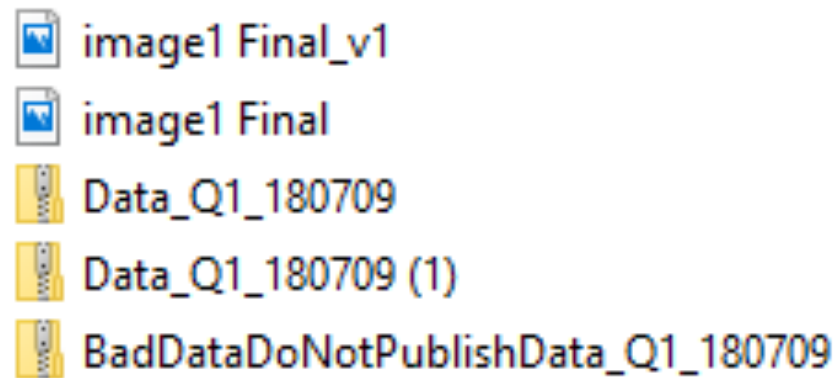




Organize and Document

What's in a Name?

Can you tell what these are without opening them?



Use *filenaming* to organize and document research files

Quick Tips for File Naming

File naming tip	Poor names	Good names
Be specific	Image1 UseThisOne_v.2 Final_LungCancerMS	StemCell_SMA WillowCreek_SpList_2012 PerceptionExp_Subj1

Quick Tips for File Naming

File naming tip	Poor names	Good names
Be specific	Image1 UseThisOne_v.2 Final_LungCancerMS	StemCell_SMA WillowCreek_SpList_2012 PerceptionExp_Subj1
Be consistent	Data_v1 ResearchData_v2 Results_v3	Azaleas_Stem Azaleas_Pollen Azaleas_Petal
Use certain characters (Stick with letters, numbers, -, _ and avoid spaces and special characters)	Perception Exp: Survey Rhododendron[Plot1] StemCell.SMA.15A*	Perception_Exp_Subj1 RhododendronPlot1 StemCell-SMA-15A
Use Standard Date/Time Format (YYYYMMDD hh:mm:ss)	April_10_2018 04102018	20180410 2018-04-10

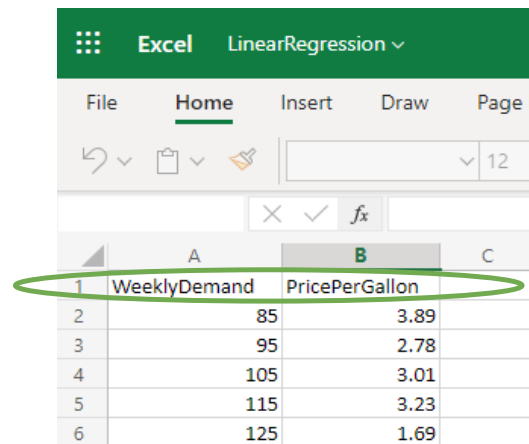
Folder, Variable and Other Names

The same file naming tips can also be applied to:

Folder names



Variable names



	A	B	C
1	WeeklyDemand	PricePerGallon	
2		85	3.89
3		95	2.78
4		105	3.01
5		115	3.23
6		125	1.69

Function names

```
function compare(a, b)
{
    return a == b;
}

var are_equals = compare(3, 5);
```



Documentation

What is it?

- Data that provides **descriptive information**
- Also known as **metadata**

What gets documented?

- Data, procedures, code, variables and values, derived data, restrictions on use of data, etc.

Why should I do it?

- Enables other to search for, trust, and **reuse** your data
- Critical for research **reproducibility**

**How much detail should I add to
my documentation?**

Enough for your future self or collaborators



Scenario: Documentation

You just came back from attending an American Heart Association conference. You learn about a similar study to yours at the University of Maryland and ask the researcher if she will share preliminary data with you. She sends you some de-identified data in a spreadsheet.

But when you see the data, you have a few questions...

<u>Patient #</u>	<u>Height</u>	<u>Weight</u>	<u>Ex. Dur</u>	<u>HR</u>	<u>PEF</u>	<u>Location</u>
154398			100	70	640	MD21218
582394					300	MD21044
814293	187	87				MD20770
39201		17				MD21202
17829		17	54	90		MD21218
239482	175	45	40	94	300	MD21001
403291		1000		96	360	MD21010
290300	175	97	33	70	490	MD21014
770543		62	43	65	510	MD21022
125765	170	50	88	98	340	MD21218

Does this patient code refer to PHI?

What is the unit for each variable?
What does each variable mean?

Aren't these identifiers that need to be removed?

What does this missing value mean? No exercise or missing record?

Is this value correct?

Am I looking at the correct sheet?

