DIVIDE & CONQUER--Practice adept by both experts and novices

Both novices and experts try to split the program into intermediate steps to understand the program

(P256) so it's nice to break it down step by step for interpretability. And then I can kind of see what part helps you understand what it does and then breaking it down to the next chunk, and then the next chunk makes it more interpretable.

(P940) [In understanding the program], at least I have a little bits and pieces flying around in my head...so I was able to, I wouldn't say, obviously I didn't get the complete picture. (38; 37)

(P312) I think when I'm reading the data, I want to know the end of my previous manipulation, whether they work and then I also want to see what I'm going to do next. So I like to print out the head, but I want to print out the whole dataset instead of only getting the rows and columns

(P312) I will try to drop- sometimes I have Pandas because that's the most straightforward, straight-to-home what I dropped. [skimming the part within the function call]

(P940) So the [saving function] breaks down each of the steps the code is doing and that's honestly, if I'm looking at a function that's not hardcore, I think that would help me a lot even in my assignments where sometimes I have to chunk of code and then they would be like, oh, debug it and then I'll have to, I have a hard time reading, so I have to screen through each of these

(P451) I think I'm going to print out the head of the new function after every change so we can see the changes. (0S: 52)

Unfamiliarity of programming language prevent novices from doing head compilation like experts

Both experts and Novices attempt to do head compilation. Experts are more successful than the novices because of their familiarity with the programming languages

(P312) [The best strategy to understand the code is to compile them in my head]

[In response to the question: "What did you find to be the most effective way to understand the code?]

(P451) there's so much going on that I don't really know what part to look at. I like to break it down part by part and looking at where I know. (27; 10)

(P907) I'm going to go line by line and I'm going to make sure that I understand what's been in the scenes. This tends to be the practice that I follow. But sometimes it leads me down these rabbit holes that I'm trying to really figure out what exactly does this look like?

(P504) I'm trying to read through the code and, you know, in a sim-ilar way, 'well, what would that produce?' Instead of actually having to do all of these operations, because I obviously don't want to.

(P312) I compile everything in my head

(P907) So when I was looking at task zero here, I'm trying to run through the code in my head

(P907) 'Dropping axis=1. How-ish' , okay, I'm trying to visualize this [in my head] right now.

[he doesn't print when he says "visualize"]

(P907) What they're doing here is they're taking the values in T1 and they're doing a transformation on it, it makes sense to me. What that does is it provides a score of all the different entries in T1 and how frequently they appear.

[Doesn't see the output, just knows]

(P241) there's a lot of random functions I don't think I've learned yet in between (48: 45)

(P451) I'll wasn't really familiar with melt and rename and transform. (18: 42)

(P720) I have a little understanding of how melt works, but I wouldn't be able to explain it to you. ...I have absolutely no idea what the function "qcut" (36; 53)

(P948) Even this is simple code, it's different[from what I have learned], so I don't think it's what I was thinking.

(P312) It is a little bit difficult for me because there's new functions I never used and it's complicated.

(P906) I know what pd.melt does, but I can't really simulate what it does in my head.

(P907) I found it more difficult because I think well-one, I didn't know what melt did. Two, grouping by, I'm not very familiar of them, and especially with the pivot tables too, it can be difficult to visually keep track of all the different moving parts.

(P907) 'Ha' re going to declare a variable that's the first few different dot commands we've doing to make this work. So it's just a lot of me going through my head and, like, I'm trying to make it a bit less it's complicated when I'm taking so many steps in a row without any sort of break.

Focus on different aspects of the context

While both experts and novices have the intensions to gather the context of program, they primarily focus in different aspects. Experts tend to explore the overall layout of data set and the final application of the program. On the other hand, novices are more interested in reasoning the variable name and writing style of the program

(P256) I think that for me, to understand what's happening in the code, I need to understand each parameter, [...] And especially when you don't have control on what you're doing, I think it's necessary to fill in those gaps to be able to do the details. But I think with more context, details matter less

(P940) [to figure out the program, I look at the variable] "missing_none", then I saw the data, then it makes sense like...we're trying to find missing ships. (0s: 39)

(P256) I think because I don't have the context of the application of this code, it's hard for me to still understand what melt does because I've never used it. So I think if I have more context and identifier variable would be necessary and stuff

(P456) when I'm writing code, I know it is good in mind, or all the steps I'm supposed to do, I can use other GitHub repositories, I can kind of understand. I know what I'm doing, but I don't know why. I'm kind of just reading code and understanding it. [...] what the final intention is.

