Transcription results:

Interviewer: 00:01	Great. So today, I'm just going to start off talking a little bit about the workshop, what you thought of it. Cast your mind back to March and then we'll take another look at your workflow that we talked about, see if anything's different, and then we'll just wrap up with that quick questionnaire that we did last time. So it probably won't be very long.
Researcher: 00:23	Okay, good.
Interviewer: 00:24	But yeah. So first, did want to cast our minds back to the workshops. So here we are in June and the workshop was in March. So first of all, were you able to attend both days?
Researcher: 00:35	Yes.
Interviewer: 00:35	Okay. And what did you think of it?
Researcher: 00:39	Do you have the schedule for that workshop? Because I--
Interviewer: 00:43	Not [crosstalk]--
Researcher: 00:44	I don't know. There were parts that were more better and parts that were worse, yeah.
Interviewer: 00:50	And the Python one.
Researcher: 00:51	The Python, yeah. You had an R one..
Interviewer: 00:55	At the same time, yeah.
Researcher: 00:56	I saw that. And then, yeah, overall, it was nice, it was paced well, teaching was done well, there was a good number of people who talked, I think. Two people. And for the most part, for mine, because everybody comes to this with different background, and I haven't been taught Python. Never ever. I just kind of picked it up when I needed. But it seemed, for the most part, too basic for me. It didn't have enough meat to it. I did find very useful but at the time, I thought it would be very useful, but it turned out I haven't used it since. It began very nicely we're just using-- was it bash shell? Bash, yeah. Just getting to know how the computer, particularly the Mac, because I got out and everybody own a PC. And ever since the Mac, it's like a mystery to me what's going on there. So it just help me oriented where are the files. I go up and go through the files. That actually helped me. It's like being given a map of where you live suddenly and then, oh, there, thats were stuff is. So that was very useful. And then it came the actual Python stuff. And for the most part, it was very easy. So there was the GitHub, third part. So GitHub, I wasn't sure if it was relevant to me and it wasn't. I have never had an urge to look it up again, what we learned there because it just doesn't apply to my needs right now. But I understand where it's coming from and when you're a data scientist and you want to-- edited, a file edited and protected and it makes sense, to be able to-- it's just not working for me. So that was not useful for me and the actual Python stuff, they were good. I mean, I understood everything but it just wasn't-- it was too rudimentary. I was already if not at that level, close to that level coming into the workshop. So it did help a little bit, just make me feel just more comfortable with looking at the code and understanding some of the basic concepts. It was just basic concepts. And it wasn't advanced applications. I want more advanced applications, and I'll pick up the concepts later. But as a teaching, if this was part of a

six-week course it would have been very excellent to opening two days. If this is what you get, this is your Python then it's just not enough for me.

Interviewer: 05:05 Yeah. Yeah. Anything besides maybe doing more advanced stuff that would have made it better for you?

Researcher: 05:14 I think if it was-- so I thought in the format of a large group that may be the best that can be done-- I don't know if it's a-- but I think just going into-- because every time there was just this-- you go to a specific scenario and your problem and you solve it. And you see-- I mean, you get how to translate it to what you need. So I think just having maybe a follow-up advanced session that you can go to more complicated stuff I think that's all that you need. I think the other one was very good

Interviewer: 06:07 Oh, good to hear. So the next piece is going back to this workflow. So this was the one that we put together back in--

Interviewer: 06:21 Yeah. So I remember we looked at your visualization you had on your laptop. But you had some kind of a-- what was it called? I don't remember the program. But you had all the programs. Yeah, which was a lot of stuff. And we chose one of them. And so this one, just reminding ourselves, you had kind of a large dataset. And a grad student or an undergrad had come in--

Researcher: 06:47 Yeah. And he holds a bunch of code in python. And I wanted to-- I just had a problem just reading these code and just following what he did. And then now that I think about it, I think coming out of the workshop the first thing I did is go back and read his code. And it did make more sense. So in that sense, the workshop did help a lot. But paradoxically, what happened was that I understood his code. And then I realized that it should be done in R, this type of project to just built for R much more than python. So it actually convinced me to use R instead of going to modify his code which was my original idea. I just translated it to R. And wrote the the code to do the similar stuff in R.

Interviewer: 07:51 And what convinced you that would be better in R?

Researcher: 07:55 Because essentially, it's a data frame. It's a rectangular data frame and it's actually, I-- and so it's funny. So, my first reaction was that this should be in R. And then after I started translating it in R, I said, "This is actually more of an SQL problem." And they are going to-- I said, "Okay." So I said, "Okay." Dropped out and I went online and looked for SQL course and I started doing that. And I was really happy. But I just had that feeling that I'd just learned Python and I'm now going R and now SQL. I'm going to be, I don't even want to do this thing. I want to collaborate with somebody who does thing. I just want to know enough to make it a useful collaboration. So now I'm learning three different languages to organize the same dataset. So I stepped on the brakes and stopped my SQL aspirations. It's lovely language I really liked how-- and I realized later that you can just write SQL in R. There was a library that you could just--

Interviewer: 09:15 Oh, cool.

