Search entries or author...

All

Sort

Marcus Birkenkrahe

AUTHOR | TEACHER

Dec 13 10:17am

Next steps / Your results / Final word

Dear students - I'm about to submit the final grades for this class today. I've enjoyed working with you, and I have especially enjoyed your term projects and all your efforts during the many interactive exercises in class. I am looking forward to hear specific comments and suggestions for future iterations of this introductory course.

As always, I'd like to use this last message to 1) outline what you could do next - with links, 2) visualize your results - with charts, and 3) pass on some final words of wisdom. So this is a little longer - but data science is a huge field!

Next steps: if you're enrolled in DSC 205, I'll help you. If not, I suggest you complete the DataCamp course **"Intermediate R" (https://app.datacamp.com/learn/courses/intermediate-r)** while you still can (your subscription ends on January 16, 2023).

If you're more inclined towards computer science, take a look at Matloff's "**The Art of R Programming (https://nostarch.com/artofr.htm)**", which covers the next level of sophistication when it comes to R, or learn some Python. DataCamp has an **excellent introductory course (https://app.datacamp.com/learn/courses/introduction-to-data-science-in-python)** - you can cover a lot of ground in no more than 1/2 day.

If you've got the heart and mind of a data engineer, you need to go deeper into the shell and learn data science on the command line: most of will transcend Windows, the poor excuse for an operating system, and will lead you to Linux. DataCamp offers an "**introduction to shell (https://app.datacamp.com/learn/courses/introduction-to-shell)**" and a course on "**data processing in shell (https://app.datacamp.com/learn/courses/data-processing-in-shell)**", which you'll find useful.

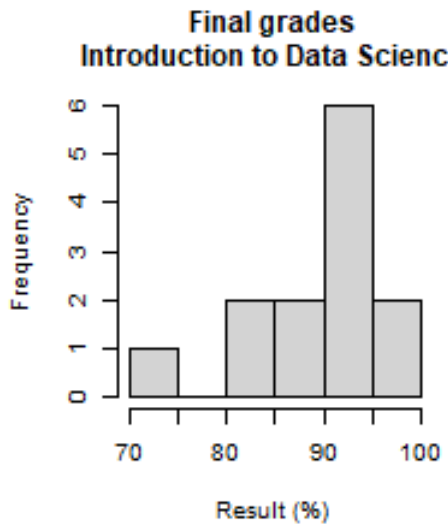
If you're more interested in data analysis and visualization, you must take a look at the `data.table` package for big data and fast data processing (**here in DataCamp (https://app.datacamp.com/learn/courses/data-manipulation-with-datatable-in-r)**), and at the DataCamp course "**Exploratory Data Analysis in R (https://app.datacamp.com/learn/courses/exploratory-data-analysis-in-r)**", which uses `dplyr` and `ggplot2`. Neither of these are really necessary for EDA but you still should know about them and the course is quite fun.

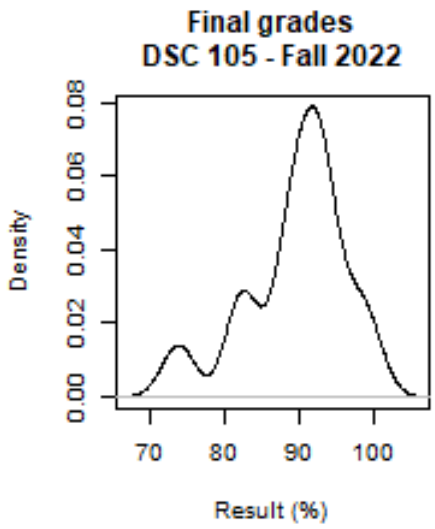
Whatever you do, if you haven't done it yet, **you must learn SQL basics**. You can do this in DataCamp or elsewhere. I strongly recommend SQLite to save time and hassle (you save yourself user management and concurrency, both of which are unnecessary at the start). Look at the free **sqlitetutorial.net (https://www.sqlitetutorial.net/)** and then check out the `RSQLite` package to create attractive visualizations with R.

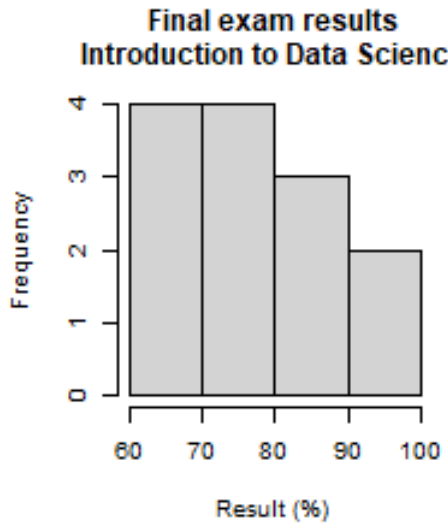
Do not forget your daily practice exercises on the **mobile DataCamp app** (https://www.datacamp.com/mobile) to keep your fingers nimble and your knowledge fresh! Watch their webinars **(I'm watching a webinar on DataCamp competitions** (https://www.datacamp.com/webinars/datacamp-competitions) right now - may do some of this next term). And subscribe to **R-Bloggers** (https://www.r-bloggers.com/) to see what's going on.