(P451) maybe the variable name stretched makes sense contextually, I don't know what melting is, but I don't know what melting is. (41: 19)

(P940) [Based on the variable name], the ships were missing, so there's some ships now or was there a melting task that I missed? (30; 50)

(P940) I'm trying to understand the graph, because I still don't know. It says scripts, so I was expecting something related to oceanography (7; 00)

(P256) So it's really nice to see that for T1, all of the entries are single character. I can scroll through, it so, it's not very informative how useful that is.

(P720) [Is express my understanding of the program], the ships were only only kind of missing from this discipline, which is your variable naming. if they pre-processed the abbreviations of their code [sic], the coding grammar in a sense...it's very to read it [for me] (43; 53)

(P720) Back in high school, I used pandas [until] I did remember there's a function that transforms it to a different type and [I] assume this one wants to use memory and looking at the ones above is non-numeric value. (26: 00)

(P940) [The first thing I look at] is, I'm trying to understand what kind of data set it is...I see that there's a lot of these kinds of lists to understand what the person who wrote this code is doing...[since] there is no comments (4: 10)

(P451) Researcher asked: "what are some things that you might more open or closer to the top of the book of code?" Participants replied: "I mostly relied on previous knowledge and then context wise of the words." (10; 21)

(P948) Even though I don't know what the variable is, I think it's a good variable name, because it's descriptive enough that I can kind of get a sense of what it's doing. (19; 35)

(P948) Yeah, yeah. [I see the large dataset] to double check assumptions about the data

(P312) It is a little bit difficult for me because there's new functions I never used and it's complicated.

(P948) The graphs I worked with so far are for graphs or bar charts, which I think that's the best we did in data science.[0; 56; this is different kinds of graph I'm seeing. I thought first it was a line graph but with color shading. I'm not sure what I was thinking. (27; 00)

(P907) So when I was looking at task zero here. I'm trying to run through the code in my head

(P907) So you're just filtering here. And if I run a save real quick, I can kind of visualize what this might look like, but if I run save right here and then I run save right here and so I can see a before and after [...] I think I'll be able to get a good sense of this.

Expert DO CONFIRM THEIR THOUGHTS

Despite the fact that experts can do head compilation, they always confirm their through the visualization of datasets either at the intermediate steps or the final results. On the other hand, novices like to make assumptions without any confirmations

(P907) I can take a snapshot of that [data set] before an operation and then I could take a look at what it's like after an operation and that's exactly what I'm doing here to be able to confirm my beliefs about what is actually happening to the data

(P451) I think for this instance, I approached [the code] by just making an educated guess on what melting would be. (44; 48)

(P956) Yeah, yeah. [I see the large dataset] to double check assumptions about the data

(P308) I'm just going to look at the data set that I want to make sure, Look at the whole thing because I'm wondering...where is that information.

(P504) I think I know the other differences, but let's see [more saved1 function]

(P907) I have no idea what the actual dataset even looks like, and then I'm applying all these steps to it. So I don't have a better understanding of just what this is doing in general, certain steps that I may not understand. ...[looking at the data] would help me to confirm a lot of my beliefs.

(P956) you can kind of intuitively know, 'okay, so I know' what the first element's supposed to be and I know that this would be the number that would be surprised after the second element. So I know the 'word' - if that makes sense. It's confirming to me that what I think is happening does indeed worked here

(P941) [I did not use visualization in understanding this chunk of code because] I guess either would work, but it seems like comparing visually takes longer (17; 35)

(P256) [I printing out this set of all the values within this column, none of which are the string 'false'. And for it's manipulations, I that's what it's looking for, but it doesn't exist.

(P256) this is more of a debugging thing, [...] kind of looking at things and making sure each line of code does what you want it to do

(P256) [I] print[ed] out the set of the values in the column to make sure that that value exists here. And then that just kind of confirms my assumption of what I thought this function did was correct

(P256) And then we're creating a pivot table by counting the values of that time within the bins that were created by directions. So I'm just looking to check that by printing that out

(P941) The saving function is actually saving me from doing the unnecessary work...plus it's also visualizing what I am thinking in my head with that piece, which gives me a little flow to the understanding. (19; 33)

(P451) I'm assuming somebody would be the same as method. (21; 53)

(P948) researcher asked: "walked me through your thinking process." Participants said, "I try to visualize, but, almost all the time...I'm trying to figure out the backstory to what we're doing with the data" (22; 40)

(P940) [After going through a large chunk of code], the visualization helps me to confirm my assumption. (25; 22)

(P720) Researcher asked 'how did you figure out this chunk of code?' Participants said, "Documentation, Google and failure and the results of ship. There are also [columns] on time, speed, and wind direction..." (46; 29)

(P451) I think I'm a visual person, so the save function was especially helpful for me sometimes, because I could see the super-familiar with all these terms. (50; 33)

