Tutorial 6: Refactoring R Code

Introduction

In this tutorial, you will refactor the code into separate scripts corresponding to each section. The dataset we will use comes from the `palmerpenguins` package, which contains measurements of penguins from three species.

The R programming language (R Core Team 2019) and the following R packages were used to perform the analysis: knitr (Xie 2014), tidyverse (Wickham 2017), and Quarto (Allaire et al. 2022). *Note: this report is adapted from Timbers (Timbers 2020).*

Load Libraries and Data

```
Rows: 333 Columns: 8
-- Column specification --------------------------------------------------------
Delimiter: ","
chr (3): species, island, sex
dbl (5): bill_length_mm, bill_depth_mm, flipper_length_mm, body_mass_g, year

i Use `spec()` to retrieve the full column specification for this data.
i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

Table 1: Initial penguins dataset

species	island	bill_length_mm	bill_depth_mm	flipper_length_mm	body_mass_g	sex	year
Adelie	Torgersen	39.1	18.7	181	3750	male	2007
Adelie	Torgersen	39.5	17.4	186	3800	female	2007
Adelie	Torgersen	40.3	18.0	195	3250	female	2007
Adelie	Torgersen	36.7	19.3	193	3450	female	2007
Adelie	Torgersen	39.3	20.6	190	3650	male	2007

species	island	bill_length_mm	bill_depth_mm	flipper_length_mm	body_mass_g	sex	year
Adelie	Torgersen	38.9	17.8	181	3625	female	2007

Methods

In this section, we perform exploratory data analysis (EDA) and prepare the data for modeling. Below provides tables (**glimpse?**), (**summary?**), (**clean?**), and Figure 1

```
Rows: 333
Columns: 8
$ species           <chr> "Adelie", "Adelie", "Adelie", "Adelie", "Adelie", "A~
$ island            <chr> "Torgersen", "Torgersen", "Torgersen", "Torgersen", ~
$ bill_length_mm    <dbl> 39.1, 39.5, 40.3, 36.7, 39.3, 38.9, 39.2, 41.1, 38.6~
$ bill_depth_mm     <dbl> 18.7, 17.4, 18.0, 19.3, 20.6, 17.8, 19.6, 17.6, 21.2~
$ flipper_length_mm <dbl> 181, 186, 195, 193, 190, 181, 195, 182, 191, 198, 18~
$ body_mass_g       <dbl> 3750, 3800, 3250, 3450, 3650, 3625, 4675, 3200, 3800~
$ sex               <chr> "male", "female", "female", "female", "male", "femal~
$ year              <dbl> 2007, 2007, 2007, 2007, 2007, 2007, 2007, 2007, 2007~
```

```
Rows: 1 Columns: 4
-- Column specification --------------------------------------------------------
Delimiter: ","
dbl (4): mean_bill_length, mean_bill_depth, mean_flipper_length, mean_body_mass

i Use `spec()` to retrieve the full column specification for this data.
i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

Table 2: Summary of initial dataset

mean_bill_length	mean_bill_depth	mean_flipper_length	mean_body_mass
43.99279	17.16486	200.967	4207.057

```
Rows: 333 Columns: 5
-- Column specification --------------------------------------------------------
Delimiter: ","
chr (1): species
dbl (4): bill_length_mm, bill_depth_mm, flipper_length_mm, body_mass_g
```

