

Topic: Poor Woman's Documentation of an ETL Pipeline

Portfolio Projects
with
Pizzazz



Welcome to Today's Lecture!

Today's Lecture

- What is ETL doco?
 - What is “doco”?
 - What is “ETL” (especially in a research context)?
- How do you make it?
 - Can use PowerPoint. Takes practice.
 - A few use-cases, and thoughts on best practices.

Explore our online learning resources!

Visit us: <https://buff.ly/45amjaQ>

- Online tutorials
- Video demonstration
- Online courses
- Mentoring program
- ...and more!



What is ETL Doco?

- “Doco” is short for “documentation”.
 - Documentation includes data curation files
 - Documentation includes anything seen as “metadata”.
- “ETL” is short for “extract, transform, and load”
 - Common terminology in data warehousing
 - Not really used much in research (although this is changing)

Resources

Use-case: Keeping track of NHANES datasets in ETL for an analytic research dataset: <https://dethwench.com/nhanes-datasets-for-use-in-data-analysis/>

Some overall best practices when keeping ETL doco:

<https://dethwench.com/etl-pipeline-documentation-best-practices/>

More ETL use-cases:

<https://dethwench.com/keep-dataset-source-documentation-to-make-science-replicable/>



Flow Charts

- **Poor woman's documentation:** Use PowerPoint
- **Other software:** MS Visio, Gliffy, other online apps
- Flow chart shapes can be useful for communication
- But you can put other shapes in there (clip art, icons, shapes you make up)
- Include minimal text to help understanding
- **Goal:** Visual communication

