Search entries or author...

All

Sort

**Marcus Birkenkrahe**

AUTHOR | TEACHER

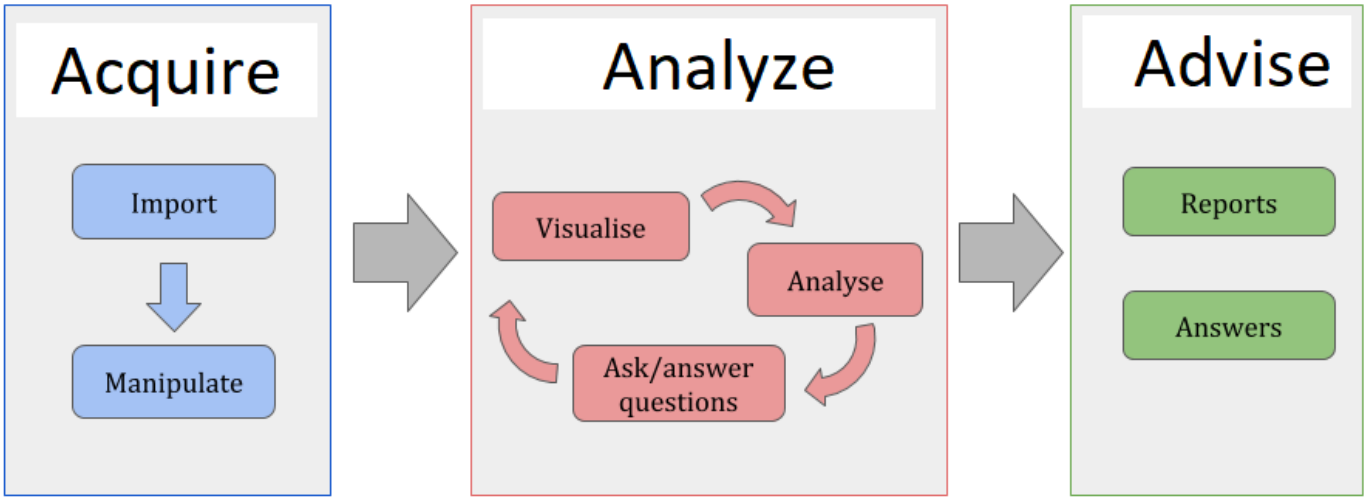
Dec 13 5:16pm

## The Final Message: Next steps / your results / final words of wisdom

Dear students! I'm about to submit the final grades for this class. I've enjoyed working with you, and I have especially enjoyed your term projects. I am looking forward to hear specific comments and suggestions for future iterations of this introductory course.

I'd like to use this last message to *1) outline what you could do next,* 2) *visualize your results*, and 3) *pass on some final words of wisdom.* So this message will be a little longer - but data science is a growing field, and visualization is a core topic!

1. **Next steps:** You know that "data visualization" is officially but one step in the data science process (**see figure (https://github.com/birkenkrahe/dviz/blob/piHome/img/2_workflow.png)** ). But 'visualization' cannot easily be isolated - you need to keep the whole "gestalt" of data science, the whole, in mind when working on creating graphs.




In your projects, you've already gone through the whole process! The best thing you could do is to identify your weaknesses based on this experience. When working with R, it is important to have alternative options - for example, if you did all your plots with `ggplot2` (because Google gives you more examples based on this package's current popularity), code the graphs using `base R` instead. If you imported your data in `CSV` format, save the data as an `Excel` file and import that, too. For deeper analysis consider additional plots. Even better: broaden (or narrow) your research question and see how far you can go using your data. Some of you already identified followup questions that you might use.

Our textbook, Pearson's "Explorative Data Analysis Using R" contains a chapter "Crafting Data Stories" with three advanced examples to try out. DataCamp has a course on EDA using `ggplot2` and `dplyr` (instead of base R) with a nice application (spam protection) at the end.

