Example Manuscript Template for a Data Analysis Project

Marguerite Butler

3/22/23

# 1. Project Instructions

*The structure below is one possible setup for a data analysis project (including the course project). For a manuscript, adjust as needed. You don’t need to have exactly these sections, but the content covering those sections should be addressed.*

*This uses MS Word as output format as specified in the YAML header.* [*See here*](https://quarto.org/docs/output-formats/ms-word.html) *for more information. You can switch to other formats, like html or pdf. See* [*the Quarto documentation*](https://quarto.org/) *for other formats.*

*For Project 2, it is fine to have very brief sections, just to practice the structure of the manuscript and get a feel for how you would set it up. The data cleaning should have been documented carefully in Project 1. Project 2 should focus on reporting the analysis, so you might have a few sentences for the data and results sections, and maybe just a sentence for the other sections like the intro, etc.*

*For Project 3 on your own data, I hope youʻll write a few more words, if it helps you think through your research project.*

\_Note on strategy: *Whereas in the exploratory\_analysis.qmd we read in* ***code chunks*** *to produce the figures and tables de novo, here we will read in the saved figures and tables. This is because in order to use code chunks, you generally have to run the whole script. This is good to do in the first case, where you are going through all of the steps but with more commentary. But here, were we are writing up the final manuscript, we will only include the best finished products and it can be cumbersome to rerun the entire analysis every time you want to see what the paper looks after each edit. It is more efficient to just read in* ***graphics files*** *(.pdf, .png, .jpg, etc.) for the figures and* ***.rds*** *files for the tables, both from the Results directory.*

# 2. Summary/Abstract

*Write a summary of your project.*

# 3. Introduction

## 3.1 General Background Information

*Provide enough background on your topic that others can understand the why and how of your analysis*

## 3.2 Description of data and data source

*Describe what the data is, what it contains, where it is from, etc. Eventually this might be part of a methods section.*

*To cite other work (usually done in the introduction, methods, and discussion), make sure your references are in the bibtex file specified in the YAML header above (here references.bib) and have the right bibtex key. Then you can include like below. Remember that enclosing the reference key in [ ] will make a parenthetical citation:*

Here we describe an analysis of the Palmer Penguins dataset (Horst et al. 2020), which contains data on three species of penguins from the Palmer Archipelago, Antarctica. Data were collected and made available by Dr. Kristen Gorman and the Palmer Station, Antarctica LTER, a member of the Long Term Ecological Research Network (Gorman 2014).

## 3.3 Questions/Hypotheses to be addressed

*State the research questions you plan to answer with this analysis. To get you started:*

Our research questions are:

1. Are the species different in size?

# 4. Methods

*Describe your methods. Breifly describe the data, the cleaning processes, and the analysis approaches. You might want to provide a shorter description here and all the details in the supplement (which could be a copy of your code or your .qmd).*

## 4.1 Data acquisition

*As applicable, explain where and how you got the data. If you directly import the data from an online source, you can combine this section with the next.*

## 4.2 Data import and cleaning

*Code that reads in the file and cleans it so it’s ready for analysis. Since this will be fairly long code for most datasets, it might be a good idea to have it in one or several R scripts. If that is the case, explain here briefly what kind of cleaning/processing you do, and provide more details and well documented code somewhere (e.g. in the supplement). All materials, including files that contain code, should be commented well so everyone can follow along.*

*For project 2, you will want to write a brief description of your cleaning and processing steps and include a reference (link to) to your cleaning script or quarto document (or both).*

## 4.3 Statistical analysis

*Explain anything related to your statistical analyses.*

# 5. Results

## 5.1 Exploratory/Descriptive analysis

*Use a combination of text/tables/figures to explore and describe your data. Show the most important descriptive results here. Additional ones should go in the supplement. Even more can be in the R and Quarto files that are part of your project.*

[Table 1](#tbl-summary_table) shows a summary of the data.

*Note the data is loaded using a* ***relative*** *path indicated by the ../../ notation. (Two dots means a folder up). You never want to specify an* ***absolute*** *path like C:\user\myproject\results\ because if you share this with someone, it won’t work for them since they don’t have that path.*

Table 1: Data summary table.

| Variable | N | Mean | SE | Counts |
| --- | --- | --- | --- | --- |
| Individual ID | 339 |  |  |  |
| Clutch Completion | 339 |  |  |  |
| Date Egg | 339 |  |  |  |
| Species | 339 |  |  | Ade: 151, Gen: 120, Chi: 68 |
| Island | 339 |  |  | Bis: 164, Dre: 124, Tor: 51 |
| Sex | 330 |  |  | MAL: 168, FEM: 162 |
| Culmen Length (mm) | 339 | 43.90 | 0.30 |  |
| Culmen Depth (mm) | 339 | 17.17 | 0.11 |  |
| Flipper Length (mm) | 339 | 200.79 | 0.76 |  |
| Body Mass (g) | 339 | 4196.17 | 43.62 |  |
| Delta 15 N (o/oo) | 327 | 8.74 | 0.03 |  |
| Delta 13 C (o/oo) | 328 | -25.68 | 0.04 |  |