Example Documentation: Data Dictionary

	A	B	C	D	E
1	Variable Name	Variable Label	Value	Value Label	Notes
2	Patient #	Unique patient ID	Randomly-generated ID. No PHI.
3	Height	Height measurement of the patient in centrimeters	Measured in every patient's visit
4	Weight	Height measurement of the patient in kilograms	Measured in every patient's visit
5	Ex. Dur	Daily exercise duration in minutes	Exercise duration is recorded by patient's wearable device from the beginning to the end of an exercise.
6	HR	Heart rate (beats per minute)	Heart rate is the peak measured by patient's wearable device during exercise
7	PEF	Peak expiratory flow (L/min)	Every patient uses the same model of peak flow meter issued by the provider to measure their peak expiratory flow. The measurement is taken first thing in the morning and follow the steps here: https://www.hopkinsmedicine.org/health/treatment-tests-and-therapies/peak-flow-measurement
8	Location (identifier, delete or recode before sharing)	Zip code for patient's home	MD	Maryland	Official postal abbreviations for each state in USA (https://pe.usps.com/text/pub28/28apb.htm). Zip code can be found here: https://tools.usps.com/zip-code-lookup.html .
9			DE	Delaware	
10			VA	Virginia	
11			
12			
13			blank	missing value	

data metadata

Separate Metadata Sheet

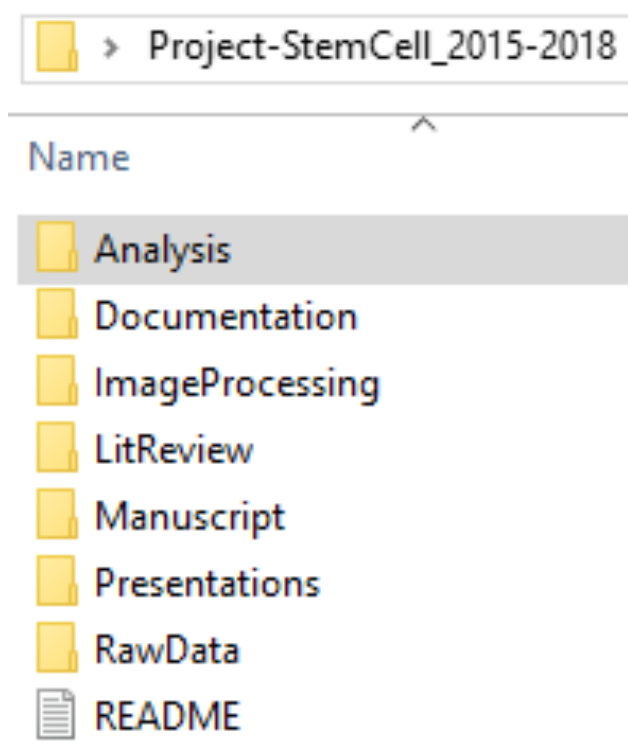
File Level Documentation

	A	B	C	D	E
1	Variable Name	Variable Label	Value	Value Label	Notes
2	Patient #	Unique patient ID	Randomly-generated ID. No PHI.
3	Height	Height measurement of the patient in centrimeters	Measured in every patient's visit
4	Weight	Height measurement of the patient in kilograms	Measured in every patient's visit
5	Ex. Dur	Daily exercise duration in minutes	Exercise duration is recorded by patient's wearable device from the beginning to the end of an exercise.
6	HR	Heart rate (beats per minute)	Heart rate is the peak measured by patient's wearable device during exercise
7	PEF	Peak expiratory flow (L/min)	Every patient uses the same model of peak flow meter issued by the provider to measure their peak expiratory flow. The measurement is taken first thing in the morning and follow the steps here: https://www.hopkinsmedicine.org/health/treatment-tests-and-therapies/peak-flow-measurement
8	Location (identifier, delete or recode before sharing)	Zip code for patient's home	MD	Maryland	Official postal abbreviations for each state in USA (https://pe.usps.com/text/pub28/28apb.htm). Zip code can be found here: https://tools.usps.com/zip-code-lookup.html .
9			DE	Delaware	
10			VA	Virginia	
11			
12			
13		blank	missing value		
14					
15					
16					

Examples

- Codebook/Data dictionary to define values in a spreadsheet
- README explaining how to run a code file

Project Level Documentation



Examples

- Author's name, PI's name, file location, etc.
- Permanent identifier, such as DOI for your dataset
- Written description of a dataset, such as a README for a project

Project Level Documentation in a README

Project: The effect of mild exercise on lung cancer surgery recovery

Funder and grant number: NSF Grant # BIO-12345678

PI(s): Dr. Ama Nobel, Johns Hopkins University

Dates: August 2021 to August 2026

Name and location of key files:

Code – <https://github.com/exercise-cancer> and published in JHU Data Archive (doi: 10.7281/T10Z715B)

Protocol – published Nature Protocols (<https://doi.org/10.1038/s0587-245-01>)

Data – exerciseSurvey_de-id.zip published in JHU Data Archive (doi: 10.7281/T10Z715B)

Codebook – In same zip file as data

File naming convention:

Dates recorded as YYYYMMDD

All files should start with exercise name, then research team's name, whether it is raw or processed, and date

Use standard abbreviations for variable names

Resource: [ReadMe file template and best practices](#) by Cornell University

A Tip for Documenting Data

Use **standards in your research fields** when available or develop your own standards!

What are YOUR standards?



NIH Common Data Elements: <https://www.nlm.nih.gov/cde/>



NIH LINCS
PROGRAM

Standards: <https://lincsproject.org/LINCS/data/standards>

LIBRARY OF INTEGRATED NETWORK-BASED CELLULAR SIGNATURES



Metadata Standards Directory Working Group:
<http://rd-alliance.github.io/metadata-directory/>



FAIRsharing metadata standards: <https://fairsharing.org/standards/>