Example Manuscript Template for a Data Analysis Project

Doreen Kibuule Kalembe

The structure below is one possible setup for a manuscript, or a general data analysis project (including the course project). 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. [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.

Warning: package 'here' was built under R version 4.4.2

**Authors**

* First Autor (ORCID: 0000-0000-1234-5678)
* Second Author
* Third Autor
* Last Author

**Author affiliations**

1. College of Public Health, University of Georgia, Athens, GA, USA.
2. Another Affiliation, Athens, GA, USA.
3. Yet another Affiliation.

These authors contributed equally to this work.

Corresponding author: some@email.com

Disclaimer: The opinions expressed in this article are the author’s own and don’t reflect their employer.

# 1. Summary/Abstract

*Write a summary of your project.* MADA PROJECT PART 1:

TITLE: Sex-based analysis/Differences in TB Immune Activation.

This project is aimed at looking at the differences in the Tuberculosis immune activation between males and females.( Sex-based analysis/Differences in TB Immune Activation Research Question)

we shall use data from a tuberculosis public repository. • Data source: Tidyverse https://dataverse.unc.edu/dataset.xhtml?persistentId=doi:10.15139/S3/AYOFEU with a Sample size of 60 individuals. The data includes 60 different individual that accepted to take part in the study.

The key variables that we shall look at are;

Key variables: - CD4 Immune activation count - Disease severity indicators(Extent of lung disease andNumber of lungs involved) -Fat in kgs - Sex/Gender - BMI - Age of the participant

#load packages  
library("readxl")  
library("readr")  
library("dplyr")

Attaching package: 'dplyr'

The following objects are masked from 'package:stats':  
  
 filter, lag

The following objects are masked from 'package:base':  
  
 intersect, setdiff, setequal, union

library("ggplot2")  
library("tidyverse")

Warning: package 'tidyverse' was built under R version 4.4.2

Warning: package 'stringr' was built under R version 4.4.2

── Attaching core tidyverse packages ──────────────────────── tidyverse 2.0.0 ──  
✔ forcats 1.0.0 ✔ stringr 1.5.1  
✔ lubridate 1.9.3 ✔ tibble 3.2.1  
✔ purrr 1.0.2 ✔ tidyr 1.3.1

── Conflicts ────────────────────────────────────────── tidyverse\_conflicts() ──  
✖ dplyr::filter() masks stats::filter()  
✖ dplyr::lag() masks stats::lag()  
ℹ Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors

library("here")  
  
# follow through with the code folder to see the #variables in the data and how i selected the #variables make this data set.   
#location of the data  
data\_location <- here::here("data","processed-data","Tb\_immune\_data.rds")  
  
#loading the data  
data<- read\_rds(data\_location)  
  
#summary of the data  
summary(data)

ID Sex Participant age BMI   
 Min. :50004 Length:61 Min. :17.00 Min. :15.55   
 1st Qu.:50020 Class :character 1st Qu.:23.00 1st Qu.:18.09   
 Median :50035 Mode :character Median :28.00 Median :19.25   
 Mean :50035 Mean :30.02 Mean :19.83   
 3rd Qu.:50050 3rd Qu.:35.00 3rd Qu.:21.47   
 Max. :50067 Max. :70.00 Max. :28.15   
 NA's :1 NA's :1   
 Dose Fat in kg CD4+   
 Min. :29.61 Min. :-96.640 Min. : 8.636   
 1st Qu.:29.99 1st Qu.: 3.554 1st Qu.:185.006   
 Median :30.00 Median : 8.704 Median :317.941   
 Mean :29.99 Mean : 9.615 Mean :337.364   
 3rd Qu.:30.00 3rd Qu.: 13.364 3rd Qu.:484.889   
 Max. :30.15 Max. : 80.979 Max. :785.896   
 NA's :1 NA's :1 NA's :1   
 CD4 Immune activation count  
 Min. : 84.0   
 1st Qu.: 914.8   
 Median : 1384.5   
 Mean : 1948.1   
 3rd Qu.: 2057.5   
 Max. :17872.0   
 NA's :1

Proposed Analysis Method

• Compare immune activation levels between males and females

• Correlate immune markers with disease severity measures/Correlation between activation levels and disease severity

• Use linear regression to analyze relationships between variables

• Control for potential confounding factors

These are some of our Expected Outcomes.

• Check and Identify any sex-based differences in immune response.

• Quantify the relationship between immune activation and disease severity.

