Introduction to Research Data Management

Data documentation

LIBRARIES
TEXAS A&M UNIVERSITY

Introduction

Focus on documenting your data to help you

- share information about your project with your advisor,
- write your thesis or dissertation,
- and ensure your data are useful later.

Why document data

Documenting your data enables:

- You to get a mess of details out of your head.
- You to reproduce and improve workflows in the future.
- Your collaborators to find the right data and use them properly.

Discussion

As you are working on a research project, what sort of information do you document?

How do you document that information?

What to document

High-level documentation explains your research goals and the progress of your project.

Low-level documentation explains the details of the data, how it has been collected, stored, and changed over time.

Project documentation

- Rationale and context for data collection.
- Research questions, goals, and hypotheses.
- Data sources, collection methodology, protocols.
- Data validation and quality assurance actions.
- Transformation of raw or derived data for integration or analysis.
- Data confidentiality, access, and use conditions.

Dataset documentation

- Variable names and descriptions.
- Codes and classification schemes.
- Algorithms used to transform data.
- Structure and organization of files.
- Relationship among data files or tables in a database schema.
- Version information.
- File formats and software used.

How to document

A few documentation tools:

- Laboratory and field notebooks
- README files
- Codebooks

Notebooks

Common in laboratory settings and fieldwork.

Document context, project, and dataset level information.

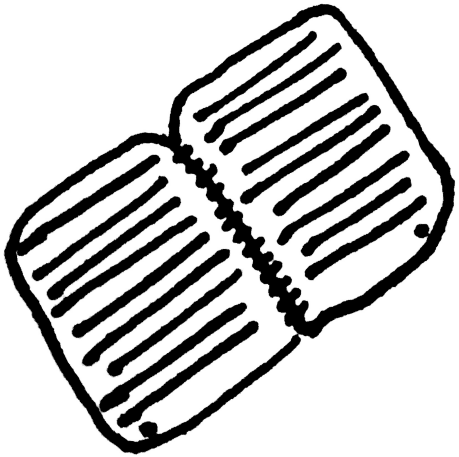
Give you one place to go throughout the research process.

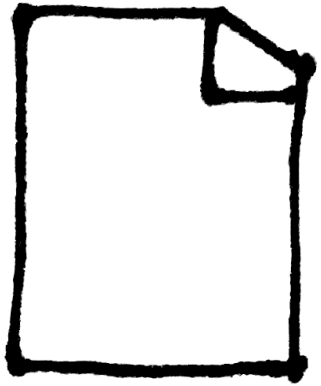
Encourage a thoughtful examination.

Enable continuity in case of unexpected events or time passing.

Establish a legal and scientific provenance (historical record) of your work.

Analog and digital notes





110

~~3 Apr~~ 1-May 1999 Exp 30 contd.

Results of Ligation/Transformation
– Colonies counted

Plate
① – Negative ctrl – 2 colonies
② Ligation ① – 0
③ " ② – 28 – YES
④ Positive control ~300
⑤ – No Amp – Confluent growth.
(test for competent bacteria)

– OK – Have some "colonies" on ①
=> Suggests need to check colonies from ③ for recombinant clones.

– Should have done a ligation control!!

– Will need to do a PCR to check colonies from ③

111

1 May 1999 Exp 31

PCR of colonies from Exp 30 – to [illegible] recombinant clones

– To see which clones are recombinant, will do PCR using primers to MCS outside insert – will get ~150bp fragment if no insert – and ~1.1Kb fragment if 900 bp insert is present.

– Stab colonies with wire loop – transfer to 20µl water in 500µl eppendorf (only did 18 colonies).

PCR Reaction setup	Exp No. - 31	Date - 1-May-99	No. Reactions - 18
	Name	µl per reaction	µl total
Buffer	5x - [illegible]	2	36
dNTP's	[illegible]	—	—
$MgCl_2$	25mM [illegible]	1	18
Primer 1	T7 - [illegible]	1	18
Primer 2	SP6 - [illegible]	1	18
TAQ	[illegible]	0.2	3.6
Other			
Water		4.8	90
TOTAL VOLUME		10	183.6

+ 2 extra!

Reactions set up with 9µl of master mix (above) and 1µl of each colony.

