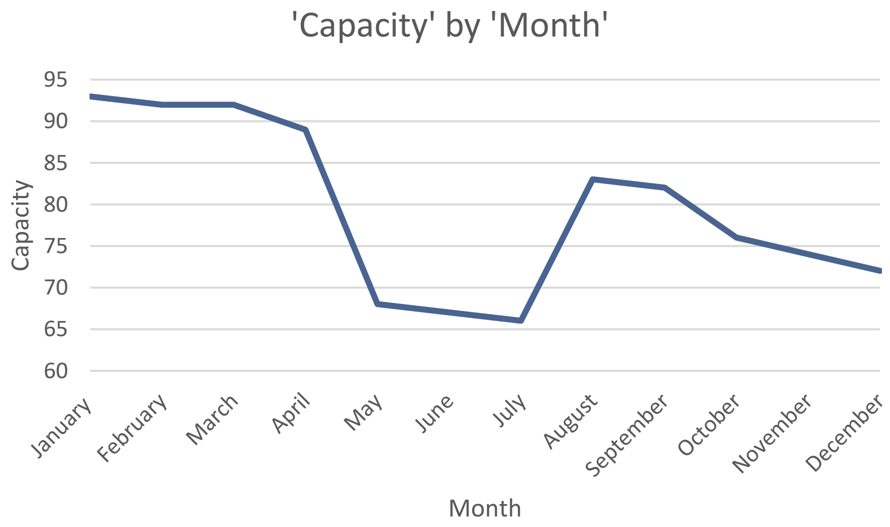


FIGURE 2.13 Bar charts must have a zero baseline





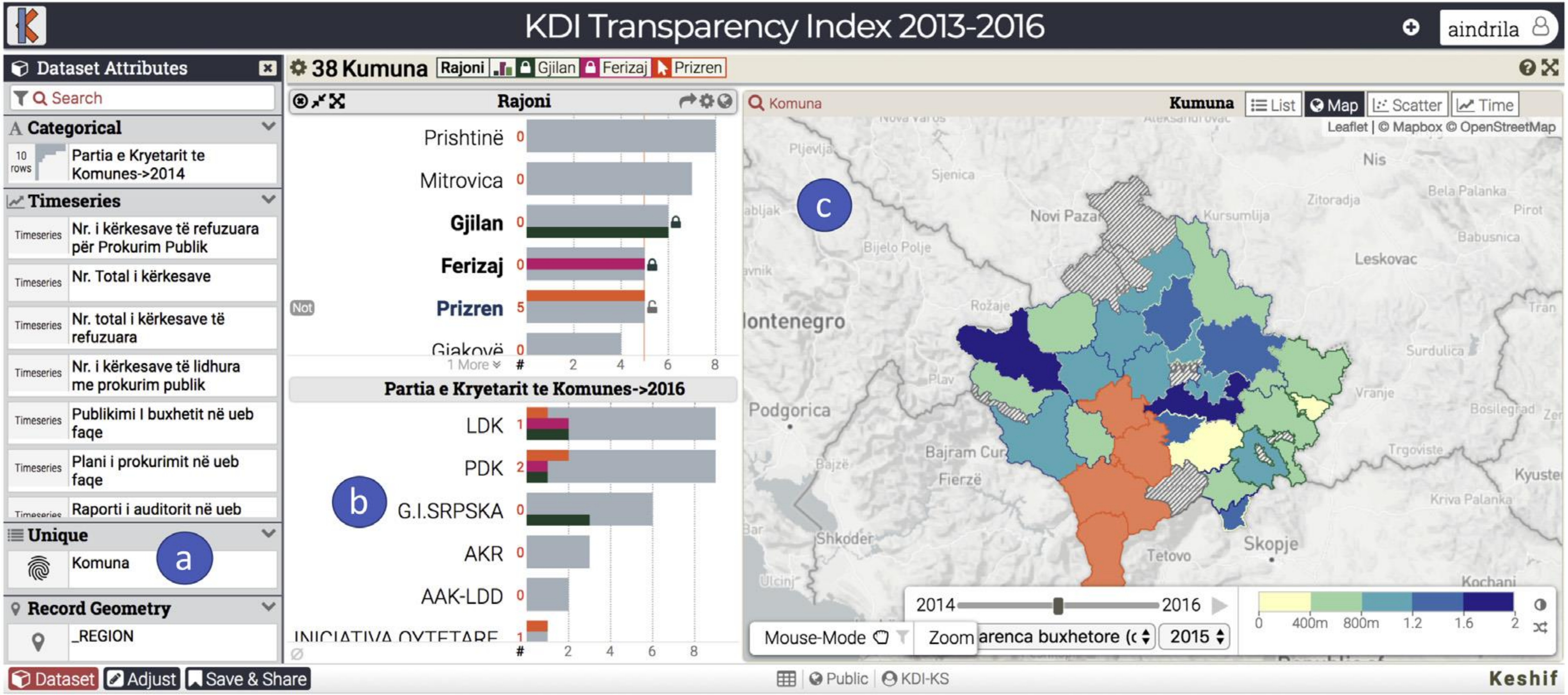
Abbreviated Axes

Source: Cole Nussbaumer Knaflic

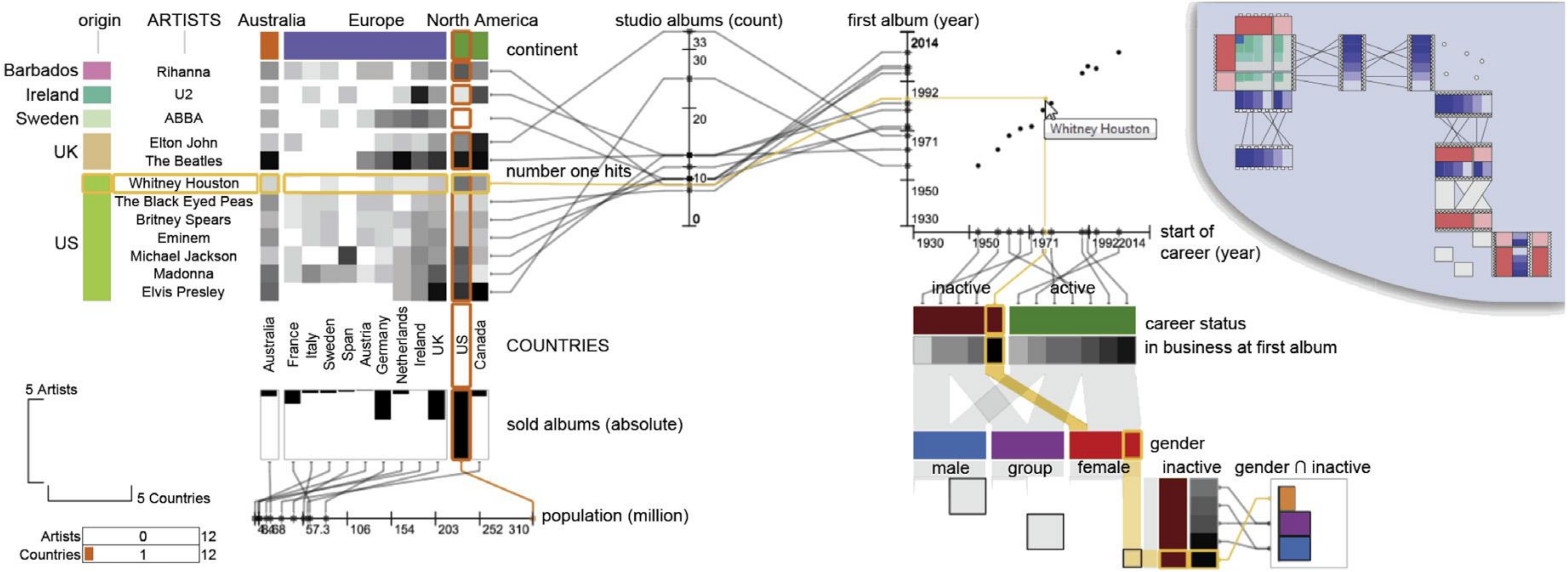
DATA VISUALISATION IN MANUFACTURING

DATA VISUALISATION IN MANUFACTURING

Source:
Silvertone Analytics



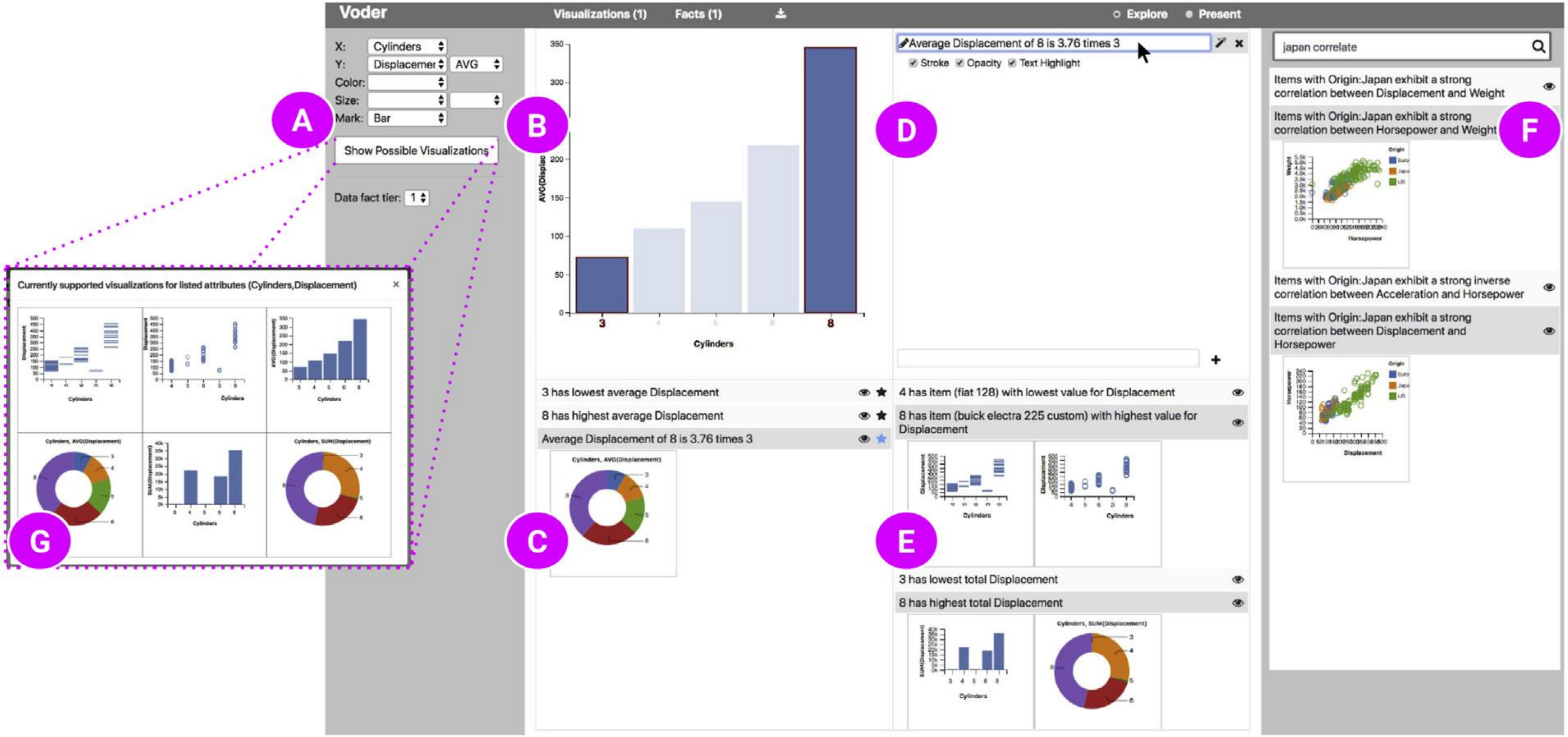
Dashboard of the Tool Keshif (Yalçin et al., 2018). In the figure:(a) Keshif