Introduction to Research
Data Management
Data documentation
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
1 may 1994 31

PCR of clones for GSP 30 - 1st selection clones

- To see when clones were combined, will do PCR mix
- 20x more primer 10x more primer

- 2 day mix - 2 1.1kb fragments from good - 1 1.1kb fragment from poor

- Send clones to the lab - Track & Reload.

PCR reaction setup

<table>
<thead>
<tr>
<th>Primer</th>
<th>No.</th>
<th>Size (bp)</th>
<th>1.1</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>10</td>
<td>1 1.1 kb</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>10</td>
<td>1 1.1 kb</td>
<td>1</td>
</tr>
</tbody>
</table>

Final product

- 1.1kb
- 100% success
- 10 replication
- 100% success
- PCR success
- 100% success
- PCR success

PCR conditions

- 94°C - 5 minutes
- 94°C - 10 sec
- 58°C - 10 sec
- 72°C - 10 sec
- 20 cycles

- 72°C - 10 minutes

- 15°C - 10 minutes

PCR success 10/10 - estimated copy @ 1.24
Note-taking software
2017-05-13

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

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');
```

![Graph example](image.png)
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 $x$, i.e. the mean of $t_k$ is defined by some function $\mu_k(x)$. Also assume that all $K$ variables share the same variance, $\sigma^2 = 1/\beta$. Assuming the network output layer has $K$ nodes where $y_k(x, w) \approx \mu_k(x)$ and letting $y(x, w) = [y_1(x, w), \ldots, y_K(x, w)]$, and that we again have $N$ training target values $t$ ($t$ is a $K \times N$ matrix of the training values), the conditional distribution of the target training values is given by

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

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

$$\frac{1}{N} \sum_{n=1}^N \| y(x_n, \hat{w}^{(ML)}) - t_n \|^2$$
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
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.

Example: Tabular data with a data dictionary

<table>
<thead>
<tr>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E Name</th>
<th>F DOB</th>
<th>G DOB_Difference</th>
<th>H Weight g</th>
<th>I Weight_Date</th>
<th>J MonthOfWeight</th>
<th>K AgeAtWt_d</th>
<th>L AgeAtWt_wk</th>
<th>M AgeAtWt_mon</th>
<th>N AgeAtWt_mon_NoDec</th>
<th>O AgeAtWt_y</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Taxon</td>
<td>Hybrid</td>
<td>DLC_JD</td>
<td>Sex</td>
<td>Name</td>
<td>DOB</td>
<td>DOB_Estimated</td>
<td>Weight_g</td>
<td>Weight_Date</td>
<td>MonthOfWeight</td>
<td>AgeAtWt_d</td>
<td>AgeAtWt_wk</td>
<td>AgeAtWt_mon</td>
<td>AgeAtWt_mon_NoDec</td>
</tr>
<tr>
<td>2</td>
<td>CMED</td>
<td>N</td>
<td>1601</td>
<td>M</td>
<td>LINUS</td>
<td>16-Jul-83</td>
<td>77</td>
<td>9-Aug-83</td>
<td>8</td>
<td>24</td>
<td>3.43</td>
<td>0.79</td>