PCR conditions
92°C – 5 minutes
92°C – 10 sec.
58°C – 20 sec } 35 cycles
72°C – 60 sec
72°C – 10 minutes
15°C – hold

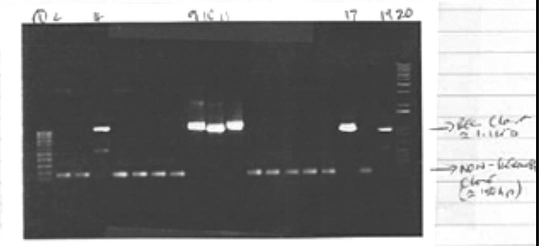
Reaction started @ 10.53 – estimated end @ 12.47

112

1 May 1999 Exp 31 contd

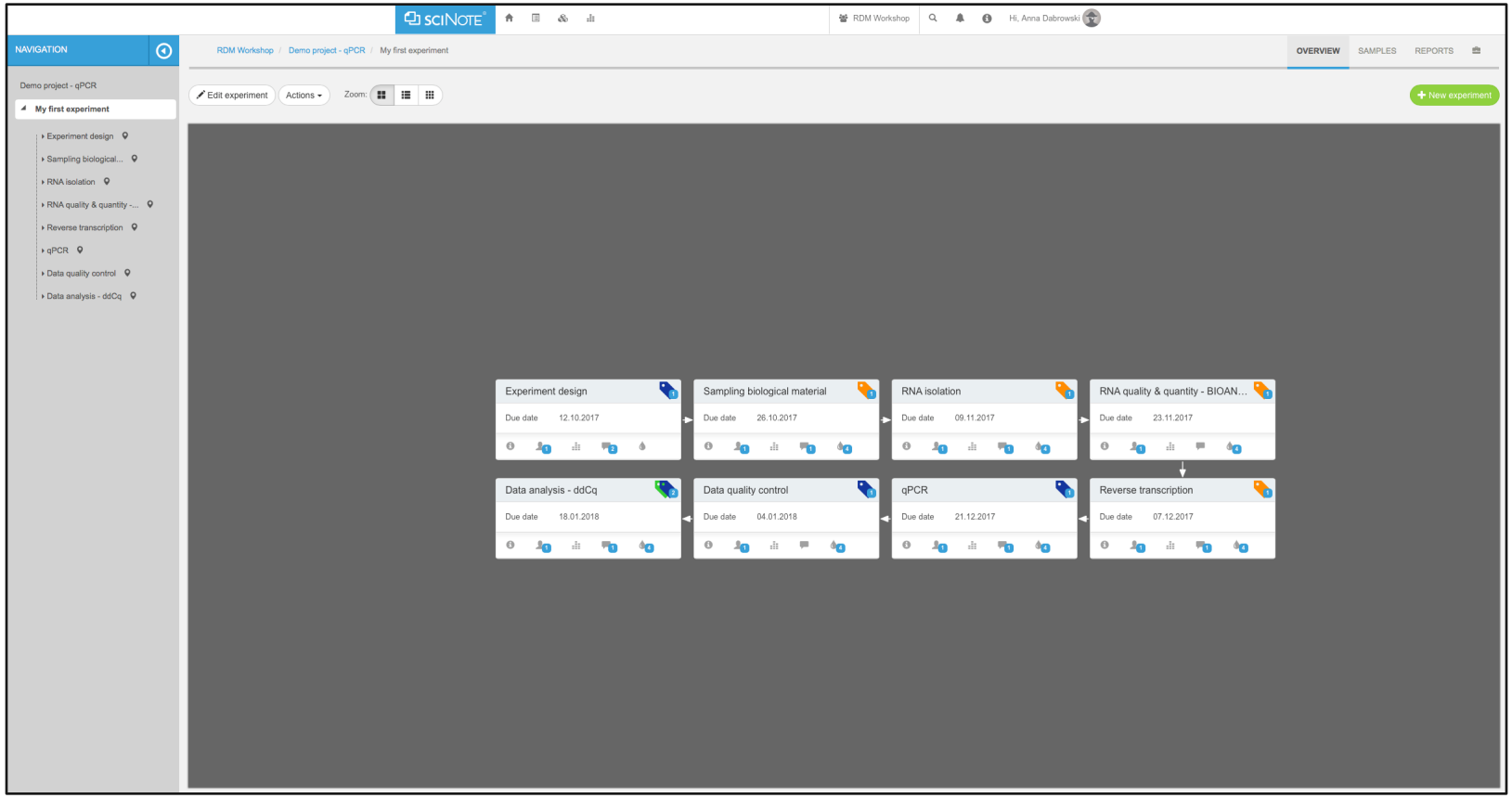
0.8% agarose gel in TAE

Lane 1 – 100bp ladder
2-19 – Sample from PCR reaction (+ 2µl gel loading buffer)
20 – 1kb plus ladder
– 100V for 60 minutes.



