Fishing for Spam

Part 1: Project Proposal

Sean O’Sullivan

2024-06-15

# 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*

## 2.2 Description of data and data source

The data set(s) that I’d liked to examine are a handful of annotated data sets pertaining to fraudulent (spam) emails. They come from a handful of sources, te largest of the bunch is from the Conference on Email and Anti-Spam. That data set and a pair of others found on [Kaggle](https://www.kaggle.com/datasets/naserabdullahalam/phishing-email-dataset/data?select=CEAS_08.csv) contain text information from the subjects and bodies of the emails as well as the senders, receivers, timestamps, and whether or not URLs are contained within.

## 2.3 Questions/Hypotheses to be addressed

I’d like to explore the features that are prevalent in the identification of spam emails. Utilizing text analysis in addition to the other email attributes supplied my end goal would be to model a classifier that has a high rate of efficacy in identifying spam emails for automated detection as used by some popular email providers and large internal IT departments across the corporate landscape.

I expect that a number of keywords common in spam emails will be relevant, but it’s also likely that potential sender domains, whether the sender and receiver domain are the same, and what day of week or time of day the emails are sent are also highly correlated with intentionally spammy behavior.

While there may be some models better suited for this than others I intend to attempt a handful of classifiers and parameter hyper-tuning in order to compare and determine the best overall predictor of spam emails.

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

## 2.4 Data import and cleaning

To begin we start by downloading our data files of interest from [Kaggle](https://www.kaggle.com/datasets/naserabdullahalam/phishing-email-dataset/data?select=CEAS_08.csv). While there are a handful provided at the prior link we’re mostly interested in the 3 CSV files that contain 6 features in addition to the spam annotation labels. Next we’ll load the data into R with the read\_csv and here functions. Once each of the 3 files is individually loaded into R we append the separate dataframes to one another by way of the dplyr::bind\_rows function and start to summarize our cumulative data so that we can assess what data cleaning steps and feature generation is necessary.

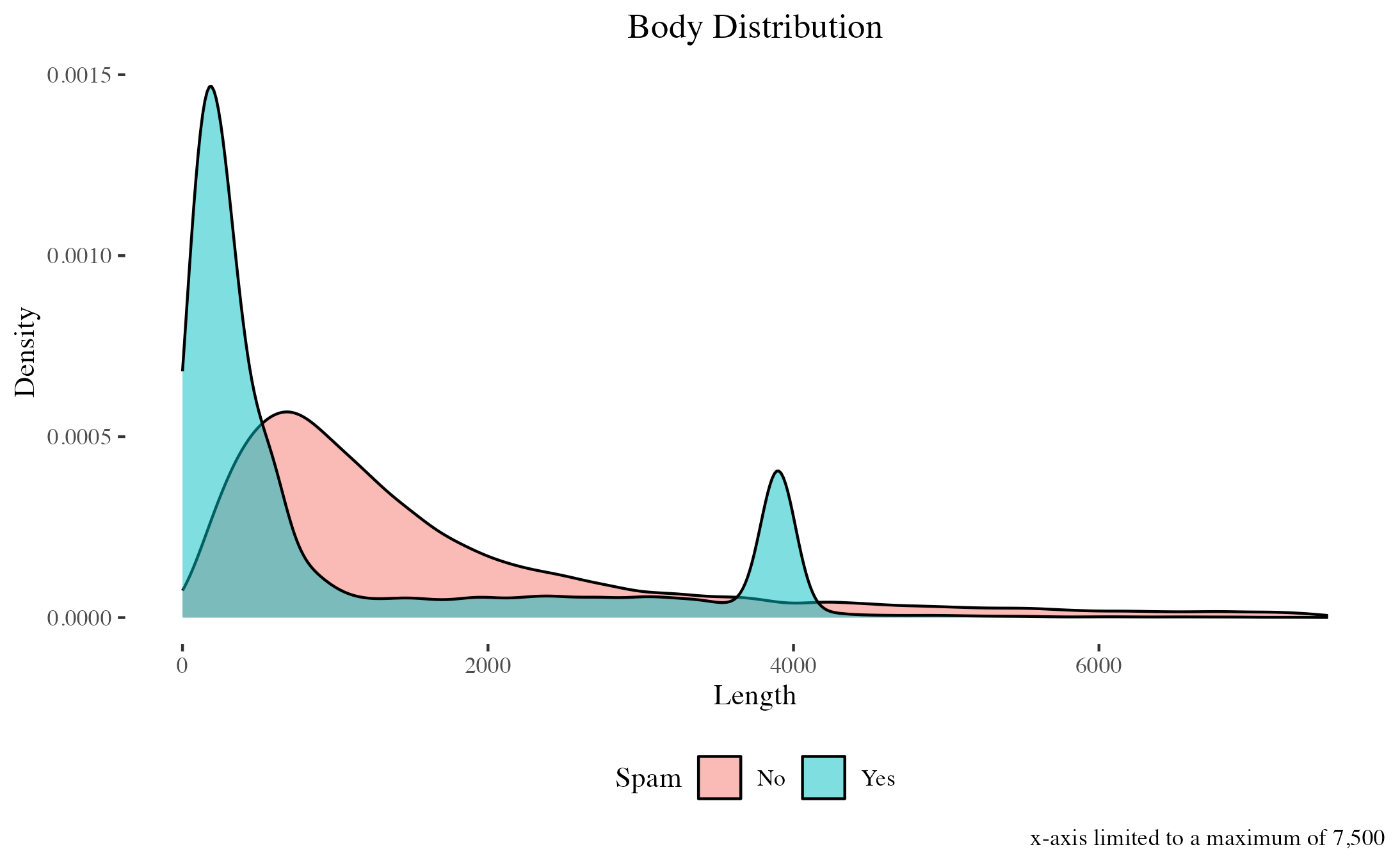
Our first step in cleaning the data centers around data type conversions, namely changing the urls and label variables to be factors rather than character data types. Next, we proceeded with cleaning and extricating sender and receiver emails from their contact names where appropriate. In order to achieve this we utilized the stringr package and regular expressions to isolate the email addresses from the contact and mutate the original columns accordingly. We’ll employee a similar technique shortly to isolate just the domain names for both sender and receiver emails.

