Which Mushroom will kill you?!

Alex Martinez

2024-06-17

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. [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 4.3.3

# 1. Summary/Abstract

**Briefly describe what the data is, how it was collected, where you will get (or got) it from. How many observations do you have, what was measured? Anything else important to report about the data?**

The dataset I will be using is from Kaggle called [Muchroom Dataset](https://www.kaggle.com/datasets/prishasawhney/mushroom-dataset/data). The kaggle version is a cleaned version of the original [Mushroom Dataset for Binary Classification](https://archive.ics.uci.edu/dataset/848/secondary+mushroom+dataset), available at UCI Library. The data set contains 9 columns and 54036 rows. There is one variable that is binary which indicates if the mushroom is edible or not. The following 8 variables are different features of the mushroom. Here are those variables.

Cap Diameter

Cap Shape

Gill Attachment

Gill Color

Stem Height

Stem Width

Stem Color Season

Target Class - Is it edible or not?

**At this stage you are not required to already have and show the data, but if you do, even better. Then add a few lines of code which load the data and using some of the commands you learned about, provide summary descriptions of the data.**

I am still working on importing and creating summary descriptions of my dataset.

**Explain the question you want to answer using the data. What will be your outcome(s) of interest (if any)? What (if any) specific predictors will you focus on? What relations/patterns are you looking for in the data?**

The main quesiton I want to answer is the following. *Determine what is the best model to predict if a mushroom is edible or not using the following 8 variables are your predictors.*

Cap Diameter

Cap Shape

Gill Attachment

Gill Color

Stem Height

Stem Width

Stem Color Season

Target Class - Is it edible or not?

I will focus on each predictor variable to see what is most significant in predicting if a mushroom is edible or not.

**As much as you know, suggest how you will analyze it. At this stage in the course, we haven’t covered analysis approaches yet, so you can keep things vague and non-technical here.**

I will run multiple different regression models to determine which is best at predicting the predictor varible. Some models I will consider are, linear/multilinear regression, logistic regression, decision tree models, etc.

**You are allowed, but not yet required, to provide background information for the question you plan to answer. For instance you can describe why it’s an interesting question, who else has done similar analyses, how your analysis will be new/different, etc. Similar to what you read in an introduction to a research paper. For the final report, you’ll need these parts. For part 1, they are not required, but you are welcome to already write down some of that.**

I want to enhance my modeling skills by trying multiple models to obtain the largest accuracy. Something unique that I will be doing compared to everything Ive seen on Kaggle for this data set is that I will be using R while everyone else used Python.

**Eventually, for your final report, what you write for this part will go into different sections of the full report. Some will go into the introduction, some in the methods section. You can already place these items there, or for now just write them as a single section.**

Ok thanks.

### 1.0.1 BELOW IS WHAT WILL GET DONE THROUGHOUT THE SEMESTER (:

# 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

This dataset is a cleaned version of the original [Mushroom Dataset for Binary Classification](https://archive.ics.uci.edu/dataset/848/secondary+mushroom+dataset) Available at UCI Library. This dataset was cleaned using various techniques such as Modal imputation, one-hot encoding, z-score normalization, and feature selection. It contains 9 columns:

1. Cap Diameter
2. Cap Shape
3. Gill Attachment
4. Gill Color
5. Stem Height
6. Stem Width
7. Stem Color
8. Season
9. Target Class - Is it edible or not?

The Target Class contains two values - 0 or 1 - where 0 refers to edible and 1 refers to poisonous.

## 2.3 Questions/Hypotheses to be addressed

*Using the Variables listed above, what is the best predictive model to determine if a mushroom is edible or not.*

# 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

*The data is from the Kaggle website that has been listed above. It is a CSV file with 9 variables and ~50,000 rows.*

## 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.*

The data is pretty cleaned already espescially since it is only 9 variables and ~50,000 rows. Here is the import

library(tidyverse)

