CAPSTONE PROJECT BY TEAM 22

A DATA SCIENCE APPROACH TO FORECAST ELECTRICITY CONSUMPTION IN AUSTRALIA

John Student (z123456), Jim Student2 (zID), Jack Student3 (zID).

School of Mathematics and Statistics
UNSW Sydney

June 2021

Submitted in partial fulfilment of the requirements of the capstone course ZZSC9020

Plagiarism statement

I declare that this thesis is my own work, except where acknowledged, and has not been submitted for academic credit elsewhere.

I acknowledge that the assessor of this thesis may, for the purpose of assessing it:

- Reproduce it and provide a copy to another member of the University; and/or,
- Communicate a copy of it to a plagiarism checking service (which may then retain a copy of it on its database for the purpose of future plagiarism checking).

I certify that I have read and understood the University Rules in respect of Student Academic Misconduct, and am aware of any potential plagiarism penalties which may apply.

By signing this declaration I am agreeing to the statements and conditions above.

Signed: _____ Date: _____

Signed: _____ Date: _____

Signed: _____ Date: _____

Acknowledgements

By far the greatest thanks must go to my supervisor for the guidance, care and support they provided.

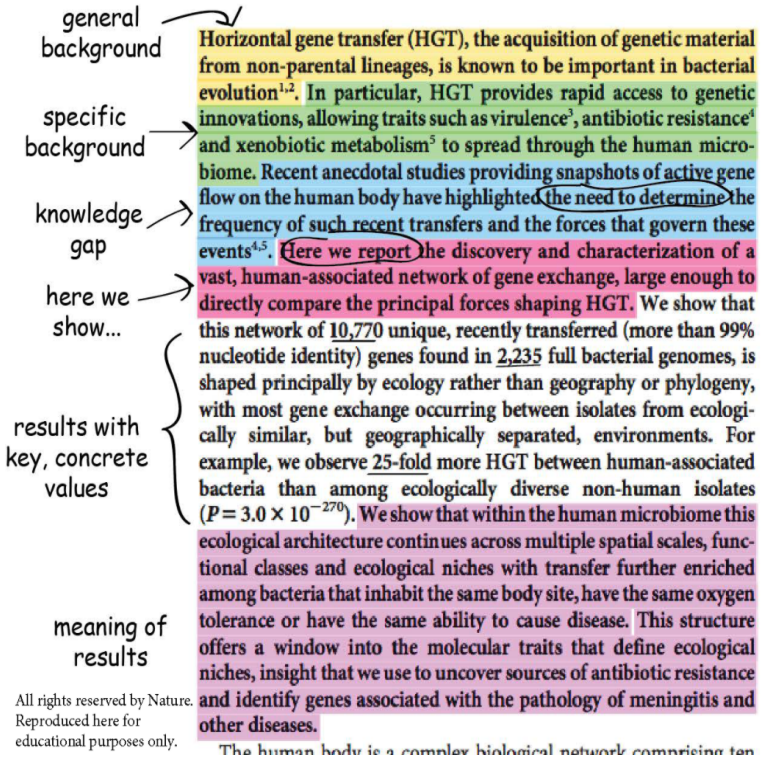
Thanks must also go to Emily, Michelle, John and Alex who helped by proof-reading the document in the final stages of preparation.

Although I have not lived with them for a number of years, my family also deserve many thanks for their encouragement. Thanks go to Robert Taggart for allowing his thesis style to be shamelessly copied.

25/07/2020.

Abstract

The image below gives you some hint about how to write a good abstract.



Horizontal gene transfer (HGT), the acquisition of genetic material from non-parental lineages, is known to be important in bacterial evolution[1,2]. In particular, HGT provides rapid access to genetic innovations, allowing traits such as virulence[3], antibiotic resistance[4] and xenobiotic metabolism[5] to spread through the human microbiome. Recent anecdotal studies providing snapshots of active gene flow on the human body have highlighted the need to determine the frequency of such recent transfers and the forces that govern these events[4,5]. Here we report the discovery and characterization of a vast, human-associated network of gene exchange, large enough to directly compare the principal forces shaping HGT. We show that this network of 10,770 unique, recently transferred (more than 99% nucleotide identity) genes found in 2,235 full bacterial genomes, is shaped principally by ecology rather than geography or phylogeny, with most gene exchange occurring between isolates from ecologically similar, but geographically separated, environments. For example, we observe 25-fold more HGT between human-associated bacteria than among ecologically diverse non-human isolates ($P = 3.0 \times 10^{-270}$). We show that within the human microbiome this ecological architecture continues across multiple spatial scales, functional classes and ecological niches with transfer further enriched among bacteria that inhabit the same body site, have the same oxygen tolerance or have the same ability to cause disease. This structure offers a window into the molecular traits that define ecological niches, insight that we use to uncover sources of antibiotic resistance and identify genes associated with the pathology of meningitis and other diseases.