Researcher: 09:15 Yeah. You could just put it in SQL order, essentially. So if I do need some very specific SQL stuff I can stay in the R environment. So I just realized it's just for my-- so Python may be faster, but I didn't need speed. And for just data analysis it just seems it was more statistical in nature. Although R I realize can do machine learning stuff as well without much difficulty. So I don't know where is the limit where you stop using R, but I think I'll go back to Python when I'm doing more classical programming like simulations and maybe image analysis or stuff that has to be more in my mind more like software programming and less data analysis. And so I will keep my data analysis in R.

Interviewer: 10:33 So with this pipeline--

Researcher: 10:35 So actually my Python workshop has convinced me to go to quit Python, yeah.

Interviewer: 10:40 That's a valid outcome. So with this--

Researcher: 10:43 So yeah.

Interviewer: 10:44 --pipeline now. You're not doing the Jupyter, you're doing it in R?

Researcher: 10:48 Yeah. So I stopped the Jupyter stuff. Although I got very comfortable Jupyter. And so what is this for?

Interviewer: 10:57 So any changes? Anything you're doing differently now we can write in pink.

Researcher: 11:01 Yeah. So let's say I have this dataset and yeah. So all I'm changing is I'm not using Jupyter. Although really I got to like it. I don't really like-- it was a good thing. And I think it was really taught well in the workshop with how to use Jupyter and it made a lot of sense to me and this was a hard thing to give up.

Interviewer: 11:26 [crosstalk]--

Researcher: 11:28 Jupyter is so fun. I thought it was fun. So I'm not using this anymore. What I do is I basically unzip and now I read into R. Into R and that's it. And then work in RStudio to do all my my data analysis. And it's actually kind of similar to Jupyter because you can run parts of your code. You have your code there. You could just run the parts that you're interested every time and it's actually fun to write too. And then, graphic description, stats. Yeah. So I do everything in R. Yeah. So it's not in Python. It's all done in R. Yup. I run stats. What I do is I-- so I do all of this thing. I run all the graphs, table, description, stats. Then, my next goal is a stats consult because I don't feel that I'm confident enough to go. So I just give that point-- I do a consult with-- which I try to do. I'm not very successful because there's not a lot of infrastructure here if you're not funded, if this is like-- see the investigator-initiated data analysis, which we probably don't have. But I mean, I have some projects that are funded that it's going to be easier to get a stats consult.

Researcher: 13:30 But I still like to get to this stage where I'm getting to know my data before I get the-- I'm doing some initial thoughts about it. So that makes the stats consult really effective. And, yeah, it kind of goes back-- like now because I don't have money to pay them. I just get the free one, and then I go back and forth and--

Interviewer: 14:00 I'm still working in R Studio and everything, and I'm writing papers in Word. And I remember you like to do the citations by hand at the end.

Researcher: 14:09 It's so much better. No, after struggling with a citation manager for so many years, and back and forth, and different people using different-- just go back to Word. So I completed now like 10 people international. I was just doing parentheses in Words, and it's so easy. It's so fun. Yeah.

Interviewer: 14:32 I think you're very much alone in that. [crosstalk].

Researcher: 14:34 Yeah. But oh yeah, I find it cool. I'm back to normal.

Interviewer: 14:38 Okay. So we were talking about how things are different now using R instead of using Python, which the grads student had done. Going forward, is there anything that you think you're going to do differently with this or you're pretty happy with the R?

Researcher: 14:52 I mean, I think R is just-- they should get a Nobel Prize, the people. It's amazing. It's just an amazing platform to analyze data. So I don't see myself going away from R. So my initial thought after the-- because I just want to do one thing. And probably if you-

- this workshop convinced me that if you just have one language, Python is probably better because you can also do R in Python. What you can do, it does in R module, whatever, you can do. But the reason, what R gives you-- especially the RStudio. It's just a little bit more geared toward data analysis. So the fact you can peak into files quickly and just a little bit more friendly for data analysis. And the Python is-- but you can do everything in Python, too. So my simulation work, because I am interested in simulation. I'm interested in data science stuff. And the simulation parts are going to be more Python. So I don't think I'm going to completely leave Python it will be interesting that the machine learning because I also want to get into doing some machine learning stuff. And that's kind of in the middle. It's not exactly just data processing, and it's not exactly simulations. So it's in the middle. So I hope it's going to be a coin toss whether I do it in Python or R.