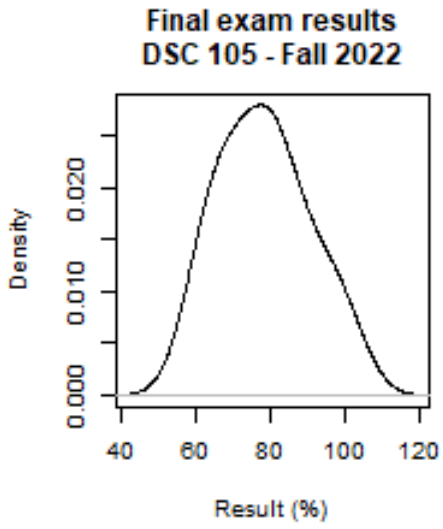
Your results: I've made a couple of charts using the tools that you learnt in this class (and then some - like pattern matching, and writing little functions):

The first set of charts (https://github.com/birkenkrahe/ds1/blob/piHome/img/ds1_f22_final.png) compares the final grade distribution as a histogram and as a density estimate.



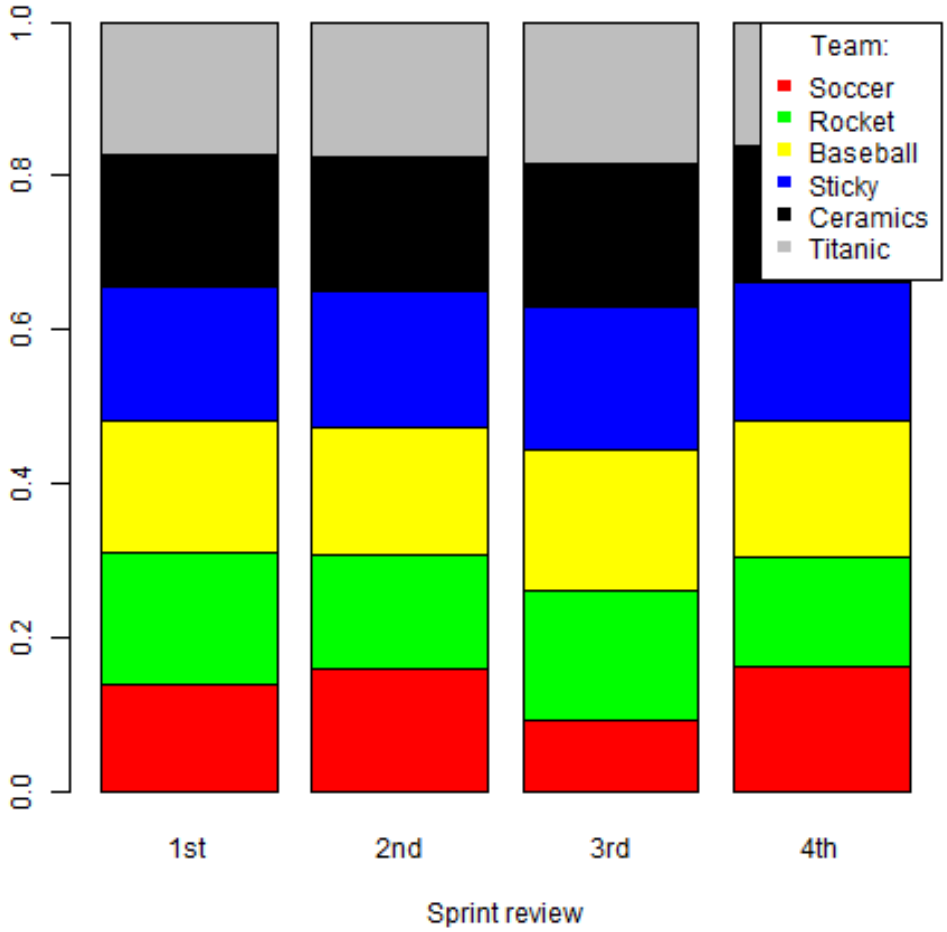






The distribution of the grades for the final exam surprised me a little given that all questions were known beforehand.

The second chart (https://github.com/birkenkrahe/ds1/blob/piHome/img/ds1_f22_projects.png) shows a stacked barplot with the results of you four sprint reviews for the projects in the course of the term.



As you can see, the effort was continuous for all teams across the term, with a few minor differences between teams.

Final words: this was an introductory course, and that's what you got. Data science continues to grow and dominate the tech-headlines (most recently with ChatGPT - transformer tech **demystified here by Google's Chief Decision Scientist) (https://kozyrkov.medium.com/introducing-chatgpt-aa824ad89623)** . More importantly for you and perhaps for everyone: it's growing in the background and in the underground. Much of the technological infrastructure of the world will be based on one of the flavors of data science like analytics (in business especially) or machine learning (in automation and production), see **Davenport and Patil (2022) (https://hbr.org/2022/07/is-data-scientist-still-the-sexiest-job-of-the-21st-century)** . At this time, you still have a chance to stay ahead of these developments! Don't be a sheep and use the opportunity to become a shepherd! *On a more practical note: many of you need to put more time into practicing outside of the classroom...this is the only way to get good!*

REFERENCES:

- Davenport and Patil (15 July 2022). Is Data Scientist Still the Sexiest Job of the 21st Century? Online: hbr.org.
- Kozyrkov (7 Dec 2022). Introducing ChatGPT! The Revolutionary New Tool for Conversation Generation. Online: kozyrkov.medium.com.
- Matloff (2011). The Art of R Programming. NoStarch.

Keep calm and row on! (Image: unsplash.com)

Reply