KEEPING IT REAL WITH DATA VISUALIZATION

Rob Carver and Volker Kraft
Stonehill College, Easton, MA 02357
SAS Institute – JMP Division, Heidelberg, Germany
volker.kraft@jmp.com

Data visualizations can excite and engage students with multivariate thinking about REAL data. Modern software tools, including professional-grade packages and free on-line tools, allow students to interact with data with social, political and environmental importance. This interactivity helps bridge the gap between theory and decision-making with data, fosters intellectual curiosity and exploration of "what if" scenarios, and facilitates the communication of findings and results, as well as addressing several recommendations of the ASA's 2016 GAISE report. In this talk we'll explore the role of interactive visualizations in statistical education and will use publicly available data to illustrate dynamic tools for data visualization.

INTRODUCTION

For a good reason (e.g. Shneiderman, 1996), today's world is full of data visualization. Developments like a "grammar of graphics" (Wickham, 2010) or a "taxonomy of tools" (Heer & Shneiderman, 2012) help us to visualize data more effectively. During the last years, easy access to technology like professional-grade software or free on-line tools allowed teachers to fully embrace visualizations for teaching statistical concepts in the classroom (Forbes et al., 2014).

A huge selection of technology exists for data simulation, data visualization and statistical concept demos. In line with teaching best practices mentioned in the following section, we will show how to use data visualization throughout all stages of a problem-solving scenario. Starting with real and engaging data in a real and active learning environment can add substantial pedagogical value to the teaching process, but also make for a lot of excitement and fun from the learner's perspective.

OBJECTIVES

This paper and the selected case aim at demonstrating how visualizations using real data can support teaching of statistical thinking, as recommended in the GAISE College Report (2016):

- "Integrate *real data* with a context and purpose": Requires flexible data import; tools for data cleanup and preparation; feature extraction like text mining or functional data analysis; big data.
- "Focus on *conceptual understanding*": Requires visualization tools for statistical data and concepts at all stages, from EDA, CDA, data modeling to sharing findings and story-telling.
- "Use *technology* to explore concepts and analyze data": Requires tools that integrate interactive visualization with standard analytic procedures.

As a further GAISE recommendation, courses should teach statistics as "an *investigative process* of problem-solving and decision-making", stimulating "statistical thinking in a multi-variable context". This requires a progressive and problem-driven workflow, an interactive and dynamic interface as well as tools to manipulate data or investigate what-if scenarios.

One technology choice to meet those requirements is JMP – *Statistical Discovery* software from SAS (Why teach with JMP?, 2018). While JMP is widely-used in industry, it is very accessible to undergraduate first learning statistics; it can also add a real-world experience and provide relevant skills to students in a statistics course.

THE CASE

Happiness and well-being concern all humans, no matter of their age and home country. Not only governments consider happiness as a proper measure of social progress, but also individuals aim to better understand what leads to a happy life (Achor, 2013).

We suggest using happiness data to stimulate statistical thinking in an undergraduate first statistics course: The World Happiness Report (2017) including datasets is publicly available and a landmark survey of the state of global happiness for many years. It ranks 155 countries by their happiness levels and relates happiness to social and economic key factors. Here we focus on chapter 2 "Social Foundations of World Happiness".

The survey results come as an Excel file, which can be directly imported into JMP or other statistical software. As the only data preparation steps, a few country names had to be recoded to match with local shapefiles for geographic mapping, and some groupings like "Countries of Interest" or high-, medium- and low score countries were added to better support data filtering, coloring etc. Rows states were assigned for colors, markers and labeling.

All datasets, analyses and visualizations from this case will be made available.

Explore recent happiness scores (2014-2016, see Fig. 2.2 in WHR2017)
Data: Happiness score by country (averaged 2014-16).
Tasks: What can you tell from the distribution of scores? Can we assume a normal distribution? (one variable at a time)
Used: Histogram, boxplot, normal quantile plot, fitting continuous distributions.
Notes: Figure 1 shows one stage in exploring the distribution of scores. Students can start with a histogram, visually vary binning intervals and locations, or show a shadowgram. Normality can be evaluated visually, following JMP's progressive and data-driven workflow. Students are motivated playing a detective's role, collecting visual and numerical clues to make discoveries.

Tasks: How do the individual countries rank? Where can we find highest and lowest happiness scores? What about some select countries of special interest?
Used: Dot plot with contour. Colored and marked data points (e.g. highest 10 green, lowest 10 red). Tagging of individual points to show their countries and scores.
Notes: While the contour provides context, students can compare individual countries simultaneously. The points also help students build understanding of what the contours mean. A data filter can focus on subsets, depending on the story to tell.