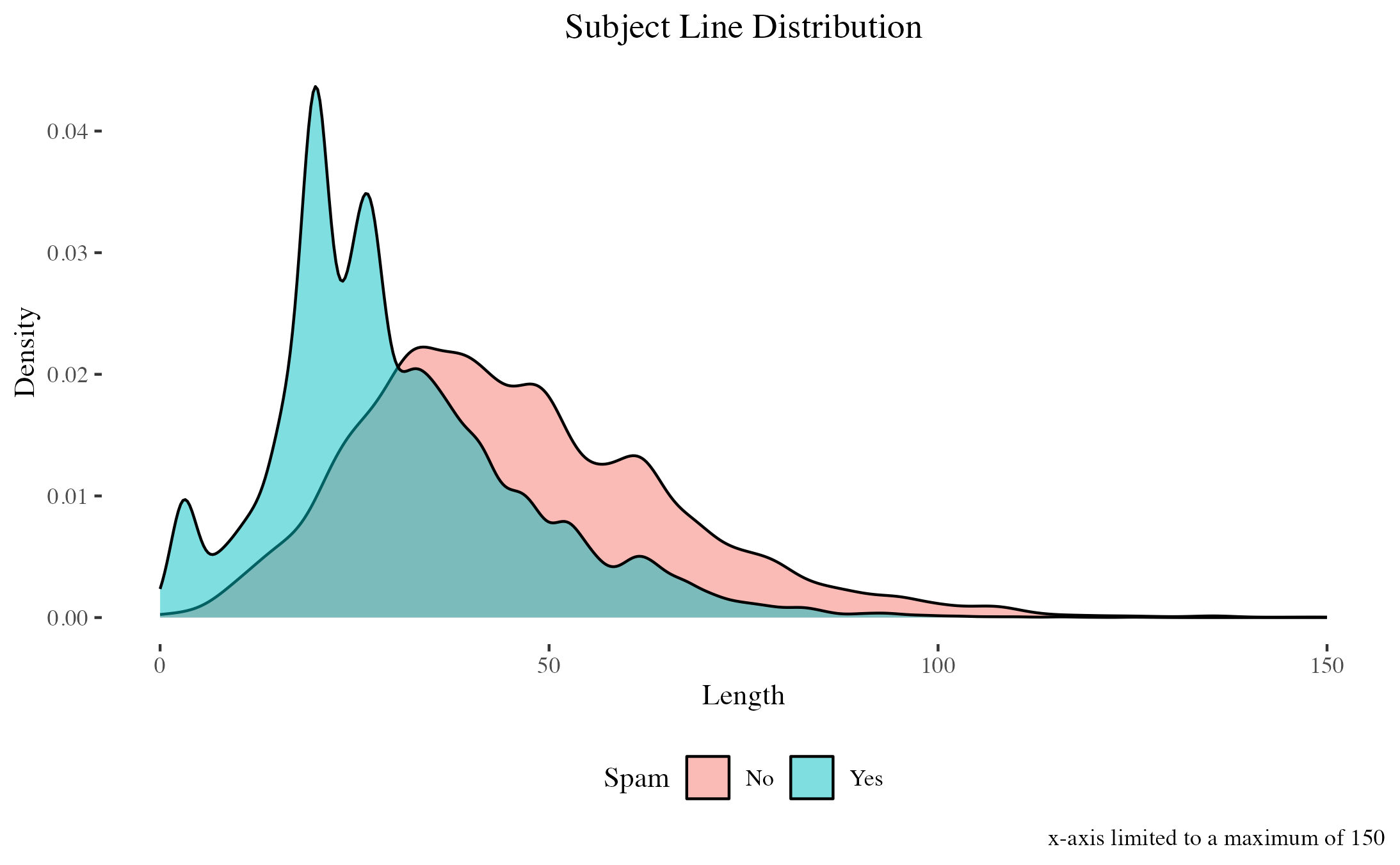
In addition to cleaning some of the existing features we also sought to generate a few simple features that we believe may be of use in our analysis and modeling stages further down the line. Specifically, we created features that measure the character length of both the subject line and email bodies as well as a boolean features that assesses whether or not the sender and receiver email were the same.