<td>0</td>
<td>0.07</td>
</tr>
<tr>
<td>3</td>
<td>CMED</td>
<td>N</td>
<td>1601</td>
<td>M</td>
<td>LINUS</td>
<td>16-Jul-83</td>
<td>242</td>
<td>21-Oct-83</td>
<td>10</td>
<td>97</td>
<td>13.86</td>
<td>3.19</td>
<td>3</td>
<td>0.27</td>
</tr>
<tr>
<td>4</td>
<td>CMED</td>
<td>N</td>
<td>1601</td>
<td>M</td>
<td>LINUS</td>
<td>16-Jul-83</td>
<td>258</td>
<td>10-Nov-83</td>
<td>11</td>
<td>117</td>
<td>16.71</td>
<td>3.85</td>
<td>3</td>
<td>0.32</td>
</tr>
<tr>
<td>5</td>
<td>CMED</td>
<td>N</td>
<td>1601</td>
<td>M</td>
<td>LINUS</td>
<td>16-Jul-83</td>
<td>357</td>
<td>8-Dec-85</td>
<td>12</td>
<td>876</td>
<td>125.14</td>
<td>28.8</td>
<td>28</td>
<td>2.4</td>
</tr>
<tr>
<td>6</td>
<td>CMED</td>
<td>N</td>
<td>1602</td>
<td>M</td>
<td>LUCAS</td>
<td>16-Jul-83</td>
<td>79</td>
<td>9-Aug-83</td>
<td>8</td>
<td>24</td>
<td>3.43</td>
<td>0.79</td>
<td>3</td>
<td>0.07</td>
</tr>
<tr>
<td>7</td>
<td>CMED</td>
<td>N</td>
<td>1602</td>
<td>M</td>
<td>LUCAS</td>
<td>16-Jul-83</td>
<td>176</td>
<td>25-Jul-93</td>
<td>7</td>
<td>3662</td>
<td>523.14</td>
<td>120.39</td>
<td>120</td>
<td>10.03</td>
</tr>
<tr>
<td>8</td>
<td>CMED</td>
<td>N</td>
<td>1602</td>
<td>M</td>
<td>LUCAS</td>
<td>16-Jul-83</td>
<td>177.7</td>
<td>7-Dec-93</td>
<td>12</td>
<td>3797</td>
<td>542.43</td>
<td>124.83</td>
<td>124</td>
<td>10.4</td>
</tr>
<tr>
<td>9</td>
<td>CMED</td>
<td>N</td>
<td>1602</td>
<td>M</td>
<td>LUCAS</td>
<td>16-Jul-83</td>
<td>181</td>
<td>31-Aug-93</td>
<td>8</td>
<td>3999</td>
<td>528.43</td>
<td>121.61</td>
<td>121</td>
<td>10.13</td>
</tr>
<tr>
<td>10</td>
<td>CMED</td>
<td>N</td>
<td>1602</td>
<td>M</td>
<td>LUCAS</td>
<td>16-Jul-83</td>
<td>193</td>
<td>26-May-93</td>
<td>5</td>
<td>3602</td>
<td>514.57</td>
<td>118.42</td>
<td>118</td>
<td>9.87</td>
</tr>
<tr>
<td>11</td>
<td>CMED</td>
<td>N</td>
<td>1602</td>
<td>M</td>
<td>LUCAS</td>
<td>16-Jul-83</td>
<td>194</td>
<td>13-Aug-93</td>
<td>8</td>
<td>2950</td>
<td>421.43</td>
<td>96.99</td>
<td>96</td>
<td>8.08</td>
</tr>
<tr>
<td>12</td>
<td>CMED</td>
<td>N</td>
<td>1602</td>
<td>M</td>
<td>LUCAS</td>
<td>16-Jul-83</td>
<td>194</td>
<td>20-Feb-97</td>
<td>2</td>
<td>4968</td>
<td>705.71</td>
<td>163.33</td>
<td>153</td>
<td>13.61</td>
</tr>
<tr>
<td>13</td>
<td>CMED</td>
<td>N</td>
<td>1602</td>
<td>M</td>
<td>LUCAS</td>
<td>16-Jul-83</td>
<td>196</td>
<td>5-Jun-95</td>
<td>6</td>
<td>4342</td>
<td>620.29</td>
<td>142.75</td>
<td>142</td>
<td>11.9</td>
</tr>
<tr>
<td>15</td>
<td>CMED</td>
<td>N</td>
<td>1602</td>
<td>M</td>
<td>LUCAS</td>
<td>16-Jul-83</td>
<td>200</td>
<td>11-Sep-96</td>
<td>9</td>
<td>4806</td>
<td>686.57</td>
<td>158.01</td>
<td>158</td>
<td>13.17</td>
</tr>
<tr>
<td>16</td>
<td>CMED</td>
<td>N</td>
<td>1602</td>
<td>M</td>
<td>LUCAS</td>
<td>16-Jul-83</td>
<td>200</td>