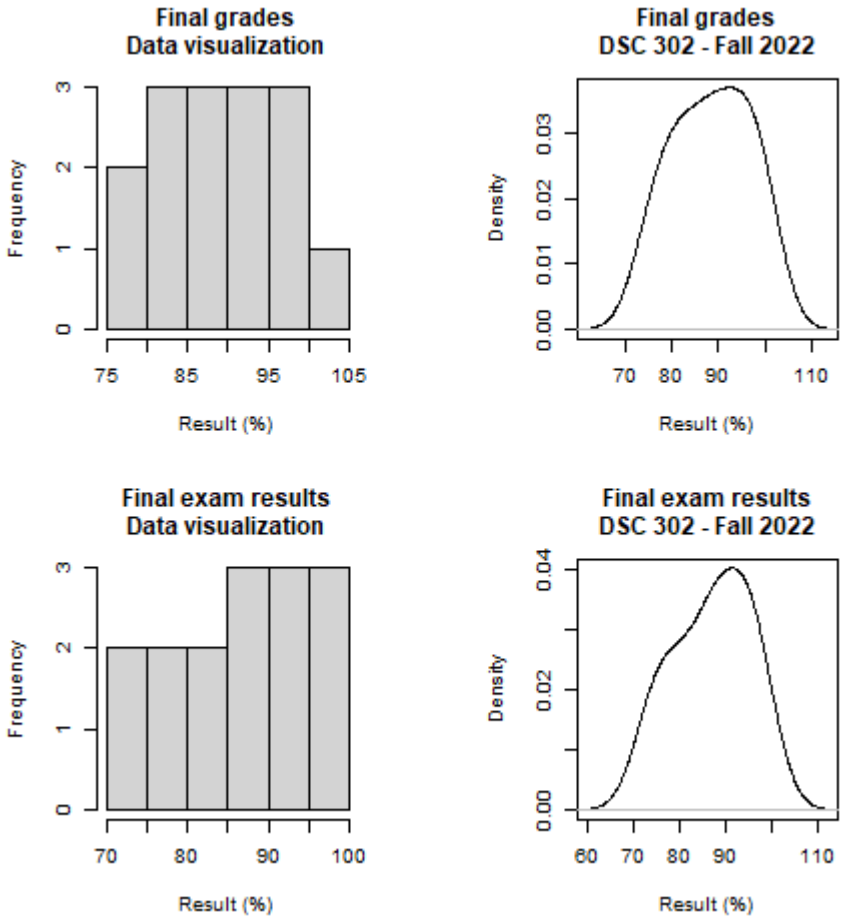
This morning I watched a **DataCamp webinar (https://www.datacamp.com/webinars)** on a new offering, "**DataCamp competitions (https://app.datacamp.com/learn/competitions)**". Take a look at some of the past results and/or try your hand on new problems. The very format of a submission for these competitions is a positive example of data storytelling, and at least the problems labeled "beginner" are totally within your reach. You create your analysis and your final report in DataCamp's "**workspace (https://app.datacamp.com/workspace/overview)**" tool, which is a**Jupyter notebook (https://jupyter-notebook-beginner-guide.readthedocs.io/en/latest/what_is_jupyter.html)** for R, Python and SQL (you can choose your language). Once you've finished, you publish your result for review, which is like presenting at a large conference! No fear! I might integrate these competitions into **next term's courses (https://github.com/birkenkrahe/org/blob/master/spring23plan.org)** (let me know what you think of that).

Remember also the **book suggestions (https://github.com/birkenkrahe/dviz/blob/piHome/org/1_overview.org#other-sources)** at the beginning of this term.

Lastly, "**Literate Programming (http://www.literateprogramming.com/)**" (LP): you've seen it, and you've done it throughout this course, and if you come back to my classes next term, you'll get more of it. It's at the heart of "data storytelling" and hence of "data visualization". But LP also has a story of its own, which includes the creation of the world's most advanced typesetting software (TeX), an adventure game, and plenty of C programming and smart algorithms. In data science, a whole industry has sprung up around interactive notebooks and hence literate programming - this may be worth your while, too.