enlists the attributes in the dataset in groups such as categorical, quantitative, time-series data. (b) For bivariate and multivariate analysis Keshif allows users to lock histograms of up to three attributes. (c) Attribute relationships are also shown on visual representations that allow users to switch to different visuals and/or filter the data [7].



The tool Domino (cf. Gratzl et al. Fig. 1 Gratzl et al., 2014) showing the relationships between data subsets using parallel coordinates and scatter plots [7].

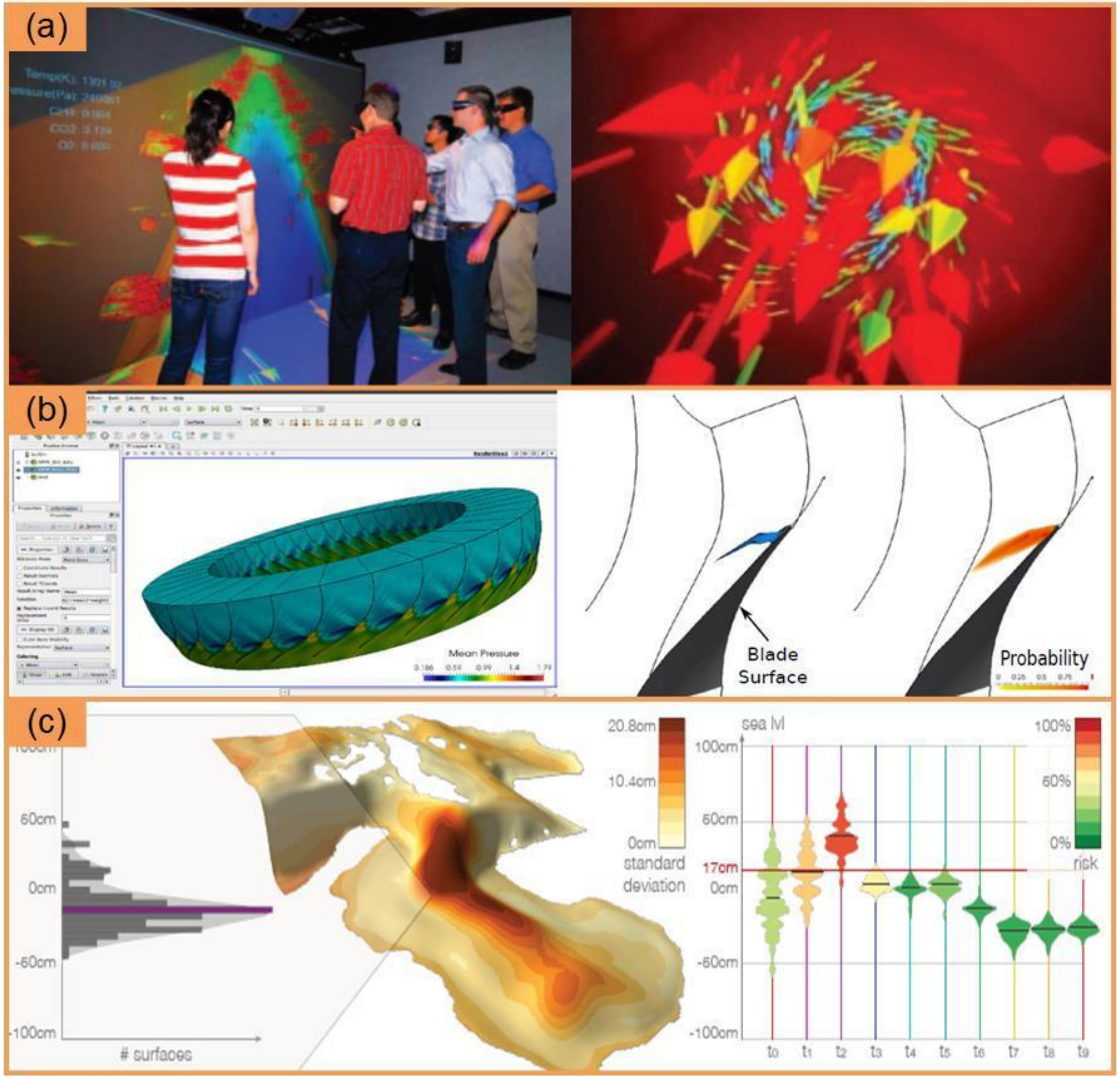
Country	DwellingsWithoutBasicFacilities	HousingExpenditure	RoomsPerPerson	HouseholdNetAdjustedDisposableIncome	HouseholdNetFinancialWealth	EmploymentRate	JobSecurity	LongtermUnemploymentRate	PersonalEarnings	QualityOfSupportNetwork	EducationalAttainment	StudentSkills	YearsInEducation	AirP
Australia	1.1	20	2.3	31588	47657	72	4.8	1.08	50449	92	76	512	19.4	

