Gmail **Andrew Ho <kironide@gmail.com>**

Finishing up Day 1 Week 1

5 messages

Jonah Sinick <signaldatascience@gmail.com> Mon, Feb 15, 2016 at 6:45 PM
To: david@bolin.at, Ali Bagherpour <ali.bagherp@gmail.com>, Andrew Ho <Kironide@gmail.com>, Chad Groft <clgroft@gmail.com>, Jacob Pekarek <jpekarek@trinity.edu>, Jaiwithani <jaiwithani@gmail.com>, James Cook <cookjw@gmail.com>, Linchuan Zhang <email.linch@gmail.com>, Matthew Gentzel <magw6270@terpmail.umd.edu>, Olivia Schaefer <taygetea@gmail.com>, Satvik Beri <satvik.beri@gmail.com>, Tom Guo <tomguo4@gmail.com>, Trevor Murphy <trevor.m.murphy@gmail.com>

Congratulations on getting through the first day! :-)

Please send me a private email with:

- Your code
- Your favorite thing about today.
- Your least favorite thing about today.
- At least one technical question.

Tomorrow you'll switch partners, and we'll be getting keyboards and doing pair programming in earnest.

Look forward to seeing you tomorrow,
Jonah

Andrew <kironide@gmail.com> Mon, Feb 15, 2016 at 6:55 PM
Reply-To: Kironide@gmail.com
To: Jonah Sinick <signaldatascience@gmail.com>

Hi Jonah,

Here's my code.

My favorite thing about today was the pair programming model & the assignment. It worked really well, and I enjoyed it a lot & also significantly improved my level of comfort with R. It was probably one of the best (top 3) structured educational experiences I've had in the past ~21 years (for reference, my life is ~21 years thus far).

My least favorite thing about today was realizing that multiple R libraries had overlapping function names, so sometimes I would get mysterious unexpected behavior until I realized that I had to use e.g. dplyr::select(...) instead of just select(...). This took up a lot of time for Ali and I to debug at first (but it was very satisfying once we realized what was going on, which was the silver lining to the cloud).

My technical question is: What are some ways to speed up really slow functions like ggpairs()?

Best,

Andrew
[Quoted text hidden]

3 attachments

beauty_own.R
1K

day1Assignment.R
3K

day1Example.R
7K

Andrew <kironide@gmail.com> Mon, Feb 15, 2016 at 6:55 PM
Reply-To: Kironide@gmail.com
To: Jeremy Li <h.jeremy.li@gmail.com>

[Quoted text hidden]

3 attachments

beauty_own.R
1K

day1Assignment.R
3K

day1Example.R
7K

Jonah Sinick <signaldatascience@gmail.com> Mon, Feb 15, 2016 at 7:45 PM
To: Andrew Ho <Kironide@gmail.com>

Glad that you had such a positive experience :-)

Yes, the dplyr select namespace conflict thing is really bad.

I don't know much about speeding up ggpairs. I think that the slowness is probably in significant part a function of R just not being that efficient. A few things to keep in mind:

- If you have a categorical variable with many categories, that will slow things down a lot when using ggpairs, because it will try to plot interactions of all of the categories with the other variables.
- This wasn't the case of the data that we had today, but if you have a dataset with a very large number of examples (e.g. 10 million) , doing scatterplots can be very slow because it's plotting every point.
- It looks like some computers can parallelize plotting http://stackoverflow.com/questions/8364288/what-hardware-limits-plotting-speed-in-r.

[Quoted text hidden]

Jonah Sinick <signaldatascience@gmail.com> Tue, Feb 16, 2016 at 12:22 AM
To: Andrew Ho <Kironide@gmail.com>

http://stackoverflow.com/questions/29946087/why-is-ggallyggpairs-significantly-slower-in-rstudio-vs-base-r

On Mon, Feb 15, 2016 at 6:55 PM, Andrew <kironide@gmail.com> wrote:
[Quoted text hidden]