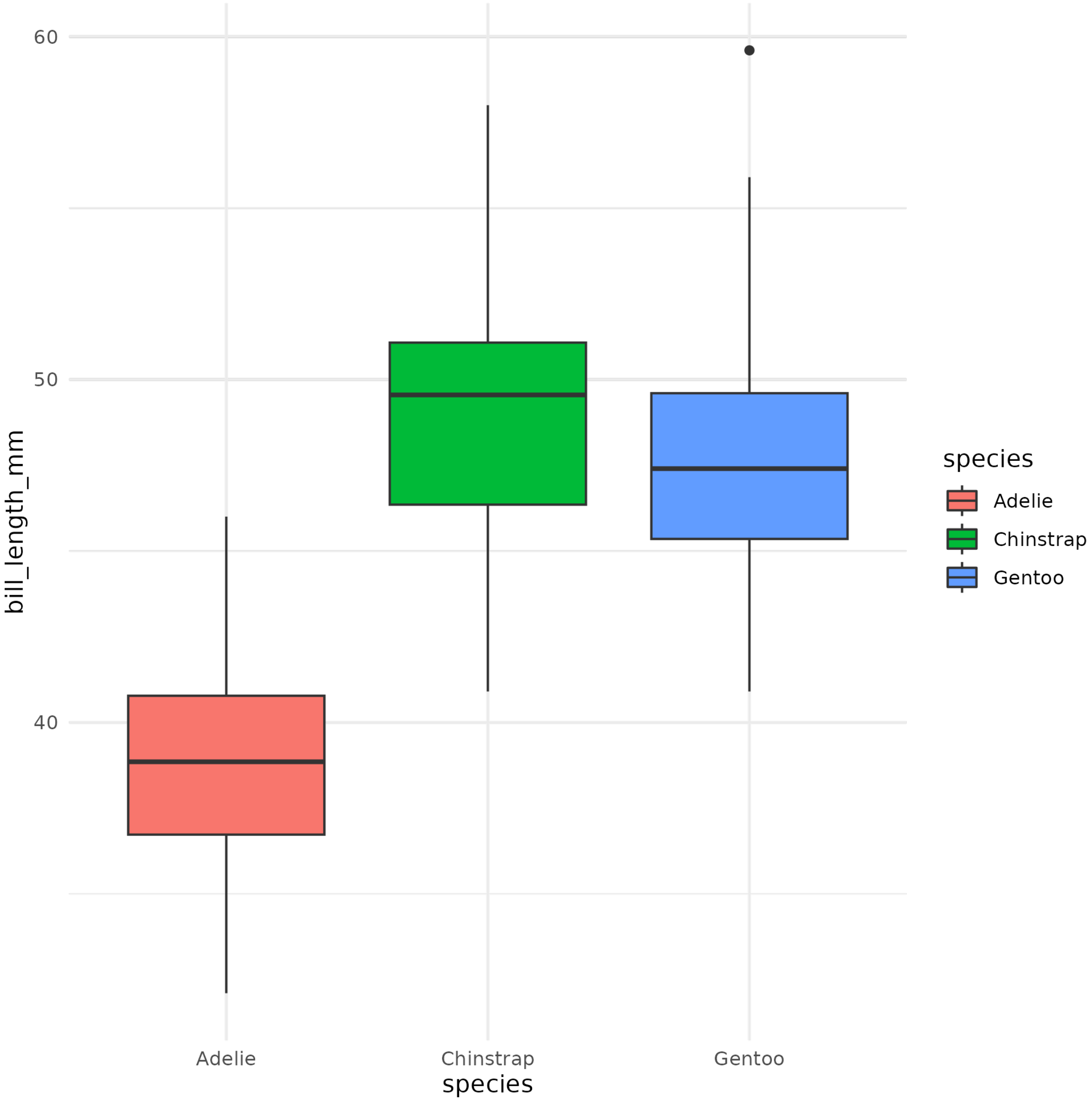

Figure 1: Boxplot of species against bill_length_mm

```
i Use `spec()` to retrieve the full column specification for this data.
i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

Table 3: Cleaned penguins dataset

species	bill_length_mm	bill_depth_mm	flipper_length_mm	body_mass_g
Adelie	39.1	18.7	181	3750
Adelie	39.5	17.4	186	3800
Adelie	40.3	18.0	195	3250
Adelie	36.7	19.3	193	3450
Adelie	39.3	20.6	190	3650
Adelie	38.9	17.8	181	3625

Model

We will fit a classification model using `tidymodels` to predict the species of a penguin based on its physical characteristics.

Table 4: Summary of fitted model

	Length	Class	Mode
pre	3	stage_pre	list
fit	2	stage_fit	list
post	1	stage_post	list
trained	1	-none-	logical

Results

We evaluate the performance of the model using the test dataset.

Table 5: Summary of fitted model

	Adelie	Chinstrap	Gentoo
Adelie	36	0	0
Chinstrap	1	17	0
Gentoo	0	0	30

Libraries Run

Test the usage of packages in the report.

```
Rows: 1 Columns: 4
-- Column specification --------------------------------------------------------
Delimiter: ","
dbl (4): mean_bill_length, mean_bill_depth, mean_flipper_length, mean_body_mass

i Use `spec()` to retrieve the full column specification for this data.
i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

Table 6: Package testings

mean_bill_length	mean_bill_depth	mean_flipper_length	mean_body_mass
43.99279	17.16486	200.967	4207.057

Conclusion

In this tutorial, we:

- Loaded and cleaned the `palmerpenguins` dataset.
- Performed exploratory data analysis.
- Built a k-Nearest Neighbors classification model using `tidymodels`.
- Evaluated the model's performance.

References

Allaire, J. J., Charles Teague, Carlos Scheidegger, Yihui Xie, and Christophe Dervieux. 2022. *Quarto* (version 1.2). https://doi.org/10.5281/zenodo.5960048.

R Core Team. 2019. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Timbers, Tiffany. 2020. *Historical Horse Population in Canada*. https://github.com/ttimbers/equine_numbers_value_canada_parameters.

Wickham, Hadley. 2017. *Tidyverse: Easily Install and Load the 'Tidyverse'*. https://CRAN.R-project.org/package=tidyverse.

Xie, Yihui. 2014. "Knitr: A Comprehensive Tool for Reproducible Research in R." In *Implementing Reproducible Computational Research*, edited by Victoria Stodden, Friedrich Leisch, and Roger D. Peng. Chapman; Hall/CRC. http://www.crcpress.com/product/isbn/9781466561595.