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 results are displayed in

Load Libraries and Data

```
-- Attaching packages --------------------------------------- tidyverse 1.3.2 --
v ggplot2 3.5.2     v purrr   1.0.4
v tibble  3.2.1     v dplyr   1.1.4
v tidyr   1.3.1     v stringr 1.5.1
v readr   2.1.3     v forcats 0.5.2
-- Conflicts ------------------------------------------ tidyverse_conflicts() --
x dplyr::filter() masks stats::filter()
x dplyr::lag()    masks stats::lag()
```

Table 1: Intial 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
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.

Glimpse at base dataset

```
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~
```

Analysis

Table 2: Summary statistics in base dataset.

mean_bill_length	mean_bill_depth	mean_flipper_length	mean_body_mass
43.99279	17.16486	200.967	4207.057

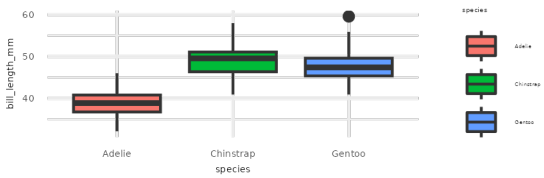

Figure 1: Boxplot of Bill Length against Species

Cleaning

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: Classification 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: Model performance.

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

Package Installation

We test out the output of the package `regexcite20250416`.

Table 6: Package usage.

Function	Output
regexcite20250416::is_leap(2000)	1
regexcite20250416::is_leap(1900)	0
regexcite20250416::temp_conv(41, 'F', 'C')	5

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.