## 5.2 Size

*As an example, here is an analysis of size by species and island. You may leave this if you wish, but do edit it to make it your own.*

We found that Gentoo penguins are substantially larger than both Adelie and Chinstrap penguins, which are similar in size ([Figure 1](#fig-mass_species_bars), ANOVA P-value < 0.001 ).

|  |
| --- |
| Figure 1: Body mass distributions of the three species of penguin. |

Table 2: Size differences between species.

|  | Df | Sum Sq | Mean Sq | F value | Pr(>F) |
| --- | --- | --- | --- | --- | --- |
| Species | 2 | 145838050 | 72919025.2 | 339.27 | 0 |
| Residuals | 336 | 72215714 | 214927.7 |  |  |

There does not seem to be any intrinsic island effect. Differences in mass across islands are related to the species which happen to reside on each island ([Figure 2](#fig-mass_species_island_density), [Table 3](#tbl-mass_species_island_anova)).

|  |
| --- |
| Figure 2: Body mass distributions of the three species of penguin. |

Table 3: Size differences between species by island.

|  | Df | Sum Sq | Mean Sq | F value | Pr(>F) |
| --- | --- | --- | --- | --- | --- |
| Species | 2 | 145838050.34 | 72919025.2 | 337.32 | 0.00 |
| Island | 2 | 13654.61 | 6827.3 | 0.03 | 0.97 |
| Residuals | 334 | 72202059.80 | 216173.8 |  |  |

## 5.3 Further analyses (use subheaders as you wish)

*Develop and explore three additional questions by any combination of figures and statistical or modeling methods to analyze your data and to produce meaningful figures, tables, etc. Make it look like a draft scientific paper or project report, appropriate for your career goals.*

*(Reminder: it is now common practice to provide code that fully reproduces the analysis with any publication. This would be your analysis scripts or .qmd, and is provided with supplementary information. Here we are writing a nice document that could be a manuscript, so we want the full analysis code to produce figures and data ready for display as tables, which are loaded here.)*

# 6. Discussion

## 6.1 Summary and Interpretation

*Summarize what you did, what you found and what it means.*

## 6.2 Strengths and Limitations

*Discuss what you perceive as strengths and limitations of your analysis.*

## 6.3 Conclusions

*What are the main take-home messages?*

*Include citations in your Rmd file using bibtex, the list of references will automatically be placed at the end*

This paper (Leek and Peng 2015) discusses types of analyses.

These papers (McKay et al. 2020a,b) are good examples of papers published using a fully reproducible setup similar to the one shown in this template.

Note that this cited reference will show up at the end of the document, the reference formatting is determined by the CSL file specified in the YAML header. Many more style files for almost any journal [are available](https://www.zotero.org/styles). You also specify the location of your bibtex reference file in the YAML. You can call your reference file anything you like, I just used the generic word references.bib but giving it a more descriptive name is probably better.

# 7. References

Gorman, T. D. A. F., Kristen B. AND Williams. 2014. [Ecological sexual dimorphism and environmental variability within a community of antarctic penguins (genus pygoscelis)](https://doi.org/10.1371/journal.pone.0090081). PLOS ONE 9:1–14. Public Library of Science.

Horst, A. M., A. P. Hill, and K. B. Gorman. 2020. [Palmerpenguins: Palmer archipelago (antarctica) penguin data](https://doi.org/10.5281/zenodo.3960218).

Leek, J. T., and R. D. Peng. 2015. [Statistics. What is the question?](https://doi.org/10.1126/science.aaa6146) Science (New York, N.Y.) 347:1314–1315.

McKay, B., M. Ebell, W. Z. Billings, A. P. Dale, Y. Shen, and A. Handel. 2020a. [Associations Between Relative Viral Load at Diagnosis and Influenza A Symptoms and Recovery.](https://doi.org/10.1093/ofid/ofaa494) Open forum infectious diseases 7:ofaa494. United States.

McKay, B., M. Ebell, A. P. Dale, Y. Shen, and A. Handel. 2020b. [Virulence-mediated infectiousness and activity trade-offs and their impact on transmission potential of influenza patients.](https://doi.org/10.1098/rspb.2020.0496) Proceedings. Biological sciences 287:20200496. England.