Tasks: What about countries in the same region? Any outliers?
Used: Geographic map-ping, colored by happiness scores. Tagging countries.
Notes: Although, historically, mapping is not mentioned in stats courses, it can be the simplest and clearest way to summarize data. Different coloring gradients can support different insights. For many countries, like Syria or Venezuela, there is a well-known reason for a conspicuous coloring.

Explain the drivers of happiness (2005-2016, see Table 2.1 in WHR2017)
Data: "Life ladder scores" (so-called Cantril ladder question, asking people to evaluate the quality of their current lives on a scale of 0 to 10), by country and year. Four social and two economic key factors for happiness.
Tasks: Evaluate changes in happiness over time.
Used: Data filter to select countries of interest. Graphing scores over time, colored by country. Smoother with confidence bands.
Notes: Supports a comparison of trends based on economical or social situations. Confidence bands help especially in case of missing data (Nigeria).

Tasks: Explore relationships between happiness and factors (multivariate). Used: Matrix with bivariate scatterplots. Density ellipses at 95% confidence.
Notes: The shape of the ellipses visually inform about correlation, e.g. between 'Log GDP per capita' and 'Social Support'. Coloring shows clustering of happier (>6), neutral and unhappier (<4) countries. Scatterplots show outliers, suspicious points can be selected and tracked for follow-up.
Other options for exploring multiple variables include dynamically linked histograms, 3D scatterplots or parallel plots.
Former univariate analysis suggested to use a log transform of 'GDP per capita'.

Fig. 6 shows correlations as a color map: 'Life Ladder' (first variable) is highly correlated with a cluster of the first three factors. Two other clusters include factor 4/5 and factor 6, respectively.

Packed bars (Fig. 7) blend a bar chart and a treemap, here for 'GDP per capita' by 'country', colored by 'Life Ladder'. It allows to focus on the ten leading countries, seeing the full context for other (skewed) data at the same time. The GDP tail shows unhappier countries (in red). Colum switcher to look at other factors.

Tasks (“teaser” about data modeling): Can happiness be predicted based on the six factors? Which factores are most important? What-if scenarios like “What can we say about countries with lower or higher life expectancy?”
Used: RSM model and stepwise regression for variable selection. Prediction profiler for understanding and exploring relationships.
Notes: The model explains 77% variation (Fig. 8). Drivers for happiness are ranked in the effect summary (Fig. 9) and can be experienced in the interactive profiler (Fig. 10), incl. interaction effects.

CONCLUSION

Although JMP software has been originally developed for professional-grade problem-solving, it’s DNA seems to match perfectly with the GAISE recommendations for teaching statistics. JMP’s dynamic output and progressive workflow provide a *direct interface* between the learner and the data, while the best practices built into JMP’s analysis and graphing tools (*Want to Make More Effective Graphs?,* 2018) open many powerful methods to less technical and more applied users.

Combining adequate technology with exciting data and real-world problems can also provide a “roadmap to happiness” in teaching statistical thinking, both for the teacher and the students.

REFERENCES

Achor, S. (2013). *Before Happiness: Five Actionable Strategies to Create a Positive Path to Success.* New York: Random House, Inc.

Forbes, S. et al. (2014). *Use of Data Visualisation in the Teaching of Statistics: A New Zealand Perspective.* IASE/ISI: Statistics Education Research Journal, 13(2), 187-201.

GAISE College Report ASA Revision Committee (2016). *Guidelines for Assessment and Instruction in Statistics Education College Report 2016.* http://www.amstat.org/education/gaise.

Heer, J. & Shneiderman, B. (2012). *Interactive Dynamics for Visual Analysis*. New York: Communications of the ACM 55(4), 45-54.

JMP course material (2018). Cary: SAS Institute – JMP Division. http://www.jmp.com/courses

JMP free trial (2018). Cary: SAS Institute – JMP Division. http://www.jmp.com/try

Shneiderman, B. (1996). *The Eyes Have It - A Task by Data Type Taxonomy for Information Visualizations*. Washington, DC: IEEE Computer Society, 336-343.

Want to Make More Effective Graphs? (2018) ANALYTICALLY SPEAKING with Xan Gregg. Cary: SAS Institute – JMP Division. http://bit.ly/2sZhTXA

Why teach with JMP? (2018) Cary: SAS Institute – JMP Division. http://www.JMP.com/why

Wickham, H. (2010). *A Layered Grammar of Graphics*. AMSTAT: Journal of Computational and Graphical Statistics 19(1), 3-28.