• Propose any insights for sex-specific treatment approaches

# 2. Introduction

## 2.1 General Background Information

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

## 2.2 Description of data and data source

## 2.3 Questions/Hypotheses to be addressed

*State the research questions you plan to answer with this analysis.*

To cite other work (important everywhere, but likely happens first in introduction), make sure your references are in the bibtex file specified in the YAML header above and have the right bibtex key. Then you can include like this:

Examples of reproducible research projects can for instance be found in (1,2).

# 3. Methods

*Describe your methods. That should 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.*

## 3.1 Schematic of workflow

Sometimes you might want to show a schematic diagram/figure that was not created with code (if you can do it with code, do it). [Figure 1](#fig-schematic) is an example of some - completely random/unrelated - schematic that was generated with Biorender. We store those figures in the assets folder.

|  |
| --- |
| Figure 1: A figure that is manually generated and shows some overview/schematic. This has nothing to do with the data, it’s just a random one from one of our projects I found and placed here. |

## 3.2 Data aquisition

*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.*

## 3.3 Data import and cleaning

*Write 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. as supplement in a paper). All materials, including files that contain code, should be commented well so everyone can follow along.*

## 3.4 Statistical analysis

*Explain anything related to your statistical analyses.*

# 4. Results

## 4.1 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-summarytable) shows a summary of the data.

Note the loading of the data providing a **relative** path using the ../../ notation. (Two dots means a folder up). You never want to specify an **absolute** path like C:\ahandel\myproject\results\ because if you share this with someone, it won’t work for them since they don’t have that path. You can also use the here R package to create paths. See examples of that below. I generally recommend the here package.

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Table 1: Data summary table.   | skim\_type | skim\_variable | n\_missing | complete\_rate | factor.ordered | factor.n\_unique | factor.top\_counts | numeric.mean | numeric.sd | numeric.p0 | numeric.p25 | numeric.p50 | numeric.p75 | numeric.p100 | numeric.hist | | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | | factor | Gender | 0 | 1 | FALSE | 3 | M: 4, F: 3, O: 2 | NA | NA | NA | NA | NA | NA | NA | NA | | numeric | Height | 0 | 1 | NA | NA | NA | 165.66667 | 15.97655 | 133 | 156 | 166 | 178 | 183 | ▂▁▃▃▇ | | numeric | Weight | 0 | 1 | NA | NA | NA | 70.11111 | 21.24526 | 45 | 55 | 70 | 80 | 110 | ▇▂▃▂▂ | |

## 4.2 Basic statistical analysis

*To get some further insight into your data, if reasonable you could compute simple statistics (e.g. simple models with 1 predictor) to look for associations between your outcome(s) and each individual predictor variable. Though note that unless you pre-specified the outcome and main exposure, any “p<0.05 means statistical significance” interpretation is not valid.*

[Figure 2](#fig-result) shows a scatterplot figure produced by one of the R scripts.

|  |
| --- |
| Figure 2: Height and weight stratified by gender. |

## 4.3 Full analysis

*Use one or several suitable statistical/machine learning methods to analyze your data and to produce meaningful figures, tables, etc. This might again be code that is best placed in one or several separate R scripts that need to be well documented. You want the code to produce figures and data ready for display as tables, and save those. Then you load them here.*

Example [Table 2](#tbl-resulttable2) shows a summary of a linear model fit.

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Table 2: Linear model fit table.   | term | estimate | std.error | statistic | p.value | | --- | --- | --- | --- | --- | | (Intercept) | 149.2726967 | 23.3823360 | 6.3839942 | 0.0013962 | | Weight | 0.2623972 | 0.3512436 | 0.7470519 | 0.4886517 | | GenderM | -2.1244913 | 15.5488953 | -0.1366329 | 0.8966520 | | GenderO | -4.7644739 | 19.0114155 | -0.2506112 | 0.8120871 | |

# 5. Discussion

## 5.1 Summary and Interpretation

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

## 5.2 Strengths and Limitations

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

## 5.3 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 (3) discusses types of analyses.

These papers (1,2) 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.

# 6. References

1. McKay B, Ebell M, Billings WZ, et al. [Associations Between Relative Viral Load at Diagnosis and Influenza A Symptoms and Recovery.](https://doi.org/10.1093/ofid/ofaa494) *Open forum infectious diseases*. 2020;7(11):ofaa494.

2. McKay B, Ebell M, Dale AP, et al. [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*. 2020;287(1927):20200496.

3. Leek JT, Peng RD. [Statistics. What is the question?](https://doi.org/10.1126/science.aaa6146) *Science (New York, N.Y.)*. 2015;347(6228):1314–1315.