Example Manuscript Template for a Data Analysis Project

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.

**Authors**

* Natalie Cann

**Author affiliations**

1. Masters in Public Health Candidate, University of Georgia
2. B.S. Microbiology, Virginia Tech

These authors contributed equally to this work.

Corresponding author: npc31984@uga.edu

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.*

# 2. Introduction

## 2.1 General Background Information

*Provide enough background on your topic that others can understand the why and how of your analysis* The novel coronvirus, SARS-CoV-2, resulted in a global pandemic unfolding in late 2019 and resulted in numerous cases and deaths worldwide (1). SARS-CoV-2, henceforth referred to as COVID-19, has resulted in 1,143,724 deaths in the United States alone between January 2020 and September 2023 (2). Of specific interest, is the virus’ spike protein (S); the S protein is the primary target for neutralizing antibodies and has been used in vaccine development (3). The S protein has accumulated mutations, resulting in COVID-19 variants. These mutations have the ability to increase the virulence and transmissibility of the virus, and therefore, decrease the effectiveness of the COVID-19 vaccine (3). It is crucial to understand the impacts that these variants have on the number of cases and deaths attributed to COVID-19. A better understanding can help lead to improved vaccine development and stronger control measures. In this project, I aim to assess the relationship between different COVID-19 variants and the number of cases, deaths, and the duration of the epidemic that they caused.

## 2.2 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.* I have decided to use a dataset that contains information on various variants of COVID-19. This data, found on kaggle, combines data from John Hopkins University (confirmed\_cases.csv and deaths\_cases.csv) and covariants.org (variants.csv). The link to the kaggle website containing the dataset can be found below. The data contains information on country, the variant, the number of days that the epidemic lasted, the mortality rate, the total number of cases, and the total number of deaths associated with the variant’s epidemic in that country.

## 2.3 Questions/Hypotheses to be addressed

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

1. How are variant strains of COVID-19 related to number of cases (worldwide)? I aim to assess if certain variants of COVID-19 are associated with higher numbers of cases. This could help to understand if some variants are more infectious and spread easier than others. I will use the total\_cases and variant variables to research this question.
2. How are variant strains of COVID-19 related to number of deaths (worldwide)? I aim to assess if certain variants of COVID-19 are associated with higher numbers of deaths. This could help to understand if some variants are more deadly than others. I will use the total\_deaths and variant variables to research this question.
3. How are variant strains of COVID-19 related to the duration of the epidemic that it caused (in days) (worldwide)? I aim to assess if certain variants of COVID-19 are associated with longer-lasting epidemics. This could provide insight into the transmissibility of the variant as well as how long it takes the cases to recover. I will use the duration and variant variables to research this question.

# 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.*

I will perform data cleaning and create a new dataset for each research question that only includes the variables of interest for that research question. I will create plots for each research question that help to display the relationship between variant and the other variable of interest.

I will perform statistial analyses such as linear fit models to see if the variant is a good predictor of number of cases, number of deaths, and the duration of the epidemic.

## 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.* This data set was downloaded from kaggle via the following link: https://www.kaggle.com/datasets/lumierebatalong/covid-19-variants-survival-data?r esource=download.

## 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.* Data Import via read\_csv()

surv\_variants <- read\_csv("~/Desktop/MADA/Cann-MADA-project/surv\_variants.csv")

Rows: 4113 Columns: 12  
── Column specification ────────────────────────────────────────────────────────  
Delimiter: ","  
chr (2): Country, variant  
dbl (6): num\_seqs, duration, mortality\_rate, total\_cases, total\_deaths, gro...  
lgl (1): censored  
date (3): first\_seq, last\_seq, censure\_date  
  
ℹ Use `spec()` to retrieve the full column specification for this data.  
ℹ Specify the column types or set `show\_col\_types = FALSE` to quiet this message.

Now that the data is imported, I will view the first rows as well as the dimensions of the dataset.

head(surv\_variants) #viewing first rows

# A tibble: 6 × 12  
 Country first\_seq num\_seqs last\_seq variant censure\_date duration censored  
 <chr> <date> <dbl> <date> <chr> <date> <dbl> <lgl>   
1 China 2019-10-22 3 2020-06-03 S.Q677 2020-06-21 225 TRUE   
2 USA 2020-03-03 26022 2021-11-19 S.Q677 2020-11-01 626 FALSE   
3 Brazil 2020-03-09 1553 2021-11-12 S.Q677 2020-11-07 613 FALSE   
4 Austral… 2020-03-20 88 2021-11-14 S.Q677 2020-11-18 604 FALSE   
5 Sweden 2020-03-20 810 2021-11-19 S.Q677 2020-11-18 609 FALSE   
6 Spain 2020-03-20 1033 2021-11-21 S.Q677 2020-11-18 611 FALSE   
# ℹ 4 more variables: mortality\_rate <dbl>, total\_cases <dbl>,  
# total\_deaths <dbl>, growth\_rate <dbl>

dim(surv\_variants) #viewing dimensions

[1] 4113 12

As you can see, the data contains 12 variables and 4113 observations. A few variables that I am interested in include: country, variant, duration (of epidemic, in days), mortality\_rate, total\_cases, and total\_deaths.

## 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 (**leek2015?**) discusses types of analyses.

These papers (**mckay2020?**,**mckay2020a?**) 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. Coronavirus COVID-19 (SARS-CoV-2) | Johns Hopkins ABX Guide. (<https://www.hopkinsguides.com/hopkins/view/Johns_Hopkins_ABX_Guide/540747/all/Coronavirus_COVID_19__SARS_CoV_2_?refer=true>). (Accessed January 28, 2025)

2. COVID-19 Data from the National Center for Health Statistics. 2023;(<https://www.cdc.gov/nchs/covid19/index.htm>). (Accessed January 29, 2025)

3. Hadj Hassine I. Covid‐19 vaccines and variants of concern: A review. *Rev Med Virol* [electronic article]. 2022;32(4):e2313. (<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8646685/>). (Accessed January 29, 2025)