Interviewer: 16:50 So you haven't given up on Python completely?

Researcher: 16:52 I haven't completely. But I think it's an amazing-- both Python and R blow me away about how flexible and forgiving the language is, right, coming from back in the days of very strict and unforgiving languages, and that you had to do a lot of-- I mean, there's so much you can-- so many shortcuts and so many stuff you can-- so it's super powerful, both of them. So I'm very impressed with both of them. But I haven't given up entirely on them. Probably, I will have to know them both.

Interviewer: 17:34 At some point?

Researcher: 17:36 Yup.

Interviewer: 17:37 I'll take a look at them so everything's-- and so one thing you mentioned when we spoke originally was that you sort of wanted to be able to look over Python script and understand it more just to not necessarily write it but be able to talk about it or talk to it. Is that something you feel like you're better at now?

Researcher: 17:59 So after the workshop, before I abandoned Python-- so I don't know if I go back now, how hard it is. But, yeah, definitely after the workshop, I went to the-- it completely achieved the goals of what I wanted because I went back to that code that was written by the summer student and I could totally understand it, follow it, and actually translate it into R so I can see exactly what it did and do the same thing. So, yeah. Absolutely.

Interviewer: 18:32 Great. And so we talked about Python. We talked about Git and maybe not being as helpful or just not fitting into this workflow

Researcher: 18:41 I mean, it's just that-- I mean, if this was my-- if I had a data scientist, probably, I'll make him work with it. But yeah for me, it's just too much at this point.

Interviewer: 18:52 Yeah. And then, you mentioned using Unix.

Researcher: 18:56 Using Unix. It was great. It was great. Yeah.

Interviewer: 18:59 And have you started using it for any of your projects?

Researcher: 19:01 No. But it was useful for me now to understand how to find files. Because even in R, you have to choose your working directory, stuff like that. And they use some of the same-- the tilde and stuff like that just make me more comfortable in finding what my stuff is.

Interviewer: 19:26 Yeah. Good. That sounds great. So then, the last thing is, we're kind of interested in barriers to using programming or things that helped use your programming. So was there anything in terms of barriers, so between March and now, things that got in the

way, reasons you weren't kind of able to make changes? Or it sounds like you did make quite a lot of changes?

Researcher: 19:53 Yeah. I mean, there was no barriers. I understood the Python better. I didn't give up, maybe the fact that you had the R workshop just next to it reminded me to there was R and then-- but I knew about R before, so I was back and forth between using it and-- but it wasn't a barrier. I could have just continued with Python. But I think the investment that I felt for this particular project which is mainly now data analysis that I'm doing, I think R will be a quicker learning curve. And to do it well in R. And it sounds also better when you publish that you're using a statistical package from a statistical program than you loaded a library in Python, we could probably make it work anyway.

Interviewer: 21:02 Yeah. And is there anything that the library did or could do to kind of support you in any of this?

Researcher: 21:10 So the problem that I have with the library is that my current schedule, which may change, just you have a lot of the pizza and stuff. You are meeting on Thursdays which is the day that I am in San Jose. So I can never make it on Thursdays. And then the meeting-- so I use [name] twice. And she's very good and she emails back. And it's great but it's not super accessible. So if there were more, if you're working genetically replicating her...

Interviewer: 21:58 Oh, man. We would love that [laughter].

Researcher: 22:01 [inaudible].

Interviewer: 22:03 Yeah.

Researcher: 22:03 If you had like 2 or 3 [name] that would be great.

Interviewer: 22:04 If only we have the money for that [laughter].

Researcher: 22:05 It's wonderful. She's great, very useful, I learned a lot just for the interaction. So yeah. To me, this is wonderful resource to make it more accessible you just need to be more of you. So I don't know if UCSF has more big data people there. I don't know if they have something similar. But this is a super good resource because I find it much more accessible that the [institute] version of it.

Interviewer: 22:49 Then the consultation

Researcher: 22:51 Yeah. consultations and also the availability all kinds of-- you don't see the statisticians doing workshops which is.

Interviewer: 23:07 Yeah. They're not as kind of education focused as we are.

Researcher: 23:09 Yeah. Yeah.

Interviewer: 23:12 Great. All right. So then the last thing is our checklist here. So here, we're just going to go through and talking about your workflow as it is now. Just want to get a sense. Are you using any programming languages like R or Python?

Researcher: 23:29 Yes.

Interviewer: 23:31 All right. Have you transformed any step by step workflows into scripts or functions?

Researcher: 23:37 So scripted like an R script?

Interviewer: 23:40 Yeah. So instead of I would have this function, this function, this function, it's one script that kind of pulls everything together.