Dashboard of ForeSight (cf. Demiralp et al., 2017). In the figure: (a) shows univariate attribute distributions, (b) shows outliers in the data, (c) linear correlations among attributes, (d) tabular access to underlying data, (e) bookmarks of data exploration, (f) related insights [7].

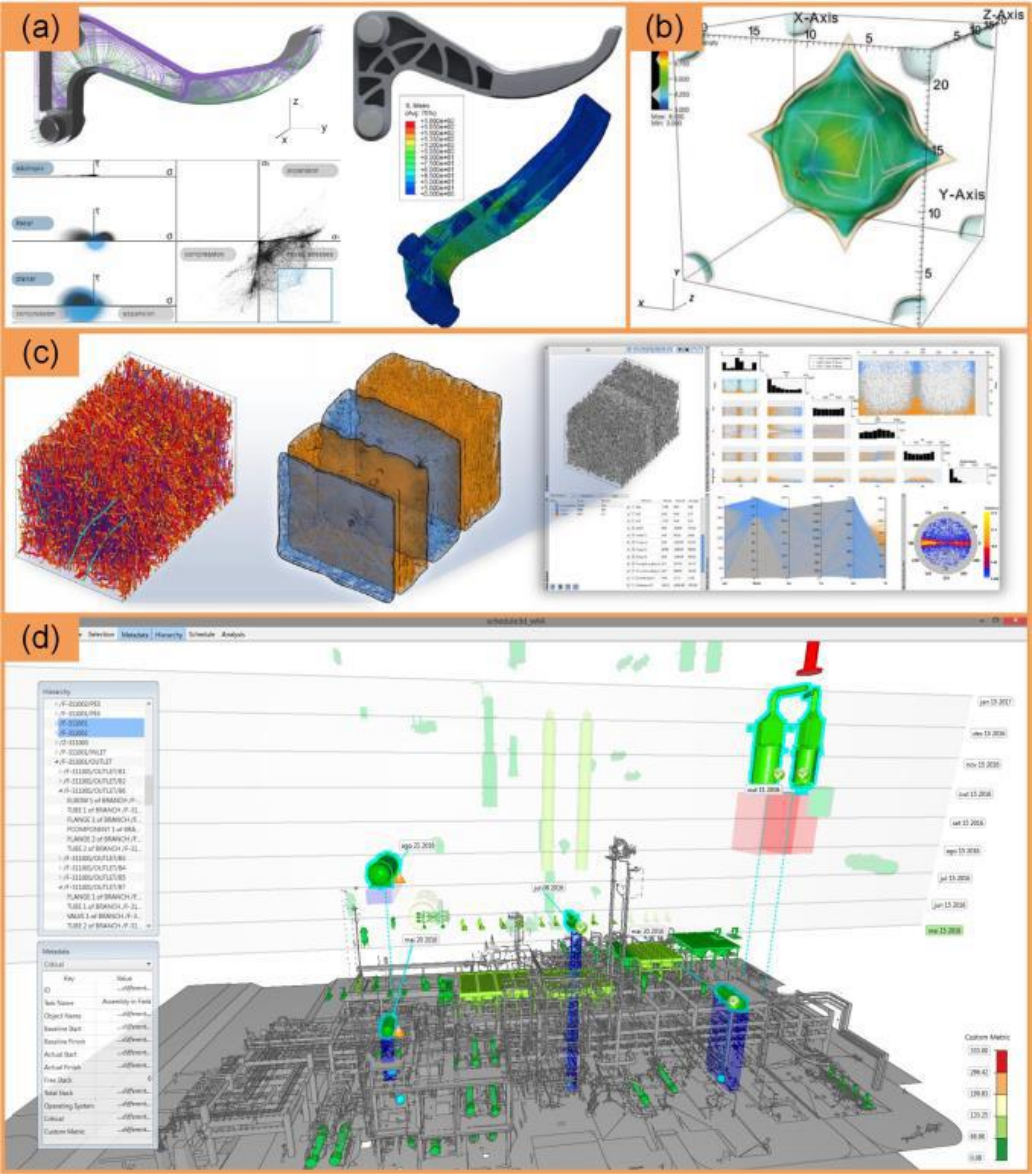


Explore view of the interface of the tool Voder (cf. Srinivasan et al.- Fig. 4 Srinivasan et al., 2018). In the figure: (A) shows specification of visualization, (B) shows active visualization, (C) automatically generated data facts, (D) starred data facts about the current visualization, (E) System generated visuals for other data facts that can be explored, (F) Query panel for data facts, (G) possible visualizations for the chosen attributes [7].

Application cases of scientific visualization in industry manufacturing. (a) Steelmaking furnace internal environment visualization; (b) Jet engine internal environment visualization; (c) Oil exploration external environment visualization [8].