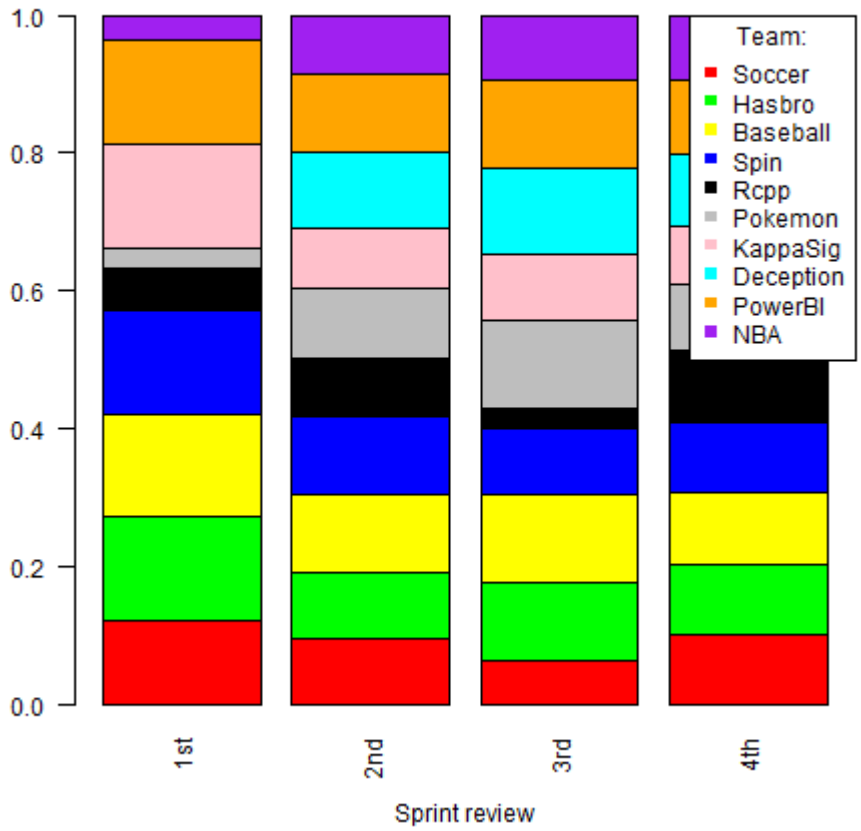
2. **Your results:** I've made a couple of charts using the tools that you learnt (and/or saw) in this class.

**The first set of charts** (https://github.com/birkenkrahe/dviz/blob/piHome/img/dviz_f22_final.png) compares the final grade distribution as a histogram (left) and as a density estimate (right). The second row shows the final exam results. You notice that the grades are almost normally distributed. The average lies at 88% (B+).




As in all courses, the distribution of the grades for the final exam surprised me a little given that all questions were known beforehand.

**The second chart** (https://github.com/birkenkrahe/dviz/raw/piHome/img/dviz_f22_projects3.png) shows a stacked barplot with the results of the four sprint reviews for the projects in the course of the term. The special aspect of this data is that many "teams" weren't actually teams at all but individuals (Baseball, Spin, Rcpp, Pokemon, KappaSig), which may account for the fluctuations in effort: it is a little easier to turn in consistent performance when you have team members who watch you or who have your back.




[You can look at the code for these graphs in GitHub (https://github.com/birkenkrahe/dviz/blob/piHome/org/grades.org)](https://github.com/birkenkrahe/dviz/blob/piHome/org/grades.org) (sanitized - all private information removed).

3. **Final message:** following on from what I said earlier in "next steps" - do not lose yourself in fancy graphics and in the details of artistically or efficiently rendering data in visual form. An emphasis on `ggplot2` in particular or the `"Tidyverse"` in general seems to foster this attitude. I believe you'll learn to tell convincing stories when you a) really have a story to tell, and b) need to convince someone. For most of you, this will be a few years out. In the meantime, do not forget about the data, and build skills **across the entire data pipeline**. Put differently: every time you read something or write some code, or look at some data, or if someone asks you to analyse something, ask yourself the [10 questions identified early on in the course (https://github.com/birkenkrahe/dviz/blob/piHome/org/2_data_eda_R.org#questions-to-ask-from-data)](https://github.com/birkenkrahe/dviz/blob/piHome/org/2_data_eda_R.org#questions-to-ask-from-data):
    1. Where does the dataset come from, and how is it documented?
    2. How many records (rows) does this dataset contain?
    3. How many fields (variables, columns) are included in each record?
    4. What kinds of variables are these (e.g. numerical, categorical)
    5. Are there missing values? (`NA`)
    6. If there are missing values: are these variables always observed?
    7. If there are missing values: how are they represented?
    8. Are the variables included in the dataset the ones we expect?
    9. Are the variable values consistent with what we expect?
    10. Do the variables exhibit the relationships we expect?

REFERENCES:

- Knuth (1984). Literate Programming. CSLI Lecture Notes No. 27, Stanford U.
- Pearson (2019). Explorative Data Analysis Using R. CRC Press/Routledge.


Reply