– Pretty good! – 6 recombinants.
– [illegible] looks strange – a [illegible]
– Select 4, 9, 18, 17 for sequencing.
– Add 100µl LB + Amp to the colony in the eppendorf tube.

Note-taking software


SCINOTE
RDM Workshop
Hi, Anna Dabrowski
NAVIGATION
RDM Workshop / Demo project - qPCR / My first experiment
OVERVIEW
SAMPLES
REPORTS
Demo project - qPCR
My first experiment
Experiment design
Sampling biological...
RNA isolation
RNA quality & quantity -...
Reverse transcription
qPCR
Data quality control
Data analysis - ddCq
Edit experiment
Actions
Zoom:
New experiment
Experiment design
Due date 12.10.2017
Sampling biological material
Due date 26.10.2017
RNA isolation
Due date 09.11.2017
RNA quality & quantity - BIOAN...
Due date 23.11.2017
Data analysis - ddCq
Due date 18.01.2018
Data quality control
Due date 04.01.2018
qPCR
Due date 21.12.2017
Reverse transcription
Due date 07.12.2017

2017-05-13

Search All Notebooks

Home | Insert | View | Audio

Paste | Cut | Copy | Format | Calibri | 11 | B I U abc X_2

Heading 1 | Heading 2

To Do | Remember for later | Contact | Important | Definition | Address | Question | Highlight | Phone number | To Do

Experiment 1 | New Section 1 | +

2017-05-13

ajd-interview

Tested interview questions with subject (id: ajd) and recorded the interview. Participant commented that question 7 was confusing.

Add Page

2017-05-13

2017-05-21

Experiment 1 — Evernote

AJDABROWSKI@LIBRARY.TAMU...

New Note in Experiment 1 | New Chat | Search notes

Shortcuts

Experiment 1

Notes

Notebooks

Tags

Trash

Upgrade

All Notes | E...

SEPTEMBER 2017 2

2017-08-29

Today Spikey the plant At 16:00 the plant is alive.

2017-08-30

Today Spikey the plant At 16:00 the plant is dead.

Experiment 1 | plant study

Share

Created: Sep 30, 2017 Updated: Sep 30, 2017

2017-08-30

Spikey the plant

At 16:00 the plant is dead.

The matplotlib object-oriented API

The main idea with object-oriented programming is to have objects that one can apply functions and actions on, and no object or program states should be global (such as the MATLAB-like API). The real advantage of this approach becomes apparent when more than one figure is created, or when a figure contains more than one subplot.

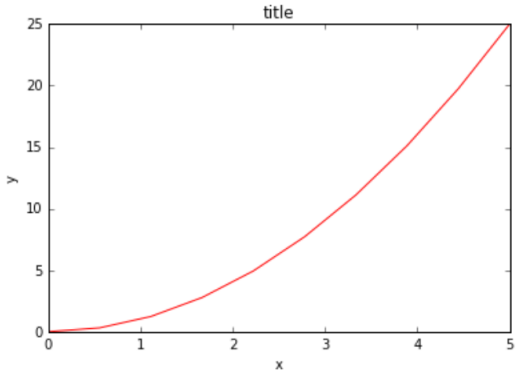
To use the object-oriented API we start out very much like in the previous example, but instead of creating a new global figure instance we store a reference to the newly created figure instance in the `fig` variable, and from it we create a new axis instance `axes` using the `add_axes` method in the `Figure` class instance `fig`:

In [9]:
```
fig = plt.figure()

axes = fig.add_axes([0.1, 0.1, 0.8, 0.8]) # left, bottom, width, height (range 0 to 1)

axes.plot(x, y, 'r')

axes.set_xlabel('x')
axes.set_ylabel('y')
axes.set_title('title');
```



```
<h3>Regression Multiple Gaussian Targets</h3>
Now *assume* that we have $K$ Gaussian distributed target variables, $\[t_1, \ldots, t_K\]$, each with a mean that is
independently conditional on $\mathbf{x}$, i.e. the mean of $t_k$ is defined by some function $\mu_k(\mathbf{x})$.
Also assume that all $K$ variables share the same variance, $\sigma^2=1/\beta$. Assuming the network output layer has
$K$ nodes where $y_k(\mathbf{x},\mathbf{w}) \approx \mu_k(\mathbf{x})$ and letting $\mathbf{y}(\mathbf{x},\mathbf{w})
= \[y_1(\mathbf{x},\mathbf{w}), \ldots, y_K(\mathbf{x},\mathbf{w})\]$, and that we again have $N$ training target
values $\mathbf{t}$ ($\mathbf{t}$ is a $K \times N$ matrix of the training values), the conditional distribution of
the target training values is given by

$p(\mathbf{t}|\mathbf{x},\mathbf{w}) = ND(\mathbf{t} | \mathbf{y}(\mathbf{x},\mathbf{w}), \beta^{-1} \mathbf{I})$

The parameter estimates, $\mathbf{w}^{(ML)}$ are again found by minimizing the sum-of-squares error function,
$E(\mathbf{w})$, and the estimate for $\beta$ is found from

$\frac{1}{\beta_{ML}} = \frac{1}{NK}\sum_{n=1}^N ||y(\mathbf{x}_n,\mathbf{w}^{(ML)}) - t_n ||^2$
```

Regression Multiple Gaussian Targets

Now *assume* that we have K Gaussian distributed target variables, $\[t_1, \ldots, t_K\]$, each with a mean that is independently conditional on $\mathbf{x}$, i.e. the mean of t_k is defined by some function $\mu_k(\mathbf{x})$. Also assume that all K variables share the same variance, $\sigma^2 = 1/\beta$. Assuming the network output layer has K nodes where $y_k(\mathbf{x}, \mathbf{w}) \approx \mu_k(\mathbf{x})$ and letting $\mathbf{y}(\mathbf{x}, \mathbf{w}) = \[y_1(\mathbf{x}, \mathbf{w}), \ldots, y_K(\mathbf{x}, \mathbf{w})\]$, and that we again have N training target values $\mathbf{t}$ ($\mathbf{t}$ is a $K \times N$ matrix of the training values), the conditional distribution of the target training values is given by

$$p(\mathbf{t}|\mathbf{x}, \mathbf{w}) = ND(\mathbf{t}|\mathbf{y}(\mathbf{x}, \mathbf{w}), \beta^{-1}\mathbf{I})$$

The parameter estimates, $\mathbf{w}^{(ML)}$ are again found by minimizing the sum-of-squares error function, $E(\mathbf{w})$, and the estimate for β is found from

$$\frac{1}{\beta_{ML}} = \frac{1}{NK}\sum_{n=1}^{N} ||y(\mathbf{x}_n, \mathbf{w}^{(ML)}) - t_n||^2$$

Good practices and tips

- Date each entry.
- List full names and contact information for collaborators.
- Take notes on meetings and discussions.
- Write justifications for research methods and data source(s).
- Include protocols.
- Capture conditions in the field, lab, interview.

Good practices and tips

- Note mistakes and corrections that need to be made.
- Annotate calculations with units.
- Record relevant digital file names and locations.
- Describe location of physical materials.
- Use images and print-outs to simplify documentation.

README files

Coopted from software development.

Describe the files and folders in a project.

Document mainly at the dataset level.

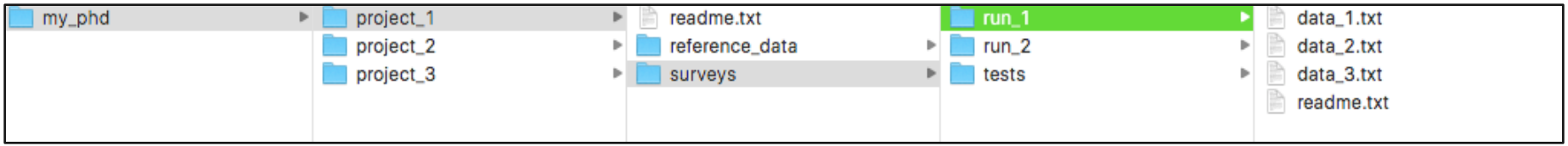
Primarily aimed at an external audience and your future self.