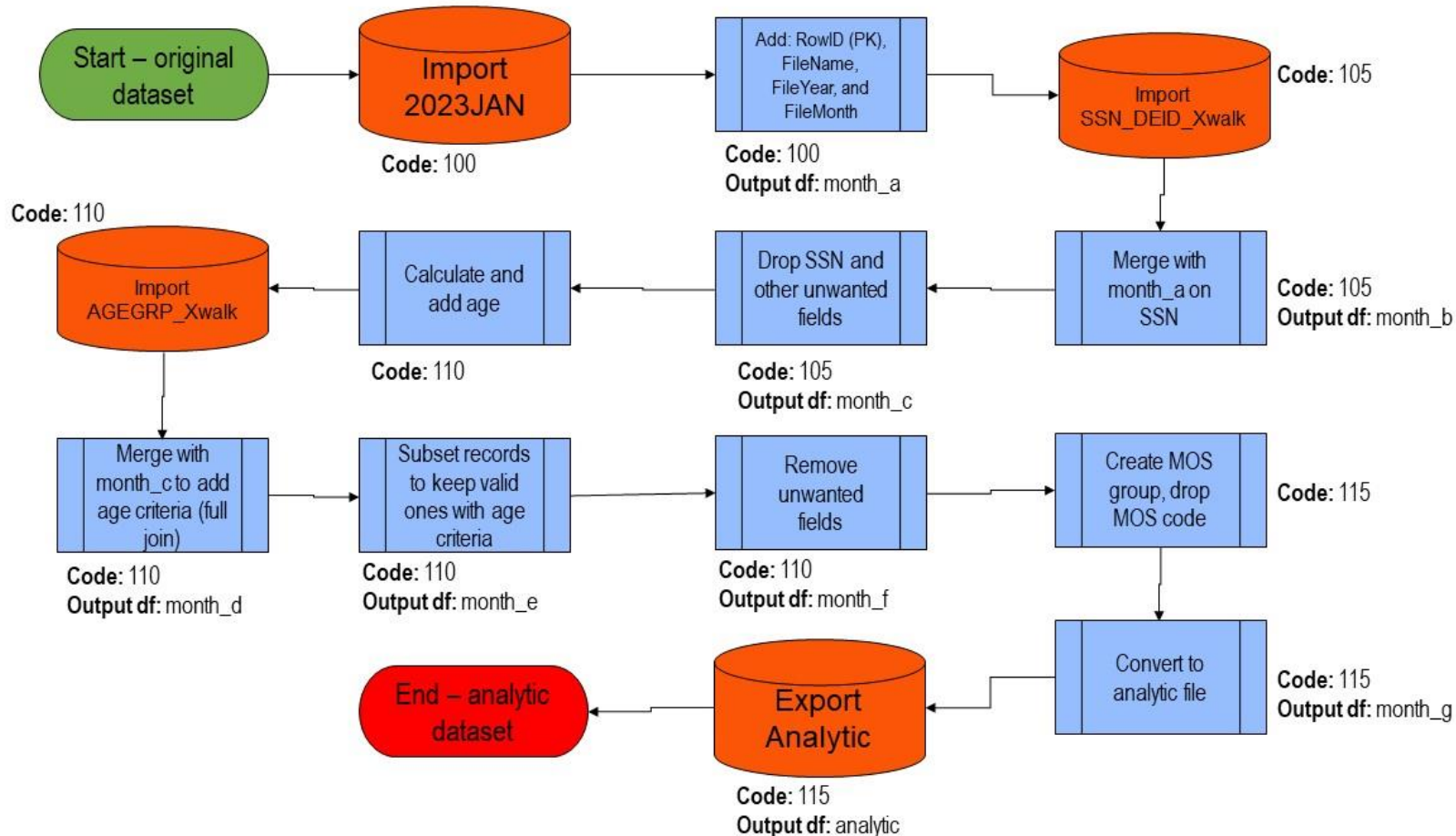
More Resources

Blog post about flow chart shapes:
<https://dethwench.com/how-to-use-shapes-and-images-in-data-visualizations/>

“PowerPoint for Visualizations” online course: <https://monika-s-school-e295.thinkific.com/courses/powerpoint-for-visualizations>