Contents

Chapter 1 Introduction 1

Chapter 2 Literature Review 2

Chapter 3 Material and Methods 3
- 3.1 Software . . . 3
- 3.2 Description of the Data . . . 3
- 3.3 Pre-processing Steps . . . 3
- 3.4 Data Cleaning . . . 3
- 3.5 Assumptions . . . 3
- 3.6 Modelling Methods . . . 3

Chapter 4 Exploratory Data Analysis 4
- 4.1 Using R . . . 4
- 4.2 Using Python . . . 4

Chapter 5 Analysis and Results 6
- 5.1 A First Model . . . 6

Chapter 6 Discussion 7

Chapter 7 Conclusion and Further Issues 8

Appendix 10
- **Codes** . . . 10
- **Tables** . . . 10

Chapter 1

Introduction

This R Markdown template can be used for the ZZSC9020 course report. You can incorporate R [1] chunks and Python chunks that will be run on the fly. You can incorporate LaTeX commands.

Before submitting the last version of your report, you might want to use https://overleaf.com to collaborate with other members of your team directly on the LaTeX version of this document (which is a byproduct you get when you Knit it from studio).

We suggest you organise your report using the following chapters but, depending on your own project, nothing prevents you to have a different organisation.

Chapter 2

Literature Review

Here are a few references that can be useful: [2] and [3]. See also https://bookdown.org/yihui/rmarkdown-cookbook/

In order to incorporate your own references in this report, we strongly advise you use BibTeX. Your references then needs to be recorded in the file `references.bib`.

Chapter 3

Material and Methods

3.1 Software

R and Python of course are great software for Data Science. Sometimes, you might want to use `bash` utilities such as `awk` or `sed`.

Of course, to ensure reproducibility, you should use something like `Git` and RMarkdown (or a Jupyter Notebook). Do **not** use Word!

3.2 Description of the Data

How are the data stored? What are the sizes of the data files? How many files? etc.

3.3 Pre-processing Steps

What did you have to do to transform the data so that they become useable?

3.4 Data Cleaning

How did you deal with missing data? etc.

3.5 Assumptions

What assumptions are you making on the data?

3.6 Modelling Methods

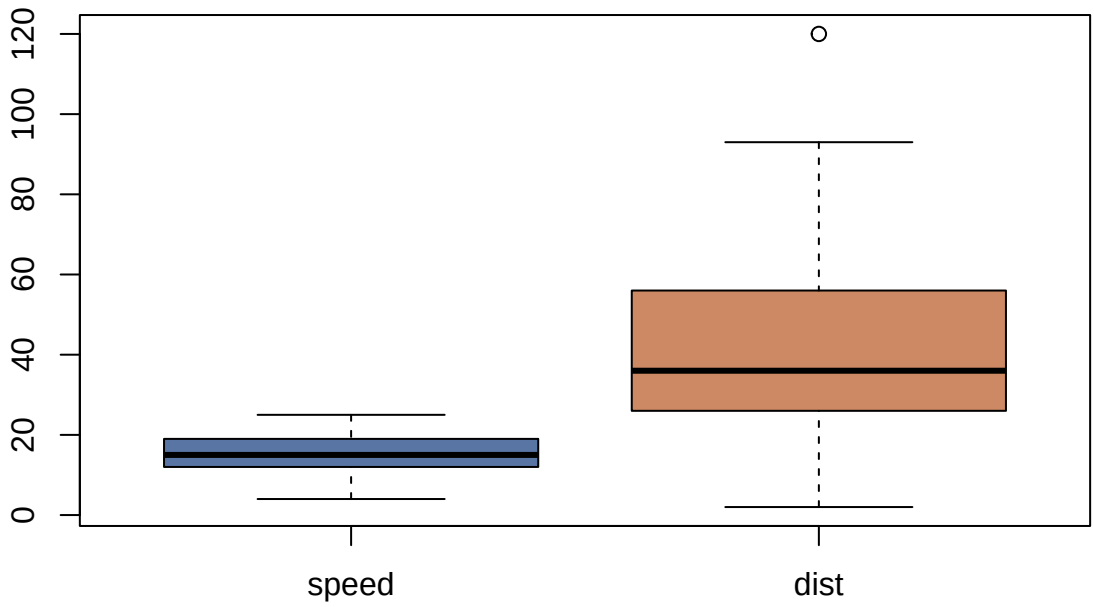
Chapter 4

Exploratory Data Analysis

This is where you explore your data using histograms, scatterplots, boxplots, numerical summaries, etc.