Example: README template

Cornell University.

https://cornell.app.box.com/v/ReadmeTemplate

README tips


my_phd
project_1
project_2
project_3
readme.txt
reference_data
surveys
run_1
run_2
tests
data_1.txt
data_2.txt
data_3.txt
readme.txt

Codebooks and data dictionaries

Document at the level of variables and data within the dataset.

Stand-alone or part of other files.

Primarily aimed at an external audience and your future self.

Example: Tabular data with a data dictionary

Data from the Duke Lemur Center about the weights of lemurs.

- Data file: http://bit.ly/2wtIeuh
- README file: http://bit.ly/2fL4UTH

Example: Tabular data with a data dictionary

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1	Taxon	Hybrid	DLC_ID	Sex	Name	DOB	DOB_Estimated	Weight_g	Weight_Date	MonthOfWeight	AgeAtWt_d	AgeAtWt_wk	AgeAtWt_mo	AgeAtWt_mo_NoDec	AgeAtWt_y
2	CMED	N	1601	M	LINUS	16-Jul-83		77	9-Aug-83	8	24	3.43	0.79	0	0.07
3	CMED	N	1601	M	LINUS	16-Jul-83		242	21-Oct-83	10	97	13.86	3.19	3	0.27
4	CMED	N	1601	M	LINUS	16-Jul-83		258	10-Nov-83	11	117	16.71	3.85	3	0.32
5	CMED	N	1601	M	LINUS	16-Jul-83		337	8-Dec-85	12	876	125.14	28.8	28	2.4
6	CMED	N	1602	M	LUCAS	16-Jul-83		79	9-Aug-83	8	24	3.43	0.79	0	0.07
7	CMED	N	1602	M	LUCAS	16-Jul-83		176	25-Jul-93	7	3662	523.14	120.39	120	10.03
8	CMED	N	1602	M	LUCAS	16-Jul-83		177.7	7-Dec-93	12	3797	542.43	124.83	124	10.4
9	CMED	N	1602	M	LUCAS	16-Jul-83		181	31-Aug-93	8	3699	528.43	121.61	121	10.13
10	CMED	N	1602	M	LUCAS	16-Jul-83		193	26-May-93	5	3602	514.57	118.42	118	9.87
11	CMED	N	1602	M	LUCAS	16-Jul-83		194	13-Aug-91	8	2950	421.43	96.99	96	8.08
12	CMED	N	1602	M	LUCAS	16-Jul-83		194	20-Feb-97	2	4968	709.71	163.33	163	13.61
13	CMED	N	1602	M	LUCAS	16-Jul-83		196	5-Jun-95	6	4342	620.29	142.75	142	11.9
14	CMED	N	1602	M	LUCAS	16-Jul-83		199	25-Oct-91	10	3023	431.86	99.39	99	8.28
15	CMED	N	1602	M	LUCAS	16-Jul-83		200	11-Sep-96	9	4806	686.57	158.01	158	13.17
16	CMED	N	1602	M	LUCAS	16-Jul-83		200	16-Jan-97	1	4933	704.71	162.18	162	13.52
17	CMED	N	1602	M	LUCAS	16-Jul-83		200	4-Apr-97	4	5011	715.86	164.75	164	13.73
18	CMED	N	1602	M	LUCAS	16-Jul-83		206	10-Jul-96	7	4743	677.57	155.93	155	12.99
19	CMED	N	1602	M	LUCAS	16-Jul-83		207	7-Aug-96	8	4771	681.57	156.85	156	13.07
20	CMED	N	1602	M	LUCAS	16-Jul-83		208	7-Nov-96	11	4863	694.71	159.88	159	13.32
21	CMED	N	1602	M	LUCAS	16-Jul-83		213	8-Jul-92	7	3280	468.57	107.84	107	8.99
22	CMED	N	1602	M	LUCAS	16-Jul-83		219	8-Jun-95	6	4345	620.71	142.85	142	11.8
23	CMED	N	1602	M	LUCAS	16-Jul-83		221.7	2-May-84	5	291	41.57	9.57	9	0.8
24	CMED	N	1602	M	LUCAS	16-Jul-83		225	1-Apr-93	4	3547	506.71	116.61	116	9.72