Warning: package 'tidyverse' was built under R version 4.3.3

Warning: package 'ggplot2' was built under R version 4.3.3

Warning: package 'tibble' was built under R version 4.3.3

Warning: package 'tidyr' was built under R version 4.3.3

Warning: package 'readr' was built under R version 4.3.3

Warning: package 'purrr' was built under R version 4.3.3

Warning: package 'dplyr' was built under R version 4.3.3

Warning: package 'stringr' was built under R version 4.3.3

Warning: package 'forcats' was built under R version 4.3.3

Warning: package 'lubridate' was built under R version 4.3.3

── Attaching core tidyverse packages ──────────────────────── tidyverse 2.0.0 ──  
✔ dplyr 1.1.4 ✔ readr 2.1.5  
✔ forcats 1.0.0 ✔ stringr 1.5.1  
✔ ggplot2 3.5.1 ✔ tibble 3.2.1  
✔ lubridate 1.9.3 ✔ tidyr 1.3.1  
✔ purrr 1.0.2   
── 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

getwd()

[1] "C:/Users/alexm/Downloads/Pract 2/AlexMartinez-Analysis-Pract2/products/manuscript"

mushrooms <- read.csv('C:/Users/alexm/Downloads/Pract 2/AlexMartinez-Analysis-Pract2/data/mushroom\_cleaned.csv')  
  
  
## Lets Explore the data  
head(mushrooms)

cap.diameter cap.shape gill.attachment gill.color stem.height stem.width  
1 1372 2 2 10 3.807467 1545  
2 1461 2 2 10 3.807467 1557  
3 1371 2 2 10 3.612496 1566  
4 1261 6 2 10 3.787572 1566  
5 1305 6 2 10 3.711971 1464  
6 1337 6 2 10 3.775635 1520  
 stem.color season class  
1 11 1.8042727 1  
2 11 1.8042727 1  
3 11 1.8042727 1  
4 11 1.8042727 1  
5 11 0.9431946 1  
6 11 0.9431946 1

str(mushrooms)

'data.frame': 54035 obs. of 9 variables:  
 $ cap.diameter : int 1372 1461 1371 1261 1305 1337 1300 1354 1222 1085 ...  
 $ cap.shape : int 2 2 2 6 6 6 2 6 6 6 ...  
 $ gill.attachment: int 2 2 2 2 2 2 2 2 2 2 ...  
 $ gill.color : int 10 10 10 10 10 10 10 10 10 10 ...  
 $ stem.height : num 3.81 3.81 3.61 3.79 3.71 ...  
 $ stem.width : int 1545 1557 1566 1566 1464 1520 1563 1532 1476 1581 ...  
 $ stem.color : int 11 11 11 11 11 11 11 11 11 11 ...  
 $ season : num 1.804 1.804 1.804 1.804 0.943 ...  
 $ class : int 1 1 1 1 1 1 1 1 1 1 ...

summary(mushrooms)

cap.diameter cap.shape gill.attachment gill.color   
 Min. : 0.0 Min. :0 Min. :0.000 Min. : 0.00   
 1st Qu.: 289.0 1st Qu.:2 1st Qu.:0.000 1st Qu.: 5.00   
 Median : 525.0 Median :5 Median :1.000 Median : 8.00   
 Mean : 567.3 Mean :4 Mean :2.142 Mean : 7.33   
 3rd Qu.: 781.0 3rd Qu.:6 3rd Qu.:4.000 3rd Qu.:10.00   
 Max. :1891.0 Max. :6 Max. :6.000 Max. :11.00   
 stem.height stem.width stem.color season   
 Min. :0.000426 Min. : 0 Min. : 0.000 Min. :0.02737   
 1st Qu.:0.270997 1st Qu.: 421 1st Qu.: 6.000 1st Qu.:0.88845   
 Median :0.593295 Median : 923 Median :11.000 Median :0.94319   
 Mean :0.759110 Mean :1051 Mean : 8.418 Mean :0.95216   
 3rd Qu.:1.054858 3rd Qu.:1523 3rd Qu.:11.000 3rd Qu.:0.94319   
 Max. :3.835320 Max. :3569 Max. :12.000 Max. :1.80427   
 class   
 Min. :0.0000   
 1st Qu.:0.0000   
 Median :1.0000   
 Mean :0.5492   
 3rd Qu.:1.0000   
 Max. :1.0000