Example of Warehouse-Style ETL Doco

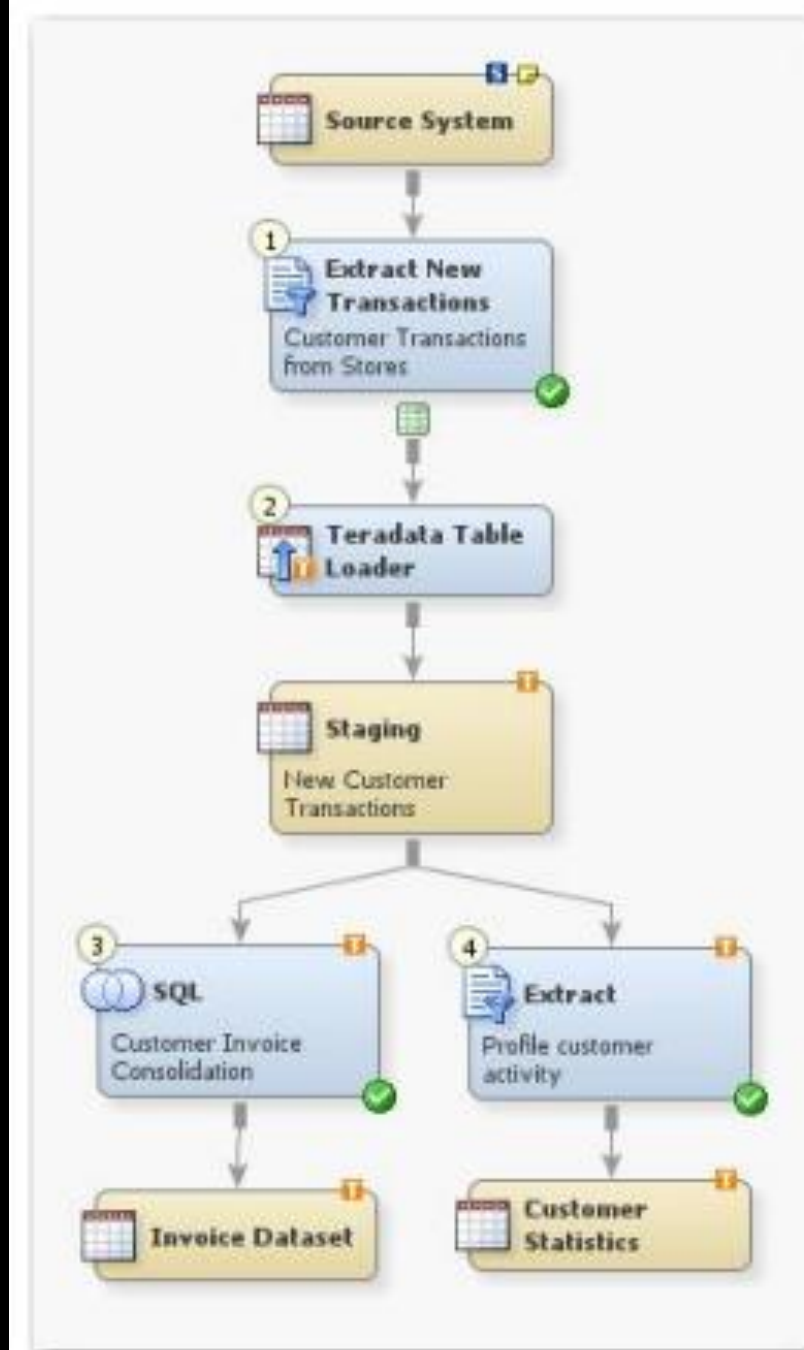


- Documents each procedure
- Includes number for code
- Should you “always” bother to make a diagram like this?



SAS Data Integration Studio Diagram

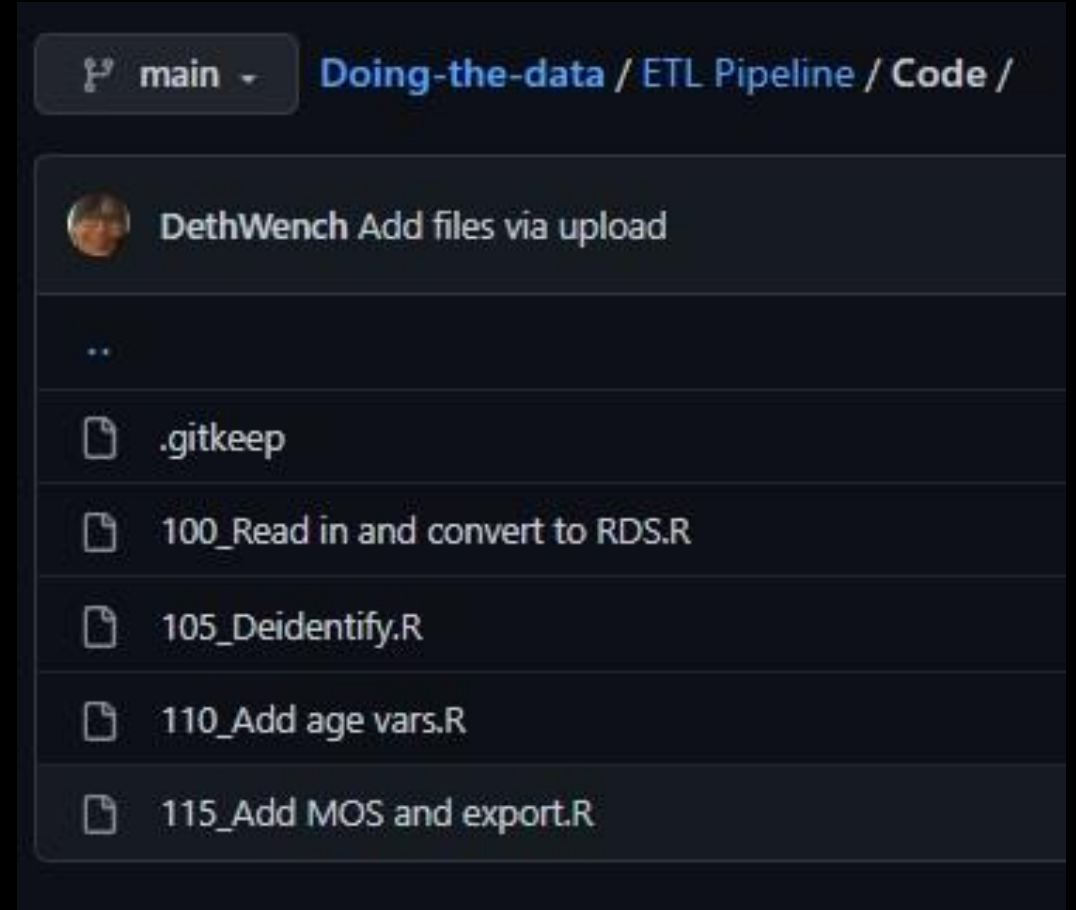
- From SAS White Paper, “SAS Data Integration Studio: Tips and Techniques for Implementing ELT” By Jeff Stander:
<https://buff.ly/44E9rKQ>
- E-L-T – not E-T-L 🤖
- This is a real pipeline – but to build this pipeline, you first have to plan it