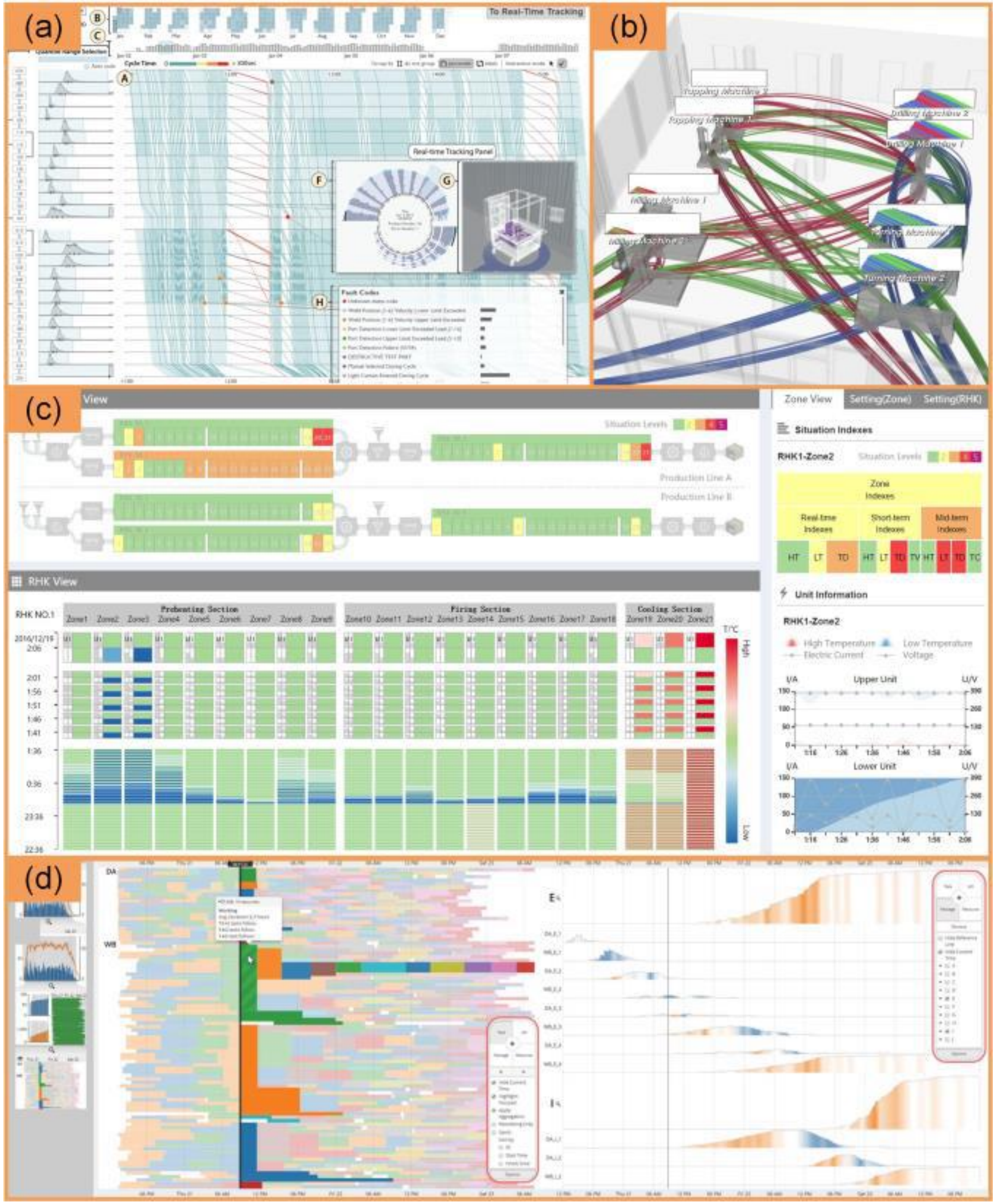


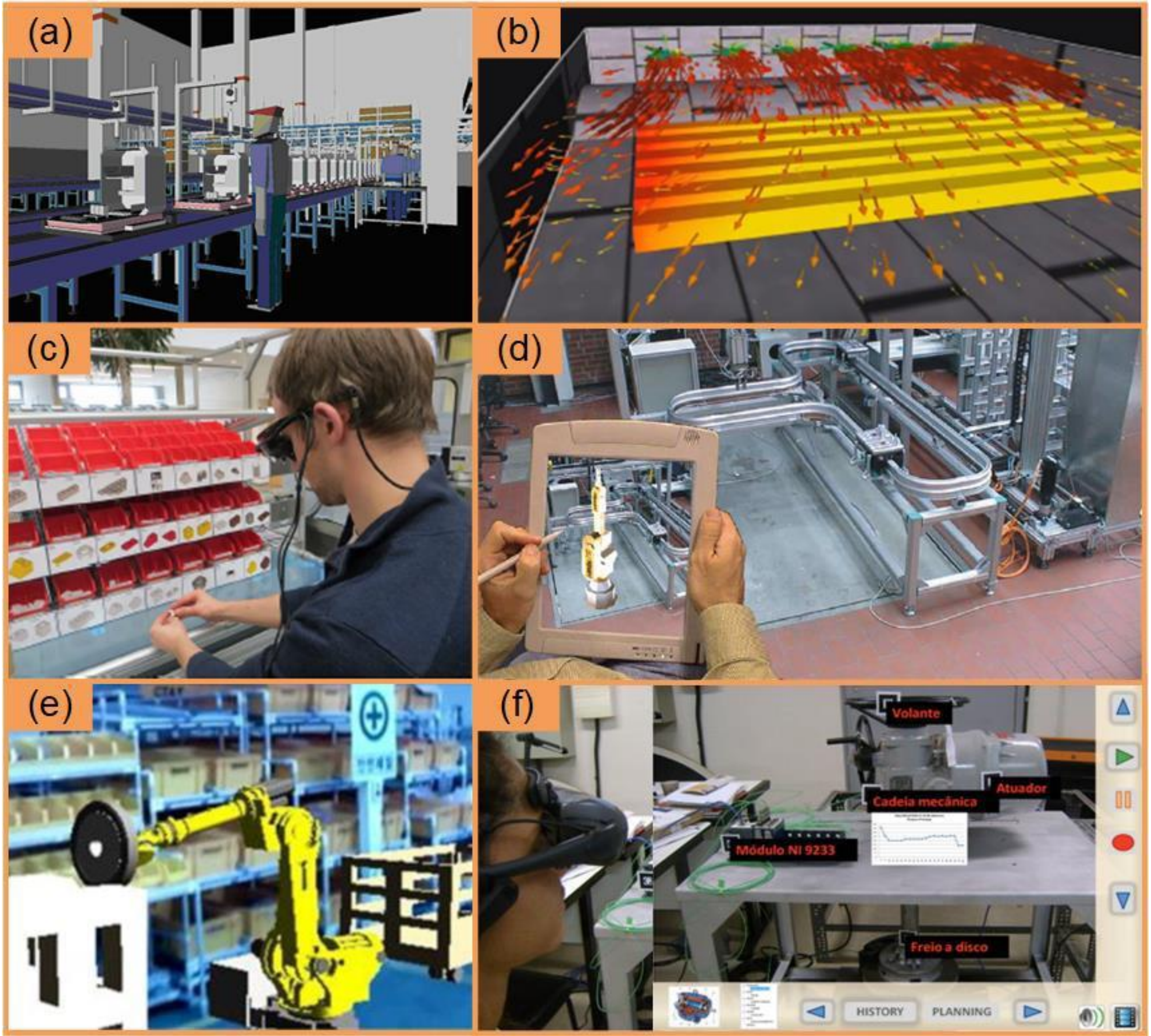
THE UNIVERSITY of EDINBURGH
School of Engineering



Visualizations for design phase: (a) Structural design of product; (c) Material characteristics analysis; (d) Production environment design [8].

