Example Manuscript Template for a Data Analysis Project

Andreas Handel

2021-09-17

#Loading required packages  
library(dplyr) #for data processing

##   
## 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(here) #to set paths

## here() starts at C:/Users/Priyanka/Desktop/new/Priyanka\_G-MADA-project

library(tidyverse) #all required data manipulation packages

## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --

## v ggplot2 3.3.5 v purrr 0.3.4  
## v tibble 3.1.3 v stringr 1.4.0  
## v tidyr 1.1.3 v forcats 0.5.1  
## v readr 2.0.1

## -- Conflicts ------------------------------------------ tidyverse\_conflicts() --  
## x dplyr::filter() masks stats::filter()  
## x dplyr::lag() masks stats::lag()

# 1 Summary/Abstract

*Write a summary of your project.*

# 2 KEY QUESTIONS

What is already known about this subject?

Residents of rural communities are at increased risk for severe COVID-19–associated morbidity and mortality. In September 2020, COVID-19 incidence (cases per 100,000 population) in rural counties surpassed that in urban counties.

What will this study add or answer?

The present analysis will attempt to estimate the percent of the population in each county that may be vaccine hesitant

Relationship between ethnicity and vaccine hesitancy

How might this impact clinical practice?

Identify and address barriers to COVID-19 vaccination in rural areas.

## 2.1 General Background Information

Coronavirus disease (COVID-19) is an infectious disease caused by a newly discovered coronavirus. Coronavirus disease 2019 (COVID-19) was declared a pandemic in March 2020. COVID-19 vaccine is the most sustainable option to manage the current pandemic. However, vaccine hesitancy by even a small subset of the population can undermine the success of this strategy.

## 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.*

The dataset used for this analysis is publicly available on CDC website. <https://data.cdc.gov/Vaccinations/Vaccine-Hesitancy-for-COVID-19-County-and-local-es/q9mh-h2tw>

The hesitancy rates at the county level using the Public Use Microdata Sample (PUMS). The data was collected by utilizing survey question, “Once a vaccine to prevent COVID-19 is available to you, would you…get a vaccine?” which provides the following options:

1. “definitely get a vaccine”;
2. “probably get a vaccine”;
3. “unsure”; 4) “probably not get a vaccine”;
4. “definitely not get a vaccine.”

We use three definitions to capture the strength of hesitancy to receive a vaccine.

Strongly hesitant: includes only survey responses indicating that they would “definitely not” receive a COVID-19 vaccine when available.

Hesitant: includes survey responses indicating that they would “probably not” or “definitely not” receive a COVID-19 vaccine when available.

Hesitant or unsure: includes survey responses indicating that they would “probably not” or “unsure” or “definitely not” receive a COVID-19 vaccine when available

## 2.3 Questions/Hypotheses to be addressed

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

The present analysis will attempt to estimate the percent of the population in each county that may be vaccine hesitant

Relationship between ethnicity and vaccine hesitancy

# 3 Future analysis

I plan to present a summary of county-wise vaccine estimate based on race and ethnicity. I will plot some graphs (box-plot, scatter plot). I am also planning to do some regression analysis too

# 4 Methods and Results

*In most research papers, results and methods are separate. You can combine them here if you find it easier. You are also welcome to structure things such that those are separate sections.*

## 4.1 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.*

## 4.2 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.*

## 4.3 Exploratory analysis

*Use a combination of text/tables/figures to explore and describe your data. You should produce plots or tables or other summary quantities for the most interesting/important quantities in your data. Depending on the total number of variables in your dataset, explore all or some of the others. FIgures produced here might be histograms or density plots, correlation plots, etc. Tables might summarize your data.*

*Continue by creating plots or tables of the outcome(s) of interest and the predictor/exposure/input variables you are most interested in. If your dataset is small, you can do that for all variables. Plots produced here can be scatterplots, boxplots, violinplots, etc. Tables can be simple 2x2 tables or larger ones.*

*To get some further insight into your data, if reasonable you could compute simple statistics (e.g. t-tests, simple regression model with 1 predictor, etc.) 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.*

Table 4.1 shows a table summarizing the data.

Table 4.1: Data summary table.

|  |  |  |
| --- | --- | --- |
|  | Height | Weight |
| Min. | 133.00 | 45.00 |
| 1st Qu. | 155.25 | 54.25 |
| Median | 166.00 | 73.00 |
| Mean | 165.50 | 72.00 |
| 3rd Qu. | 177.25 | 87.50 |
| Max. | 192.00 | 110.00 |

Figure 4.1 shows a scatterplot figure produced by one of the R scripts.

