Documentation

Documentation occurs at multiple levels.

Project Master Document – You might have a project master document with the goals or aims of the project, information on who has contributed to the project, funding sources, as well as other source of support that you might want to acknowledge in publications or presentations. Please remember to acknowledge the PRC Center and Training infrastructure grants when appropriate.

Notebook – a file with running notes on findings and rationale for research decisions. This could be helpful when you go to write the methods section of your paper.

Code Master Document – a document that provides instructions on how to run the code to create your analytical data files and run your analyses. This can be helpful to your future self, collaborators, and to individuals interested in replicating your analysis. I typically call this file README.txt.

Within code documentation – each script should have at the top a description of what the script does and either what project or paper it is for or the repository name where it is stored. Some people also have a note identifying the authors of the script. Some scripts are short, but others are complicated. For complicated scripts you might break it into sections with a description of the function or purpose of each section. Finally, sometimes code is straightforward and requires little explanation; other times it gets complicated or confusing. If someone looking over your shoulder wouldn't likely immediately understand what you are writing, add a few line-level notes of explanation. Your future self will likely appreciate it, but also explaining your logic might save you from making as many errors. Here is a place where collaborators can be helpful. In fact, I heard about on twitter (but can't currently point to) studies find that collaborating on code is more efficient and less error prone than double coding (i.e. having two people code independently to see if they produce the same results). Having one person write the code and document it well enough that a collaborator can understand it is one model. Another would be for the collaborator to write the documentation.

As part of your documentation, include a variable label describing variables, especially those with non-intuitive names. Also, add variable labels for variables that you will keep in your analysis files.

Some web resources for documentation:

https://blogs.oracle.com/datascience/how-to-write-production-level-code-for-data-science-projects

https://towardsdatascience.com/why-you-should-document-your-work-as-a-data-scientist-a265af8a373

https://medium.com/@andrewgoldis/how-to-document-source-code-responsibly-2b2f303aa525