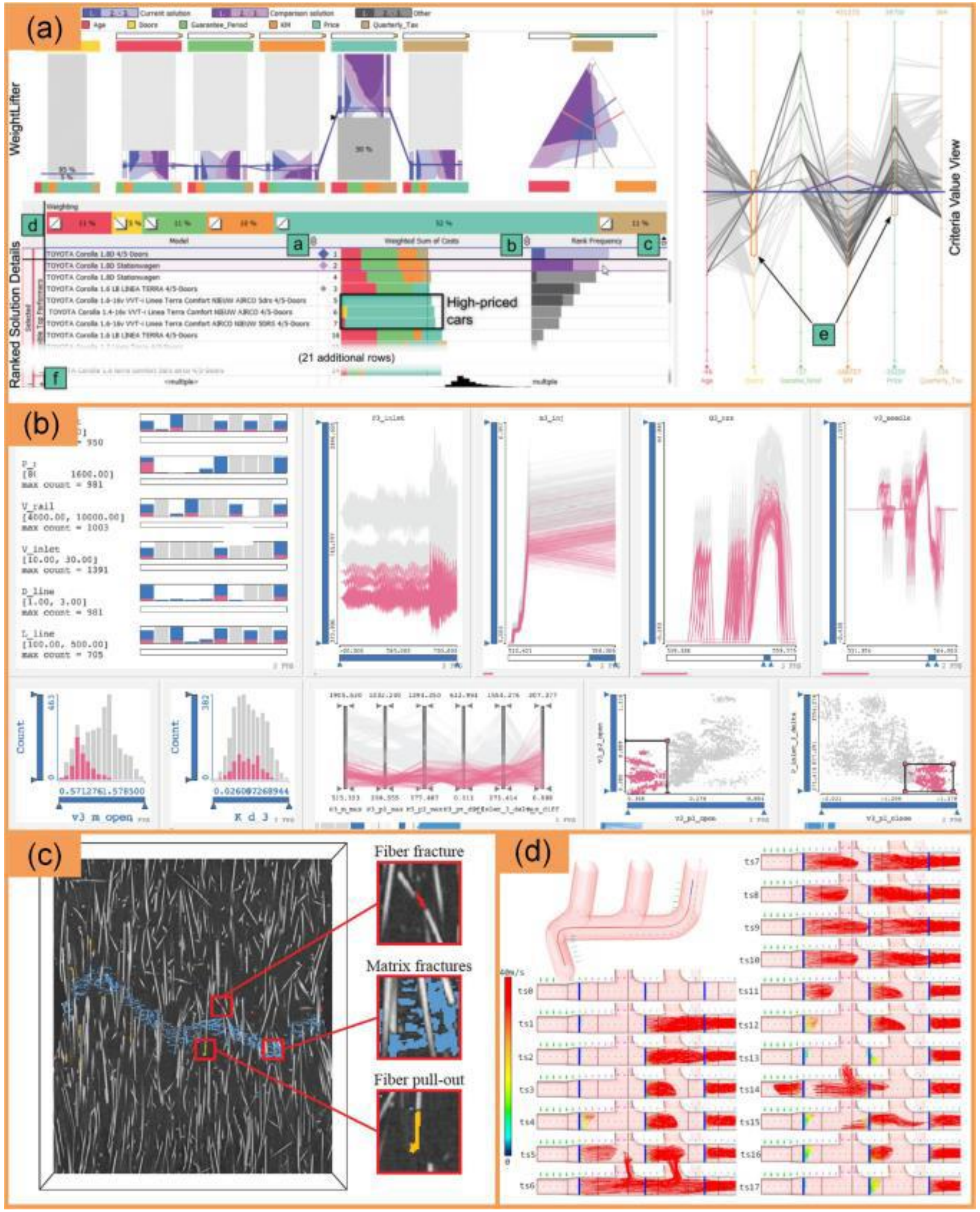
Visualizations for design phase: (a) Structural design of product; (c) Material characteristics analysis; (d) Production environment design [8].





Immersive analytics. Application cases of VR, AR and MR in industry manufacturing. (a) VR assembly factory; (b) VR furnace hot gases escaping; (c) An assembly worker wearing AR glasses; (d) AR supported production line modeling; (e) MR workshop environment; (f) MR equipment interface [8].

Visualizations for testing phase. (a) Multi-dimensional test data analysis; (b) Ensemble testing data analysis; (c) Image testing data analysis; (d) Flow testing data analysis [8].



HOW TO CHOOSE DATA VISUALISATION FORMAT

Who is your audience?

Expertise

Culture

Accessibility

What insights does the data graphics need to produce?

Change minds?

Inform?

Influence?

Statement?

How is the visualisation going to be presented?

Presentation

Storytelling or showcasing?