## 3.4 Statistical analysis

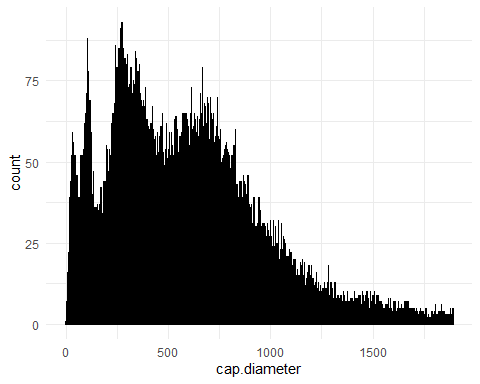
*Explain anything related to your statistical analyses.*

# 4. Results

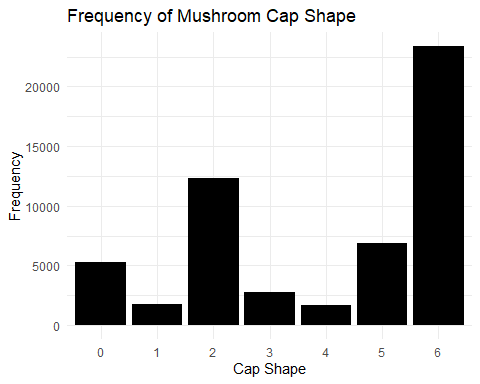
## 4.1 Exploratory/Descriptive analysis

*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.*

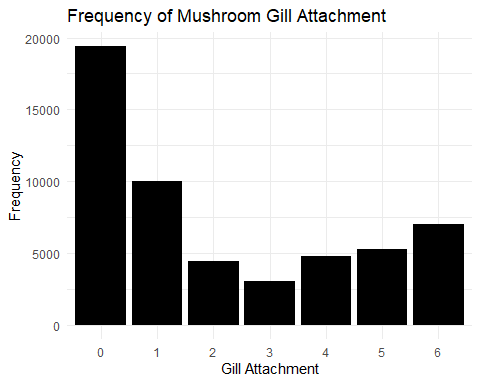
#Exploring all varibles with some visualizations  
mushrooms %>% ggplot(aes(x = cap.diameter)) +  
 geom\_histogram(binwidth = 1, fill = "blue", color = "black") +  
 theme\_minimal()



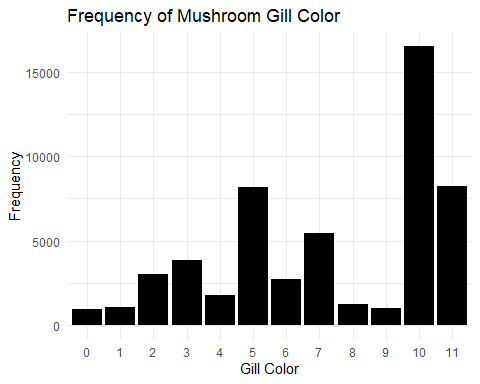
ggplot(mushrooms, aes(x = factor(cap.shape))) +  
 geom\_bar(fill = 'black') +  
 labs(x = "Cap Shape", y = "Frequency", title = "Frequency of Mushroom Cap Shape") +  
 theme\_minimal()



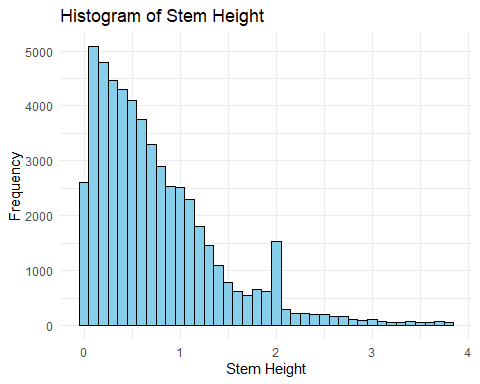
ggplot(mushrooms, aes(x = factor(gill.attachment))) +  
 geom\_bar(fill = 'black') +  
 labs(x = "Gill Attachment", y = "Frequency", title = "Frequency of Mushroom Gill Attachment") +  
 theme\_minimal()