Researcher: 23:46 I mean, the idea is, because every projects is different. So for example, now, I went from [database] to another database [database]. And it was just-- so I took the code I wrote and just started replacing it. So I'm not putting the effort into making it automated. So I just have-- because I think it's nice to do. But it started since I just do it manually, write a the script for that project. I love it, I write my script. And even scripts I don't like to run them as a program, like run them from beginning to end. I like to run portions of them step-by-step see what they're doing and to make sure what I am getting.

Interviewer: 24:43 How about using a version control to manage your code?

Researcher: 24:47 Does it say what the version is each time that one or is it keeping it kind of git style?

Interviewer: 24:54 Yeah, more of a git style.

Researcher: 24:56 No, I'm not using any of that.

Interviewer: 24:59 Do you use any open-source software?

Researcher: 25:01 R is open source.

Interviewer: 25:02 Yeah. Do you share of any your code publicly?

Researcher: 25:07 No, but it's because I don't have it. And I think it's interesting because I haven't-- actually, I'm starting to publish my first paper that -- I wrote the code, too. And I don't know if it's a requirement for the paper to share your code of how you analyze the data. It should be. It's a supplemental, yeah. They don't do that unless they require it.

Interviewer: 25:32 You haven't done that in the past?

Researcher: 25:34 I haven't done it yet. I want to do it.

Interviewer: 25:39 Yeah. Do you share your computational work or protocols publicly?

Researcher: 25:46 I don't. But again, I want to. I don't know how, and if they want to request it with the paper, that you have to write an appendix with all your code and stuff, I would gladly do it, yeah.

Interviewer: 26:03 So it's mostly--

Researcher: 26:03 I'm not against it.

Interviewer: 26:05 --yeah, you haven't been asked to do it, it's not really the standard. Great. Okay, so the last thing is we have these workshops and they teach you-- the idea is to teach the basics of programming. But one of our larger goals is we 'd love to teach people how to program so that they can make their work more reproducible, computationally reproducible. So meaning somebody could use the same code that you wrote, the same data, and get the same results in the end. And so I'm interested to hear what you think about that. If you think being a part of this workshop has helped you make your work more computationally reproducible.

Researcher: 26:41 Oh, yeah. Yeah, I think it definitely-- it emphasized some good programming habits and actually when I-- yeah, just showing how to document and how and why it, and the way that the style, and I think there's a-- both R and Python have a very unique way of style to them and how a good programming style to do them. So it was emphasized in the workshop and just seeing other people code, particularly [name], what she said to me back, her code, it was nice. That's how we do it. And that's how we do it. But also you look up the web you will find a lot of times and you realize examples of how-- clean ways to write code and it's definitely become much more computational.

Interviewer: 27:49 Great. Good to hear. Anything else about the workshop or the way you work that's different now that is worth sharing?

Researcher: 27:57 No. No. That's basically it. And I'm looking forward to participating in more relevant workshops. I think with time, yeah, I think a lot of these things will be-- I think after I do a couple of bigger data projects like this like I'm doing now, this part of data analysis will be more reproducible. I would be able to automate that part of my research. But I don't know if I want to. I don't think it's the most efficient use of my time. And going forward is to be the data scientist in the grand scheme of things. So what I am looking is to see how I get the data scientist or part of a data scientist on this project. I mean, although it's incredibly fun, it's just not the best use of my time. I think I should focus on the writing part, in the conceptualizing.

Interviewer: 29:21 Yeah. And all those other projects that you had going on. Breaking bones, was that one of them, crushing bones?

Researcher: 29:26 Breaking bones.Yeah, yeah, yeah. That was real fun stuff to do, but then [laughter].

Interviewer: 29:30 Yeah. That's great to hear. So maybe, for you, it's more wanting to be aware of what's going on and what you can do, and then [crosstalk]--

Researcher: 29:37 Yeah. I want to know how it's done and intimately know the data. I call it tamed data because it's wild when you get it, so. I want to be part of that because otherwise, if you just get a number, you just give them-- [database] give us the number, then there's a big, black box, and we don't know. And I think a problem with the data scientists or any non- clinician is that there is-- this black box can come up in a lot of different ways that are not relevant to what is actually happening in the real world. So I think it's critical that clinicians be involved intimately in that part because you just can't analyze it like just numbers. You have to know what you're doing. Every variable has two as a meaning and has a meaning that's not exactly what it says in the book. When you know how the people seeing patients feel about this variable, you have to know, then you understand its meaning. So I think it's important for clinicians to be involved. But I can't do it alone. Every Monday I have four hours that I put on my schedule. It's called the data analysis hours. And I look forward very much to them. But it gets eaten. But that's all I can spend a week, max. And that's not enough to do all of these interesting projects, so.

Interviewer: 31:29 Yeah. For sure. Great. All right. Well, that's all I have, so thanks for coming to chat with me.

[redacted talk about accessing the Library]