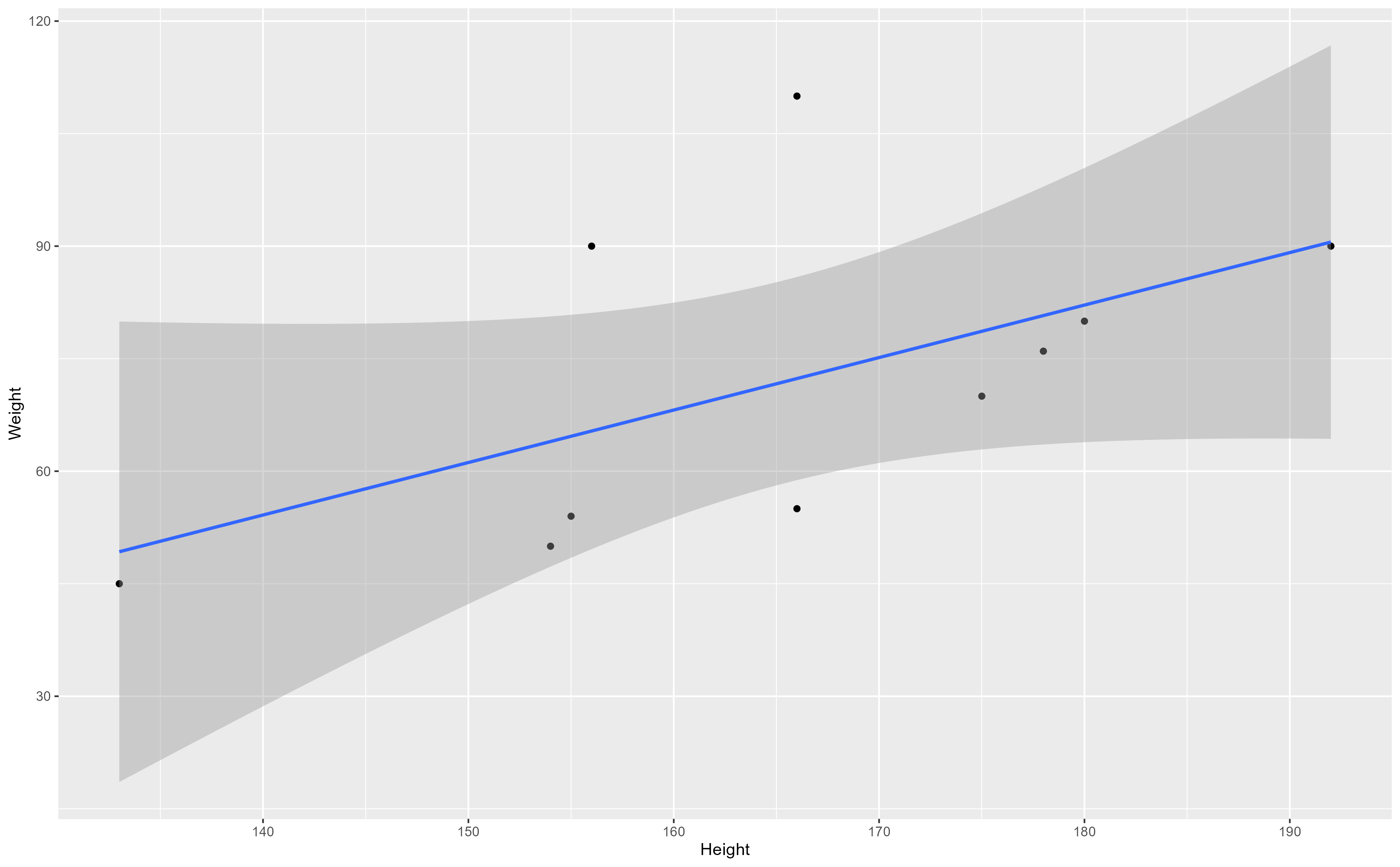


Figure 4.1: Analysis figure.

## 4.4 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 4.2 shows a table summarizing a linear model fit.

Table 4.2: Linear model fit table.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| term | estimate | std.error | statistic | p.value |
| (Intercept) | -43.7883068 | 61.1150617 | -0.7164896 | 0.4940713 |
| Height | 0.6996272 | 0.3675692 | 1.9033889 | 0.0934786 |

# 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 (Leek & Peng, 2015) discusses types of analyses.

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.

# References

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