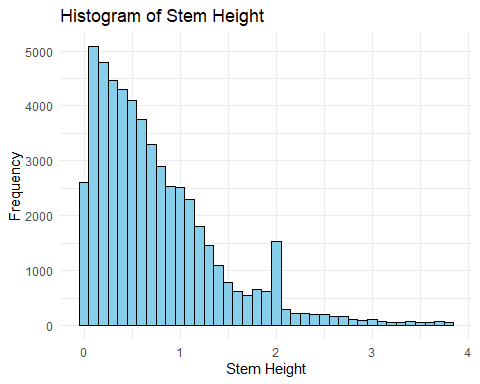
ggplot(mushrooms, aes(x = factor(gill.color))) +  
 geom\_bar(fill = 'black') +  
 labs(x = "Gill Color", y = "Frequency", title = "Frequency of Mushroom Gill Color") +  
 theme\_minimal()



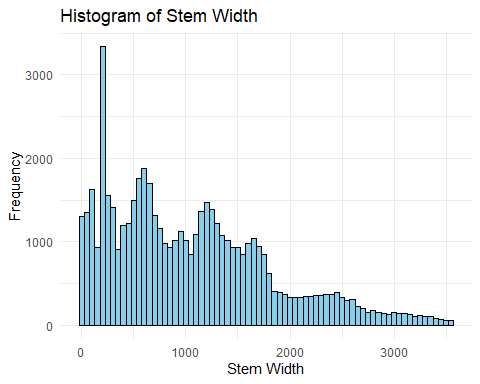
ggplot(mushrooms, aes(x = stem.height)) +  
 geom\_histogram(binwidth = 0.1, fill = "skyblue", color = "black") +  
 labs(x = "Stem Height", y = "Frequency", title = "Histogram of Stem Height") +  
 theme\_minimal()



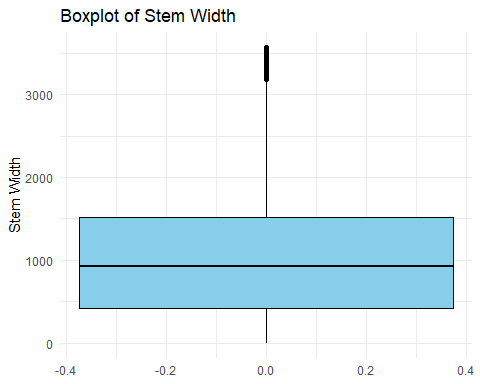
ggplot(mushrooms, aes(x = stem.height)) +  
 geom\_histogram(binwidth = 0.1, fill = "skyblue", color = "black") +  
 labs(x = "Stem Height", y = "Frequency", title = "Histogram of Stem Height") +  
 theme\_minimal()



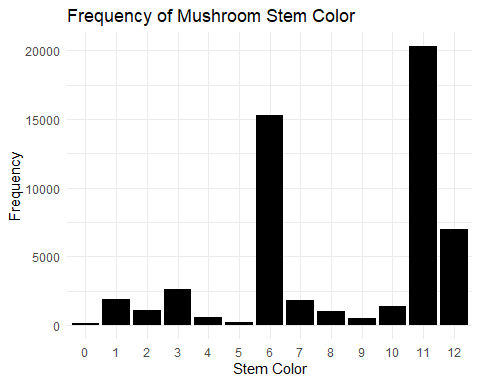
ggplot(mushrooms, aes(x = stem.width)) +  
 geom\_histogram(binwidth = 50, fill = "skyblue", color = "black") +  
 labs(x = "Stem Width", y = "Frequency", title = "Histogram of Stem Width") +  
 theme\_minimal()



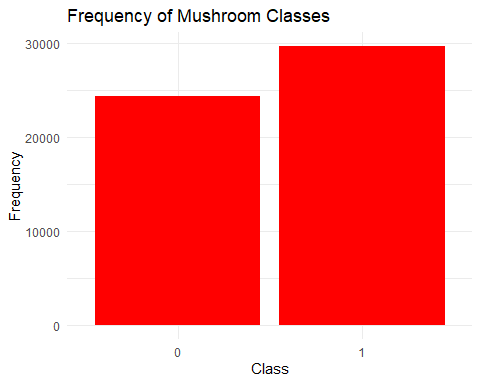
ggplot(mushrooms, aes(y = stem.width)) +  
 geom\_boxplot(fill = "skyblue", color = "black") +  
 labs(y = "Stem Width", title = "Boxplot of Stem Width") +  
 theme\_minimal()



ggplot(mushrooms, aes(x = factor(stem.color))) +  
 geom\_bar(fill = 'black') +  
 labs(x = "Stem Color", y = "Frequency", title = "Frequency of Mushroom Stem Color") +  
 theme\_minimal()



ggplot(mushrooms, aes(x = factor(class))) +  
 geom\_bar(fill = 'red') +  
 labs(x = "Class", y = "Frequency", title = "Frequency of Mushroom Classes") +  
 theme\_minimal()