<td>16-Jan-97</td>
<td>1</td>
<td>4933</td>
<td>704.71</td>
<td>162.18</td>
<td>162</td>
<td>13.52</td>
</tr>
<tr>
<td>17</td>
<td>CMED</td>
<td>N</td>
<td>1602</td>
<td>M</td>
<td>LUCAS</td>
<td>16-Jul-83</td>
<td>200</td>
<td>4-Apr-97</td>
<td>4</td>
<td>5011</td>
<td>715.86</td>
<td>164.75</td>
<td>164</td>
<td>13.73</td>
</tr>
<tr>
<td>18</td>
<td>CMED</td>
<td>N</td>
<td>1602</td>
<td>M</td>
<td>LUCAS</td>
<td>16-Jul-83</td>
<td>206</td>
<td>10-Jul-96</td>
<td>7</td>
<td>4743</td>
<td>677.57</td>
<td>155.93</td>
<td>155</td>
<td>12.99</td>
</tr>
<tr>
<td>19</td>
<td>CMED</td>
<td>N</td>
<td>1602</td>
<td>M</td>
<td>LUCAS</td>
<td>16-Jul-83</td>
<td>207</td>
<td>7-Aug-96</td>
<td>8</td>
<td>4771</td>
<td>681.57</td>
<td>156.85</td>
<td>156</td>
<td>13.07</td>
</tr>
<tr>
<td>20</td>
<td>CMED</td>
<td>N</td>
<td>1602</td>
<td>M</td>
<td>LUCAS</td>
<td>16-Jul-83</td>
<td>208</td>
<td>7-Nov-96</td>
<td>11</td>
<td>4863</td>
<td>694.71</td>
<td>159.88</td>
<td>159</td>
<td>13.32</td>
</tr>
<tr>
<td>21</td>
<td>CMED</td>
<td>N</td>
<td>1602</td>
<td>M</td>
<td>LUCAS</td>
<td>16-Jul-83</td>
<td>213</td>
<td>8-Jul-92</td>
<td>7</td>
<td>3280</td>
<td>468.57</td>
<td>107.84</td>
<td>107</td>
<td>8.99</td>
</tr>
<tr>
<td>22</td>
<td>CMED</td>
<td>N</td>
<td>1602</td>
<td>M</td>
<td>LUCAS</td>
<td>16-Jul-83</td>
<td>219</td>
<td>8-Jun-95</td>
<td>6</td>
<td>4345</td>
<td>620.71</td>
<td>142.85</td>
<td>142</td>
<td>11.85</td>
</tr>
<tr>
<td>23</td>
<td>CMED</td>
<td>N</td>
<td>1602</td>
<td>M</td>
<td>LUCAS</td>
<td>16-Jul-83</td>
<td>221.7</td>
<td>2-May-84</td>
<td>5</td>
<td>291</td>
<td>41.57</td>
<td>9.57</td>
<td>9</td>
<td>0.8</td>
</tr>
<tr>
<td>24</td>
<td>CMED</td>
<td>N</td>
<td>1602</td>
<td>M</td>
<td>LUCAS</td>
<td>16-Jul-83</td>
<td>225</td>
<td>1-Apr-93</td>
<td>4</td>
<td>3547</td>
<td>506.71</td>
<td>116.61</td>
<td>116</td>
<td>9.72</td>
</tr>
</tbody>
</table>
### 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.\(^4\,5\).

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

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

<table>
<thead>
<tr>
<th># BEDRM: Bedrooms, number of</th>
<th>Information</th>
<th>[Type= discrete] [Format=numeric] [Range= 0-8] [Missing= */8]</th>
</tr>
</thead>
<tbody>
<tr>
<td>Statistics [NW/ W]</td>
<td>[Valid=872972 / 32343616.174 ] [Invalid=14040 / 508706.825 ]</td>
<td></td>
</tr>
<tr>
<td>Definition</td>
<td>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.</td>
<td></td>
</tr>
<tr>
<td>Universe</td>
<td>Reported for: Persons in private households</td>
<td></td>
</tr>
<tr>
<td>Notes</td>
<td>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.</td>
<td></td>
</tr>
</tbody>
</table>

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

- 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)