Best Practices

Prof. Eric A. Suess

Best Practices for Data Visualization

Today we will be discussing some of my ideas about best practices for data visualization.

Please note that these are my suggestions, other people may have different suggestions. Data Visualization is *subjective*, but using good **quantitative sense** you can produce excellent visualizations that communicate with data.

CRISP-DM and Data Mining

- When I think about stepping back and looking at my own data visualizations I always think about Step 5 of CRISP-DM. This is the **Evaluation** step.
- This is the step where I ask myself, “can I show this to my boss?”
- There are a number of simple questions you can ask yourself before you turn in your work.

Question 1

- **Does the visualization look good?**
- Remove junk and think about the color used.
- Are there any things that stand out that do not look right? If there are things that do not look right they need to be fixed or removed.
- Do the colors look correct? Have you used a continuous scale, divergent scale, or categorical scale correctly?

Titles, main and subtitles

- **All visualizations need a title and it needs to be one that describes what is being visualized.**
- The main title refers to the data.
- The subtitle(s) about the variables.

Labels for the axes

- **Do not use variable names as the labels on your plots.**
- I always ask myself is anyone other than me going to know what the axes represent?

Does the verticle axis include zero?

- **When creating a scatterplot have you included the zero on the y-axis?**
- Without the zero on the y-axis the slope of the trend is not accurately presented.
- While this is a problem if the exact value of the slope is important, this is commonly not done by default in software.

Less is more

- **Do not overload your plot with data just because you can!**
- When plotting points on scatterplots, timeplot, maps, etc., when there are too many points on the plot it will be overwhelmed and will not communicate the information that is being presented. It will just look like a mess.
- When creating a dashboard, having approximately 4 plots visible at one time, is a good rule-of-thumb.

Labeling with arrows

- **When putting text into a plot to label a specific point, the label should not overlap any other part of the visualization.**
- The *less is more* suggestion also applied here.

One variable at a time

- **It is important to keep in mind that you should use each variable in your dataset only once on a plot.**
- In ggplot use a variable only once. For example, do not use a variable for both color and size. This can be confusing and is not good practice to do so.
- Be careful changing *area* in bargraphs and bubble charts.

Same scale(s) for comparable plots/faceting/dynamic visualization

- **When comparing visualization the same scales should be used.**
- When faceting the y-axis and x-axis on **all** plots should be the same.
- When putting different plots together, look closely at each pair of plots and ask yourself, “can the scales can be compared?”

Start small

- **Wrangle your data as needed, while making your visualizations.**
- **Making visualizations is an iterative process.**
- Start with a small sample of your data to make your first visualizations.
- Determine the maximum number of points that can be used.
- With *scatterplot* this is not very many, same with boxplot.
- Modern *hexplots* are maybe better for showing the relationship between variables when working with a big data set.
- Modern *violin plots* are maybe better for showing the distribution of a variable when working with a big data set.

Latitude and Longitude

- **Interview Question:** If you plot geolocation data on a scatterplot, which axis does the latitude go on? Which axis does the longitude go on?

Ask for a review

- Before delivering your visualization(s) to your boss, have someone else look at your visualization to give you feedback.
- Ask the person, “do you understand what is being presented in the visualization?” Please explain.

Tell a story

- In the **Claus O. Wilke** YouTube video that was assigned for homework, what was the main point of his discussion?
- The main point was that visualizations are used to *tell stories*.
- Often more than one plot is needed to complete the story.

Color has cultural meaning

- In the **David McCandless** YouTube video that was assigned for homework, what was the **eye** in the video at 6:13 minutes? I am referring to the Data is Beautiful video.
- This picture is about the different use of *color in different cultures*. See page 337 of the Visualize This book by Nathan Yau.

Dynamic visualization is exciting

- In the **Hans Rosling** video(s) what is the most important thing we learned from his video(s).
- That *dynamic data visualization* can be **exciting**.

Conclusion

- Make yourself a check list that you will use in the future to self **Evaluate** your own data visualizations.