WHR World Happiness Report (2017). Chapter 2: Online Data. Sustainable Development Solutions Network, New York. http://worldhappiness.report/ed/2017/

FIGURES

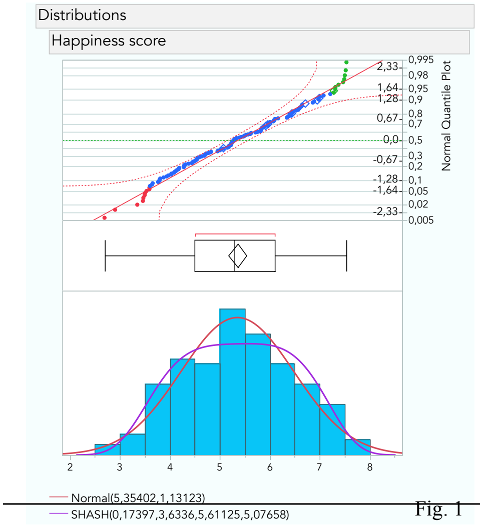


Fig. 1

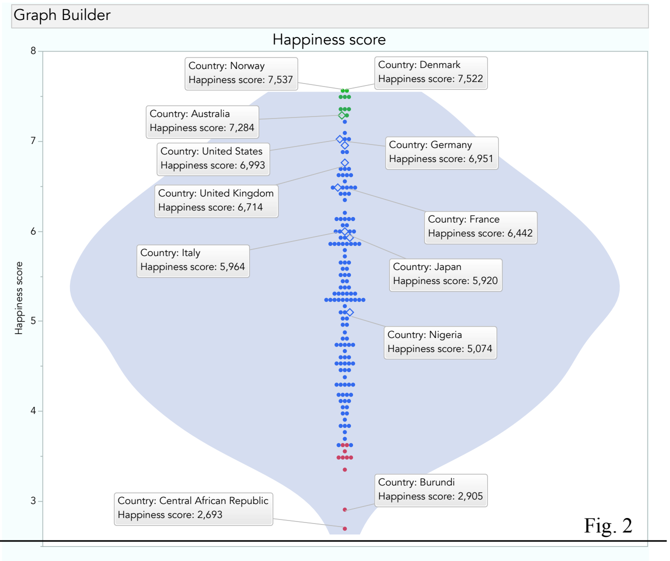


Fig. 2

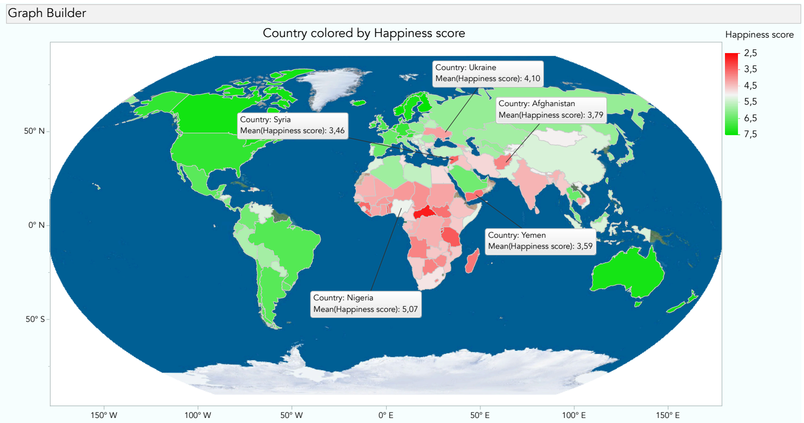

Graph Builder
Country colored by Happiness score
Happiness score
2,5
3,5
4,5
5,5
6,5
7,5
Country: Ukraine
Mean(Happiness score): 4,10
Country: Syria
Mean(Happiness score): 3,46
Country: Afghanistan
Mean(Happiness score): 3,79
Country: Yemen
Mean(Happiness score): 3,59
Country: Nigeria
Mean(Happiness score): 5,07
50° N
0° N
50° S
150° W
100° W
50° W
0° E
50° E
100° E
150° E
Map shape not found: North Cyprus.

