MADA Project Manuscript

Dawson Dobash

2021-09-17

# 1 Summary/Abstract

*Write a summary of your project.*

# 2 Introduction

## 2.1 General Background Information

COVID-19 hit the world by storm and is a possible ever-lasting change to the way society acts. It also had an impact on each countries funding towards health care. Actions within the past two years have been taken to help fight against this virus.

## 2.2 Description of data and data source

The data I plan on using is from a mixture of locations. Two sources are from kaggle.com and the other is from the Organisation for Economic Co-operation and Development (OECD) website. My data consists of eight data sets with information over each country in the world (for those that data was collected for). Six of these data sets provide the same information just for different countries where each data set is a different region. There is a data set that includes all of the countries but I want to include a region indicator variable for each country. These data sets includes total covid-19 cases, active cases, total deaths, total recovered, total tests, and population size. This data was updated on kaggle as of Sept 16th. Another data set I am using is a vaccination data set for each country. This is also data as of Sept. 16th. Finally, the last data set is of each countries healthcare funding for the past 4-5 years. Some countries provide 2020 data but others only provide up to 2019.

I plan on combining the data sets into one with the variable combining them all is the country. If there are missing data within the variables used in analysis, I will exclude those countries. Ill have to change the raw data as one data set uses 3 letter abbreviations for countries while the others use the whole name of the country. There will be a lot of cleaning for this data just to make it as I want it.

I included 3 data sets out of the 8 within the processingscript for now because one shows the same information as the other data sets not included just in a different region of the world. This is just to show the data I will be using.

I plan on trying to find more information/data sets about some demographics data for the countries like type of government and an indicator for developed vs. developing countries.

Links to the locations I received the data are below:

<https://data.oecd.org/healthres/health-spending.htm>

<https://www.kaggle.com/iamsouravbanerjee/covid19-dataset-world-and-continent-wise?select=Covid+Data+-+World.csv>

<https://www.kaggle.com/anandhuh/latest-worldwide-vaccine-data>

## 2.3 Questions/Hypotheses to be addressed

Since this data is cumulative and does not break down the amount of cases per month per country, I want to compare the recovery proportion to the death proportion and if this difference is based on variables such as the number of vaccines pushed out to society, the amount of funding the health care system has, the amount of tests taken. I also plan on adding more information as state in the description of data section. This will hopefully show if the amount of health care support really helps patients recover.

Some issues I can run into are: 1) Developed countries might have more cases as they are in contact with the world more as travelers commonly travel to certain countries. Therefore could show an inverse relationship to what I expect with more developed countries have more recoveries. 2) Unknown dates of when a country received vaccines.

# 3 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.*

## 3.1 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.*

## 3.2 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.*

## 3.3 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 3.1 shows a table summarizing the data.

Table 3.1: 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 3.1 shows a scatterplot figure produced by one of the R scripts.

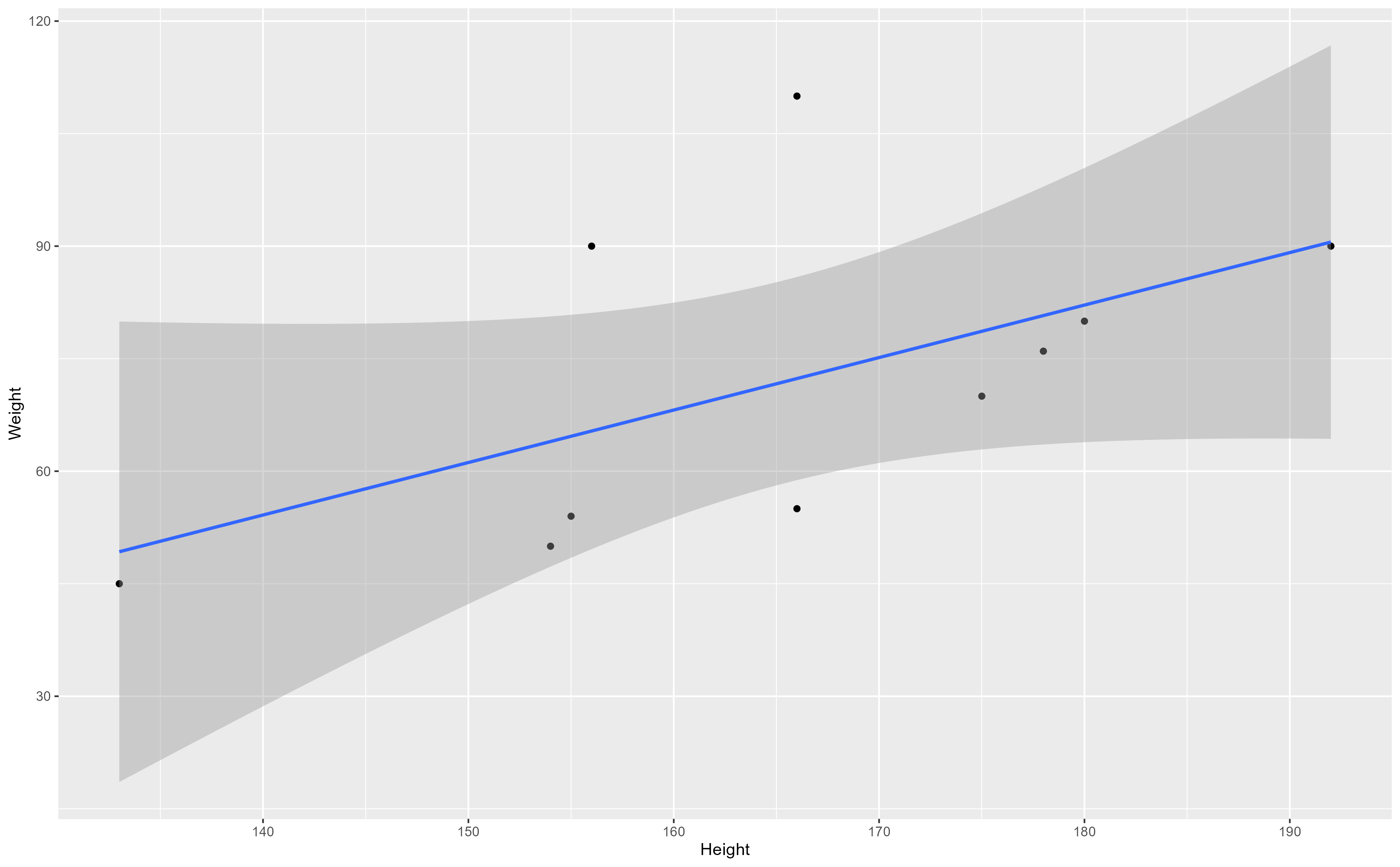


Figure 3.1: Analysis figure.

## 3.4 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 3.2 shows a table summarizing a linear model fit.

Table 3.2: 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 |

# 4 Discussion

## 4.1 Summary and Interpretation

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

## 4.2 Strengths and Limitations

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

## 4.3 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>