Code of ETL step in research

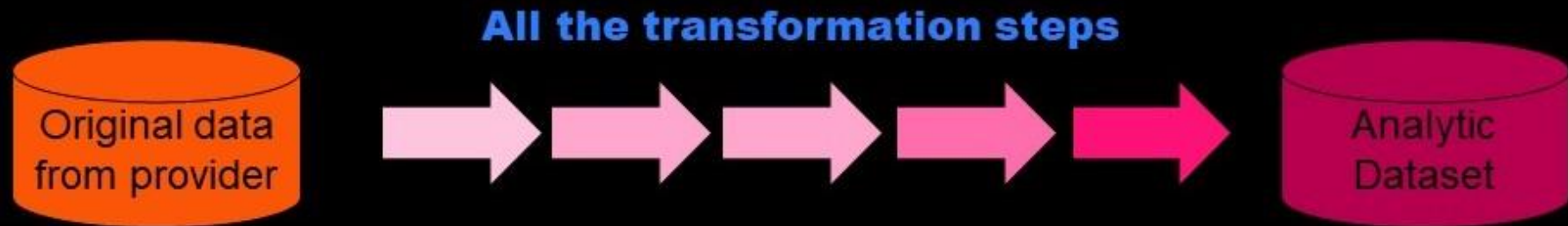
- You need to make sure you run the code in order, so why not prepend numbers?
- Modular code that does only one or a few things in each step
 - “deidentify”, “add age vars”, etc.
 - Read in *df_a*, export *df_b*. Next, read in *df_b*, export *df_c*.
 - Very clear.
 - Also, easy do to a rollback if a repair is needed.





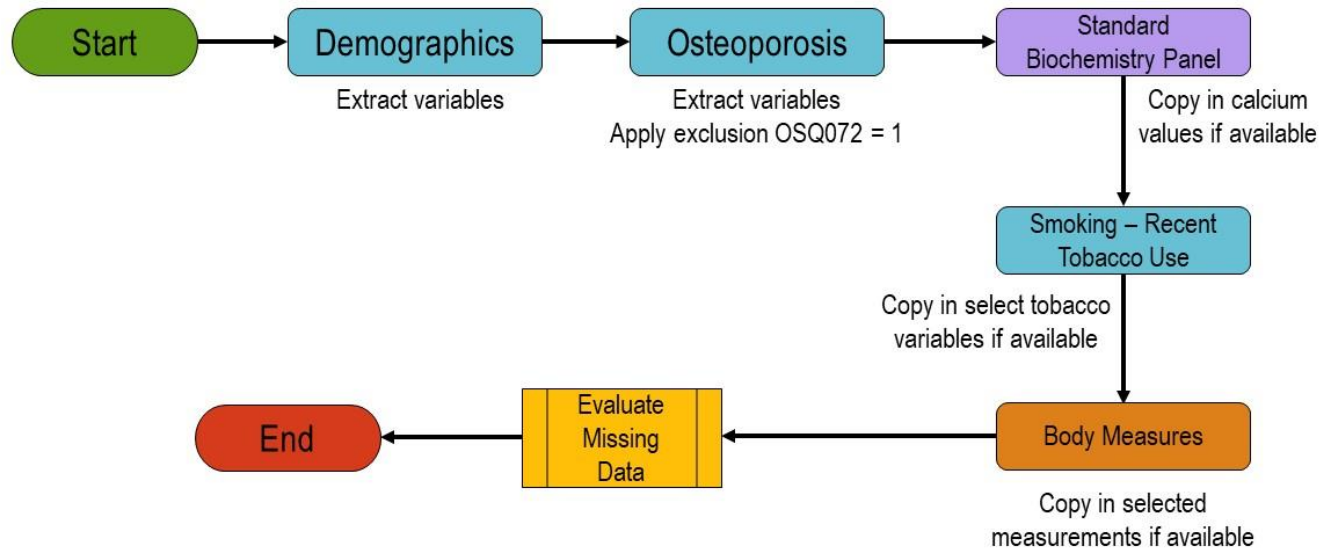
What does ETL mean in a research context?