Fig. 3

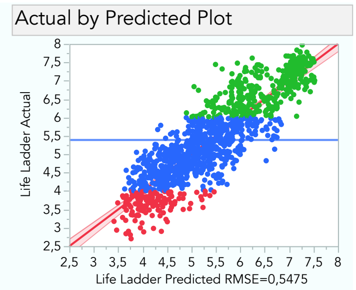


Fig. 8

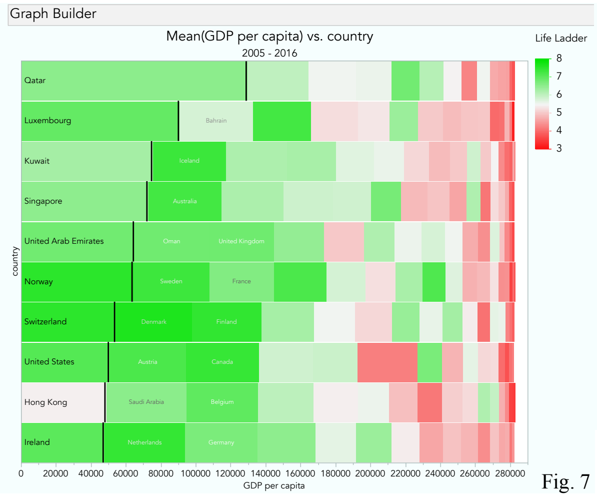


Fig. 7

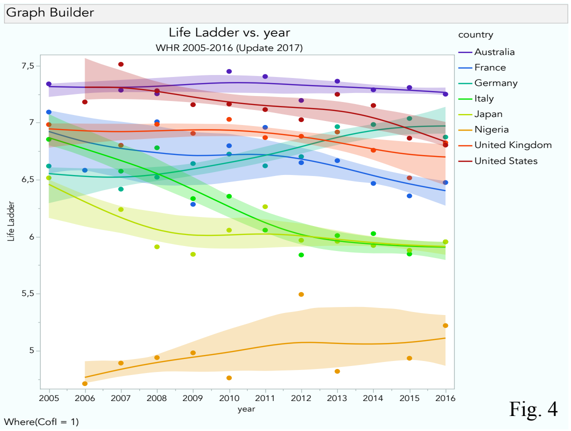


Fig. 4

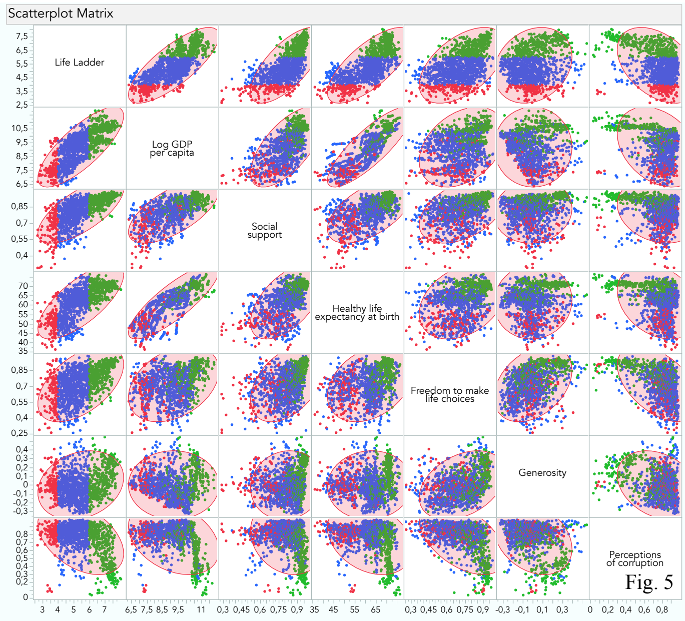
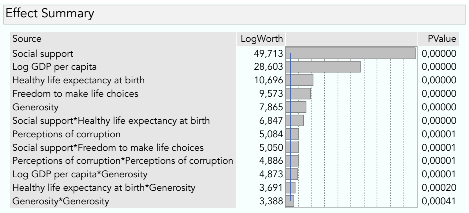


Fig. 5



Effect Summary

Source	LogWorth		PValue
Social support	49,713		0,00000
Log GDP per capita	28,603		0,00000
Healthy life expectancy at birth	10,696		0,00000
Freedom to make life choices	9,573		0,00000
Generosity	7,865		0,00000
Social support*Healthy life expectancy at birth	6,847		0,00000
Perceptions of corruption	5,084		0,00001
Social support*Freedom to make life choices	5,050		0,00001
Perceptions of corruption*Perceptions of corruption	4,886		0,00001
Log GDP per capita*Generosity	4,873		0,00001
Healthy life expectancy at birth*Generosity	3,691		0,00020
Generosity*Generosity	3,388		0,00041

Fig. 9

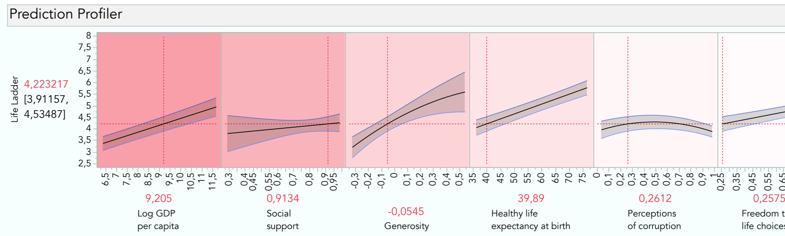


Fig. 10

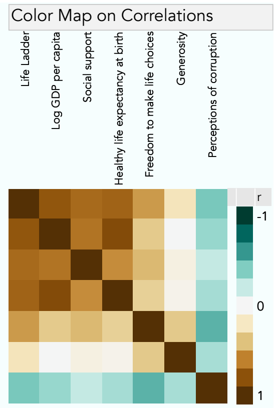


Fig. 6