REFERENCES

1. Medium. 2022. *What is Exploratory Data Analysis?*. [online] Available at: <https://towardsdatascience.com/exploratory-data-analysis-8fc1cb20fd15>
2. MacInnes, John. "Exploratory Data Analysis." SAGE Research Methods Foundations, Edited by Paul Atkinson, et al. London: SAGE Publications, Ltd., 2020, doi: https://www-doi-org.ezproxy.is.ed.ac.uk/10.4135/9781526421036889602.
3. Medium. 2022. Dramatically Improve Your Exploratory Data Analysis (EDA). [online] Available at: https://towardsdatascience.com/dramatically-improve-your-exploratory-data-analysis-eda-a2fc8c851124
4. Devopedia. 2020. "Exploratory Data Analysis." Version 29, December 7. Accessed 2021-09-09. https://devopedia.org/exploratory-data-analysis
5. Britannica, The Editors of Encyclopaedia. "Gestalt psychology". *Encyclopedia Britannica*,
6. Friendly, Michael. "A brief history of data visualization." *Handbook of data visualization*. Springer, Berlin, Heidelberg, 2008. 15-56
7. Ghosh, Aindrila & Nashaat, Mona & Miller, James & Quader, Shaikh & Marston, Chad. (2018). A comprehensive review of tools for exploratory analysis of tabular industrial datasets. Visual Informatics. 2. 10.1016/j.visinf.2018.12.004.

REFERENCES

8. Zhou, F., Lin, X., Liu, C., Zhao, Y., Xu, P., Ren, L., Xue, T. and Ren, L., 2018. A survey of visualization for smart manufacturing. *Journal of Visualization*, 22(2), pp.419-435.
9. Pierson, Lillian. *Data Science for Dummies*, John Wiley & Sons, Incorporated, 2017. *ProQuest Ebook Central*, http://ebookcentral.proquest.com/lib/ed/detail.action?docID=4812516.
10. Sharma, Megha. 2017. "Descriptive Statistics in R." Data Analytics Edge, June 16. Accessed 2018-04-15.
11. Math Open Reference. 2011. "Outlier." Accessed 2018-04-15.
12. Friendly, Michael and Daniel J. Denis. 2001. "Milestones in the history of thematic cartography, statistical graphics, and data visualization.“
13. 20 Stunning Data Visualisations from Around the Web. https://www.designbysoap.com/blog/2019/1/20/20-stunning-data-visualisations-from-around-the-web

RESOURCES

- Information is beautiful awards: https://www.informationisbeautifulawards.com/showcase?acategory=science-technology&action=index&controller=showcase&page=1&type=awards
- Behrens, J. T. (1997). Principles and procedures of exploratory data analysis. Psychological Methods, 2, 131–160.
- Cochran, W. G., Mosteller, F., & Tukey, J. W. (1954). Statistical problems of the Kinsey report on sexual behaviour in the human male. Journal of the American Statistical Association, 48, 673–716.
- Good, I. J. (1983). The philosophy of exploratory data analysis. Philosophy of Science, 50, 283–295.
- Hoaglin, D. C., Mosteller, F., & Tukey, J. W. (1983a). Fundamentals of exploratory analysis of variance. John Wiley.
- Lakshminarayan, K., Harp, S. and Samad, T., Imputation of Missing Data in Industrial Databases, 1999.
- Lee, Katherine J, and Julie A Simpson. “Introduction to multiple imputation for dealing with missing data.” *Respirology (Carlton, Vic.)* vol. 19,2 (2014): 162-167. doi:10.1111/resp.12226
- EDA: https://towardsdatascience.com/dramatically-improve-your-exploratory-data-analysis-eda-a2fc8c851124
- EDA: https://r4ds.had.co.nz/exploratory-data-analysis.html
- Seltman, Howard J. "Experimental design and analysis.” Chapter 4: https://www.stat.cmu.edu/~hseltman/309/Book/chapter4.pdf

RESOURCES

- Outliers: https://www.mathopenref.com/outlier.html
- Anscombes quartet: https://towardsdatascience.com/importance-of-data-visualization-anscombes-quartet-way-a325148b9fd2
- Python graph gallery: https://www.python-graph-gallery.com/pie-plot/
- Cheatsheets: https://www.python-graph-gallery.com/cheat-sheets/
- Colour theory basics: https://www.revunit.com/post/the-role-of-color-theory-in-data-visualization
- EDA: https://devopedia.org/exploratory-data-analysis
- Filliben, James J. and Alan Heckert. 2003. "Exploratory Data Analysis." Chapter 1 in NIST/SEMATECH e-Handbook of Statistical Methods. Updated March 2018. Accessed 2018-04-15.
- Lile, Samantha. 2017. "44 Types of Graphs Perfect for Every Top Industry." Visme Blog, July 5. Accessed 2018-04-15.
- Sander, Liz. 2016. "Telling stories with data using the grammar of graphics." CodeWords, Issue Six, March, Recurse Center. Accessed 2018-04-15.

RESOURCES

- Siddiqi, Adnan. 2018. "Introduction to Exploratory Data Analysis in Python." Python Pandemonium, March 3. Accessed 2018-04-15.
- Ganguly, Ambarish. 2017. "Little Book on Exploratory Data Analysis." October 1. Accessed 2018-04-15.
- InData Labs. 2017. "Exploratory Data Analysis: the Best way to Start a Data Science Project." Medium, June 19. Accessed 2018-05-03.
- Anscombe's Quartet: https://www.youtube.com/watch?v=Ftp3mmItV-k
- Five Ways to Mislead with Data Visualizations: https://www.tessellationtech.io/top-five-ways-mislead-data-visualization/