- Need to be able to document each transformation step from the original dataset to the final analytic dataset used in the research analysis
- Should be replicable (by another statistician)
 - Even if you did it in SAS, if you document each stop, they could rewrite it in R, or vice versa





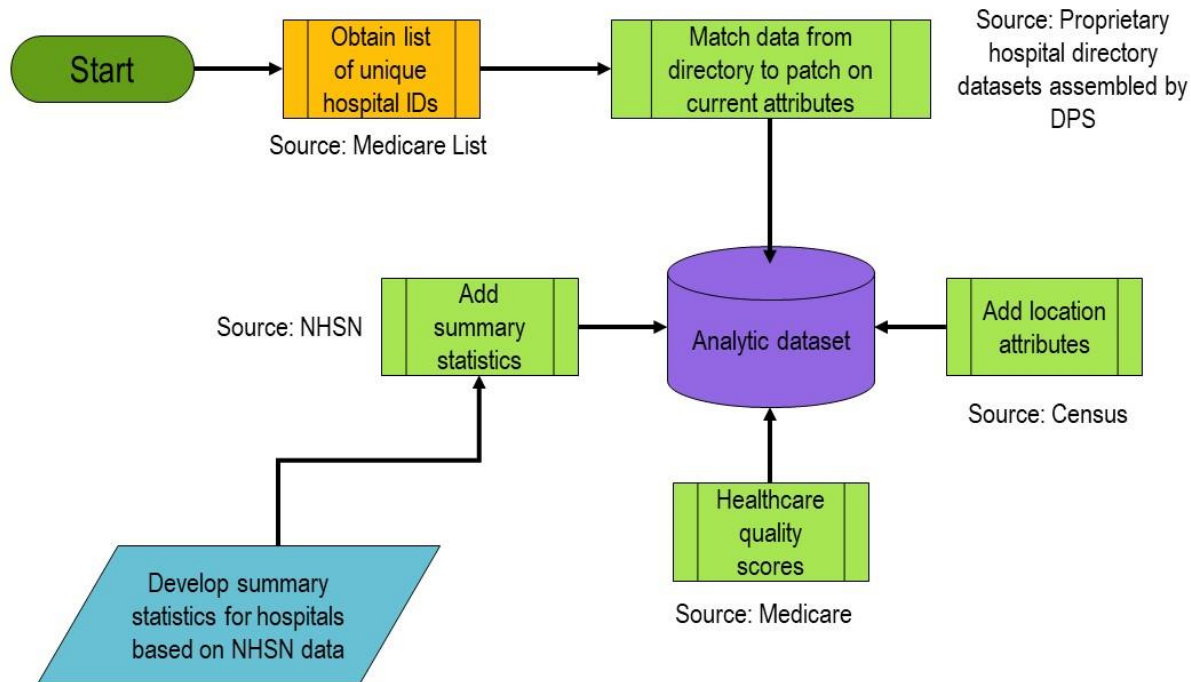
Research ETL Doco: NHANES Example



- Can be kind of simple
- The person who will love it the most in the world is **YOU!**
- The person who will love it 2nd most in the world is **your boss!**



Especially Helpful with Multiple Source Datasets



- Again, if you bother to sit down and make this diagram, the “future you” will be overjoyed to find this in your data curation files
- Notice all the items that you could easily forget (and would have to dig through code to remember)



Questions?

Facilitator: Monika M.
Wahi, MPH, CPH

Email:
dethwench@gmail.com

Blog:
<https://dethwench.com/blog/>

LinkedIn:
<https://www.linkedin.com/in/dethwench/>

Explore our online learning resources!

Visit us: <https://buff.ly/45amjaQ>

- Online tutorials
- Video demonstration
- Online courses
- Mentoring program
- ...and more!