4.1 Using R

```
boxplot(cars, col = c("#5975a4", "#cc8963"))
```



4.2 Using Python

See https://cran.r-project.org/web/packages/reticulate/vignettes/r_markdown.html for more details.

You need to install the R package `reticulate`.

```
print("Python can be used with MATHxxxx!")
```

```
## Python can be used with MATHxxxx!
```

```
import sys
print(sys.version)
```

```
## 3.6.12 |Anaconda, Inc.| (default, Sep  9 2020, 00:29:25) [MSC v.1916 64 bit
```

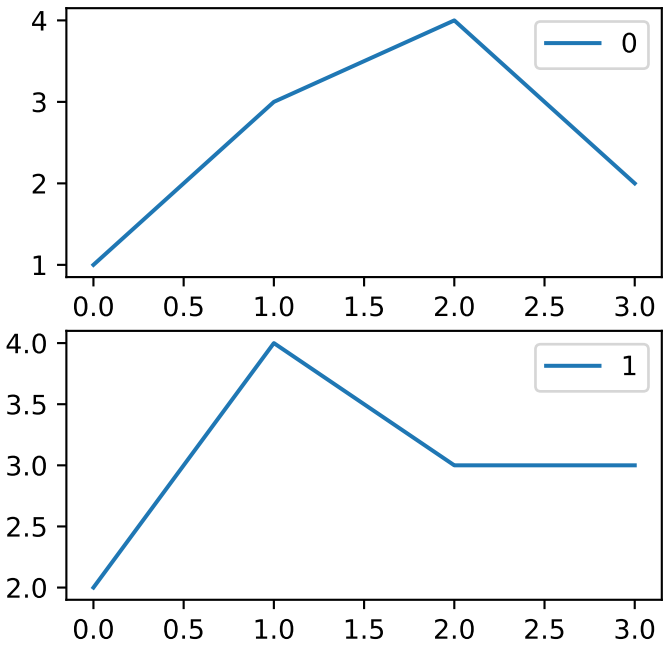
```
import numpy as np
np.random.seed(1)
np.random.normal(0.0, 1.0, size=10)
```

```
## array([ 1.62434536, -0.61175641, -0.52817175, -1.07296862,  0.86540763,
##        -2.3015387 ,  1.74481176, -0.7612069 ,  0.3190391 , -0.24937038])
```

```
import pandas as pd
import matplotlib.pyplot as plt
df=pd.DataFrame([[1, 2], [3, 4], [4, 3], [2, 3]])
fig = plt.figure(figsize=(4, 4))
for i in df.columns:
    ax=plt.subplot(2,1,i+1)
    df[[i]].plot(ax=ax)
    print(i)
```

```
## <AxesSubplot:>
## 0
## <AxesSubplot:>
## 1
```

```
plt.show()
```



Chapter 5

Analysis and Results

5.1 A First Model

Having a very simple model is always good so that you can benchmark any result you would obtain with a more elaborate model.

For example, one can use the linear regression model

$$Y_i = \beta_0 + \beta_1 x_{1i} + \cdots \beta_p x_{pi} + \epsilon_i, \qquad i = 1, \ldots, n.$$

where it is assumed that the ϵ_i's are i.i.d. $N(0, 1)$.

Chapter 6

Discussion

Put the results you got in the previous chapter in perspective with respect to the problem studied.

Chapter 7

Conclusion and Further Issues

What are the main conclusions? What are your recommendations for the "client"? What further analysis could be done in the future?

A figure:

Figure 7.1: A caption

In the text, see Figure 7.1.

References

[1] R Core Team, R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, Vienna, Austria (2017).
URL https://www.R-project.org/

[2] Y. Xie, J. Allaire, G. Grolemund, R Markdown, The Definitive Guide, Chapman and Hall/CRC, 2018.
URL https://bookdown.org/yihui/rmarkdown/

[3] P. Lafaye de Micheaux, R. Drouilhet, B. Liquet, The R Software: Fundamentals of Programming and Statistical Analysis, Statistics and Computing, Springer New York, 2013.
URL https://books.google.fr/books?id=Ji-8BAAAQBAJ

Appendix

Codes

Add you codes here.

Tables

If you have tables, you can add them here.

Use https://www.tablesgenerator.com/markdown_tables to crete very simple markdown tables, otherwise use LaTeX.

Tables	Are	Cool
col 1 is	left-aligned	$1600
col 2 is	centered	$12
col 3 is	right-aligned	$1