(P940) I just skimmed through this part[the actual code], I guesses these were missing, so that's true. And this is the graph. (31; 33)

Different approaches to DataFrame visualization

Both experts and novices reply on visualizing the DataFrame when they are puzzled toward the complex operations. Experts mainly try to figure out the functionality of individual functions via DataFrame visualization, on the other hand, novices compare tables to draw a final conclusion or even without a clear purpose

(P256) I see print because I can see exactly what [...] these operations do right after each other. I can see the information change right after each other

(P451) (What visualizing table,], made a fast comparison and first examining closer to see if there was any other similar things that happened (1h: 00)

(P256) I want to see what it_h_props is [...] I don't know what .dtvi() does. I'll look it up.

[d_h_props uses .dtvi(), she tries to use the visualization to inform what .dtvi() does]

(P312) I would like to see how, before I apply this filter, what's the original one[...] compared to what's the filtered one so I can see how it works

(P256) I still don't know what melt does. Let's look at the original

[Instantly isn't referring to documentation]

(P451) For me it was easier to notice it by looking at the frame because this very first row of values took up a lot of space in the output HTML. So I noticed it when it was missing in the next one (9: 30).

(P451) I was initially just switching fast between them just so I could see what I would notice as the difference (30: 30).

(P241) Since I don't have a lot of experience [visualizing large data set helps a lot for the most part]. We can make comparisons between step and step. (33: 40)

(P241) [When I was] visualizing, I [...] don't know[, why are some of these so long]? And then I just made that comparison from this one. (30: 00)

(P451) From seeing this outthe DataFrame Visualization[, I can tell] that there's a lot of unnecessi because it has a bunch of NaN's. (1: 14)

(P451) (P241) every time I saw a column name, I kind of knew that it was going to be a result of a transformation on that column, so I was especially looking at differences in that column because I knew if the column name didn't appear in that line of code, then nothing's happening to it. (39: 00)

(P720) I want to see what it_h_props is [...] I don't know what .dtvi() does. I'll look it up.

[d_h_props uses .dtvi(), she tries to use the visualization to inform what .dtvi() does]

(P956) it's nice to break it down step by step for interpretability. And then I can kind of see what part helps you understand what it does and then breaking it down to the next chunk, and then the next chunk makes it more interpretable.

(P256) I'm printing out this set of all the values within this column, none of which are the string 'false'. And for it's manipulations, I that's what it's looking for, but it doesn't exist.

(P312) I would like to inspect [the table] every time I do the manipulations

(P312) It is a little bit difficult for me because there's new functions I never used and it's complicated.

(P312) I want to know every intermediate step, so I need to comment out and very code to, to see what's the result of the variable names, but it's still very tedious. And then as the dataset gets bigger and the notebook becomes a mess. It contains so many unnecessary outputs trying to bring out those steps

(P312) Even though I think I understand that, I was trying to make sure that everything's right here

(P312) I think printing it out, I can guarantee that the code is no problem, like no errors.

(P907) it helps elucidate what's happening programmatically by just being able to go in and look at what's in the data, the numbers doing what I "want" or "is there anything that's off"?

(P907) 'How useful did you find the sub functions in your understanding of the code? It was quite helpful here. I would say that solid six behavior. that's what save deserves here, because I was able to take those snapshots and I was able to confirm my beliefs about certain things in the data.

(P241) [Was] visualizing table[,] I made a fast comparison and first examining closer to see if there was any other similar things that happened (1h: 00)

(P241) [When I was] visualizing, I [...] don't know[, why are some of these so long]? And then I just made that comparison from this one. (30: 00)

(P451) From seeing this outthe DataFrame Visualization[, I can tell] that there's a lot of unnecessi because it has a bunch of NaN's. (1: 14)

(P720) Researcher asked: "why did you compare these two tables?" Participants responded by saying, "I wanted to see what exactly did the first task do." (15: 02)

(P907) 'TIME != TIME'. How I'm interpreting this right now is 'I'm looking for a value that's looking for a word "TIME" here, and it seems like, when, it seems like what this test file is all about is they've stitched multiple tables together vertically so that they have the same column names and what they want to do is they want to remove basic such that 'time["TIME"] != TIME' [...] they're excluding all rows that have a header so that it just concatenates it into one long table.

(P907) I'm not going to pretend like I know what melt means, but 'groupby("variable").[dtavi]', it seems like you're doing more grouping and then you're concatenating it altogether. Perhaps I could check and see what that looks like [...] So this right here is _no_dtavi, I think it might be time to run save again. I'm going to see what this method looks like.

P451
P940
P241
P720