#lets view a correlation matrix of all the variables  
library(corrplot)

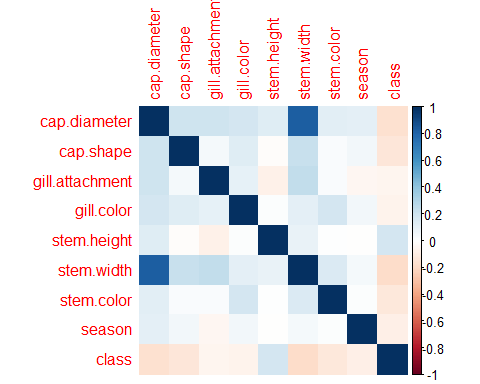
Warning: package 'corrplot' was built under R version 4.3.3

corrplot 0.92 loaded

correlation\_matrix <- cor(mushrooms)  
print(correlation\_matrix)

cap.diameter cap.shape gill.attachment gill.color  
cap.diameter 1.0000000 0.20401127 0.20048148 0.18637715  
cap.shape 0.2040113 1.00000000 0.04306624 0.13138708  
gill.attachment 0.2004815 0.04306624 1.00000000 0.10027559  
gill.color 0.1863772 0.13138708 0.10027559 1.00000000  
stem.height 0.1356524 -0.01039312 -0.07528418 0.01505654  
stem.width 0.8284693 0.22249376 0.24529991 0.11028297  
stem.color 0.1218561 0.02903463 0.02007325 0.18609020  
season 0.1133345 0.05544203 -0.04031458 0.05996487  
class -0.1656763 -0.13333831 -0.05254126 -0.06394667  
 stem.height stem.width stem.color season  
cap.diameter 0.1356524038 0.82846931 0.121856112 0.1133344523  
cap.shape -0.0103931195 0.22249376 0.029034633 0.0554420257  
gill.attachment -0.0752841798 0.24529991 0.020073247 -0.0403145793  
gill.color 0.0150565412 0.11028297 0.186090200 0.0599648711  
stem.height 1.0000000000 0.09809506 0.002623783 -0.0002919205  
stem.width 0.0980950645 1.00000000 0.157394266 0.0406794021  
stem.color 0.0026237827 0.15739427 1.000000000 0.0107499434  
season -0.0002919205 0.04067940 0.010749943 1.0000000000  
class 0.1833538889 -0.18285619 -0.128339344 -0.0829189342  
 class  
cap.diameter -0.16567630  
cap.shape -0.13333831  
gill.attachment -0.05254126  
gill.color -0.06394667  
stem.height 0.18335389  
stem.width -0.18285619  
stem.color -0.12833934  
season -0.08291893  
class 1.00000000

# correlation matrix heatmap  
corrplot(correlation\_matrix, method = "color")



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

These papers (McKay, Ebell, Billings, et al., 2020; McKay, Ebell, Dale, Shen, & Handel, 2020) 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

Leek, J. T., & Peng, R. D. (2015). Statistics. What is the question? *Science (New York, N.Y.)*, *347*(6228), 1314–1315. <https://doi.org/10.1126/science.aaa6146>

McKay, B., Ebell, M., Billings, W. Z., Dale, A. P., Shen, Y., & Handel, A. (2020). Associations Between Relative Viral Load at Diagnosis and Influenza A Symptoms and Recovery. *Open Forum Infectious Diseases*, *7*(11), ofaa494. <https://doi.org/10.1093/ofid/ofaa494>

McKay, B., Ebell, M., Dale, A. P., Shen, Y., & Handel, A. (2020). Virulence-mediated infectiousness and activity trade-offs and their impact on transmission potential of influenza patients. *Proceedings. Biological Sciences*, *287*(1927), 20200496. <https://doi.org/10.1098/rspb.2020.0496>