Table 1. List of taxa included in the data files, including taxonomic code used in all data files (Taxon), Latin name and common name of each taxon[4, 5].

count	Taxon	Latin_Name	Common_Name
1	CMED	Cheirogaleus medius	Fat-tailed dwarf lemur
2	DMAD	Daubentonia madagascariensis	Aye-aye
3	EALB	Eulemur albifrons	White-fronted brown lemur
4	ECOL	Eulemur collaris	Collared brown lemur
5	ECOR	Eulemur coronatus	Crowned lemur
6	EFLA	Eulemur flavifrons	Blue-eyed black lemur
7	EFUL	Eulemur fulvus	Common brown lemur
8	EMAC	Eulemur macaco	Black lemur
9	EMON	Eulemur mongoz	Mongoose lemur
10	ERUB	Eulemur rubriventer	Red-bellied lemur
11	ERUF	Eulemur rufus	Red-fronted brown lemur
12	ESAN	Eulemur sanfordi	Sanford's brown lemur
13	EUL	Eulemur	Eulemur hybrid

11	AgeAtWt_d	Age in days: Age of the animal when the weight was taken, in days (Weight_Date-DOB)
12	AgeAtWt_wk	Age in weeks: Age of the animal when the weight was taken, in weeks ((Weight_Date-DOB)/7))
13	AgeAtWt_mo	Age in months: Age of the animal when the weight was taken, in months (((Weight_Date-DOB)/365)*12)
14	AgeAtWt_mo_NoDec	Age in months with no decimal: AgeAtWt_mo value rounded down to a whole number for use in computing average individual weights (FLOOR(AgeAtWt_mo))
15	AgeAtWt_y	Age in years: Age of the animal when the weight was taken, in years ((weight date-DOB)/365)
16	Days_Since_PrevWt	Days difference: Difference, in days, between the date of this weight and the date of the animal's previous weight

Example: Data Documentation Initiative codebook

# BEDRM: Bedrooms, number of	
Information	[Type= discrete] [Format=numeric] [Range= 0-8] [Missing=*/8]
Statistics [NW/ W]	[Valid=872972 / 32343616.174] [Invalid=14040 / 508706.825]
Definition	Refers to all rooms designed mainly for sleeping purposes even if they are now used for other purposes, such as guest rooms and television rooms.
Universe	Reported for: Persons in private households
Notes	Data quality note – In the 2011 National Household Survey (NHS), a large proportion of records with 0 bedroom dwellings and 1 room dwellings was affected by respondent error (such as reporting more bedrooms than rooms). These errors were resolved during data processing and the results are consistent with other surveys. However, it is possible that in some instances or in small geographic areas the processed result is not consistent with the respondent's true situation. For more information, please consult the Housing Reference Guide, National Household Survey, Catalogue no. 99-014-X2011007.

Value	Label	Cases	Weighted	Percentage (Weighted)
0	No bedroom	4889	178016.9	0.6%
1	1 bedroom	65590	2470756.9	7.6%
2	2 bedrooms	169285	6421315.9	19.9%
3	3 bedrooms	333708	12422735.6	38.4%
4	4 bedrooms	215433	7816595.4	24.2%
5	5 bedrooms or more	84067	3034195.4	9.4%
8	Not available	14040	508706.8	

Good practices and tips

- Variable name.
- Variable meaning.
- Variable format and how the variable was recorded.
- Units of measurement for scale variables.
- Numeric codes for categorical variables, and what they represent.
- Known issues and relationships.

Conclusion

- Identified relevant project-level and data-level documentation.
- Reviewed documentation tools including lab notebooks, README files, and codebooks.

References and resources

- Briney, Kristin. “Data Ab Initio” [Blog](http://dataabinitio.com/?p=378), [Blog](http://dataabinitio.com/?p=454)
- Cornell University. "Guide to writing "readme" style metadata" [Webpage](https://data.research.cornell.edu/content/readme)
- DDI. “Create a codebook” [Website](http://www.ddialliance.org/training/getting-started-new-content/create-a-codebook)
- DMPTool "Data Management General Guidance" [Website](https://dmptool.org/dm_guidance)
- Jupyter Team. “The Jupyter Notebook" [Website](http://jupyter-notebook.readthedocs.io/en/stable/notebook.html)