By the end of our data loading and processing we arrived at a dataset that contains 7 character columns, 2 numeric columns, 2 factor columns, and a single boolean column – taking care to ensure that our steps prevented generating any missing values where there hadn’t previously been any. Ultimately this makes for a dataset with 48,295 rows across 11 features and our variable of interest. Complete steps and code can be located in the .../code/processing/processingcode.R file found within the project.

## 2.5 Exploratory/Descriptive analysis

Once the data had been loaded and cleaned we began to examine some of the relationships between anticipated predictor variables and our variable of interest.

 First examining the distribution of email body length between spam and legitimate emails we observe a tendency for spam emails to be shorter in length. However, interestingly the distribution of spam emails is bi-modal – with a second peak at around the 4,000 character count – which may indicate that spam vendors are aware of the identification potential of email length and have begun targeting a longer length to avoid notice.

 Similarly, when we examine subject line length we observe the same tendency for spam emails to trend shorter. Subject line does not share the same distinct bi-modal shape for spam emails that the bodies do, but subject line length distribution between both spam and legitimate emails overlap in large part which may reduce it’s efficacy for classification modeling.

|  |  |
| --- | --- |
|  |  |

Finally, we examined the most popular sender domain names for vendors of both spam and legitimate emails. In doing so we see some clear indications of which domains are favorable to spam vendors and which are not. Notably Gmail fails to make the top 25 domains for spam vendors despite being the far-and-away most common choice for legitimate emails. Similarly, Yahoo and Hotmail lead the pack for spam and are not among common domains for legitimate emails. This bodes well for the classification power of sender domain when we enter the modeling phase and begin to incorporate bag-of-words and text processing.

A complete set of exploratory visualization steps utilizing ggplot2 can be located in the .../code/eda-code/eda.qmd file found within the project.

## 2.6 Statistical analysis

*Explain anything related to your statistical analyses.*

# 3. Results

**?@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.

## 3.1 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.*

**?@fig-result** shows a scatterplot figure produced by one of the R scripts.

## 3.2 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 **?@tbl-resulttable2** shows a summary of a linear model fit.

# 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.

These papers (McKay, Ebell, Billings, et al., 2020; McKay, Ebell, Dale, et al., 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.

# 5. 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>