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.

Load Libraries and Data

```
```{r setup}
# Run 01_load_data.R
library(tidyverse)
```
```

```
-- 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()
```

```
```{r setup}
data <- readr::read_csv("~/work/data/penguins.csv")
```
```

```
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.
```

````
```{r setup}
head(data)
```
````

```
# A tibble: 6 x 8
  species island    bill_length_mm bill_depth_mm flipper_length_mm body_mass_g
  <chr>   <chr>              <dbl>         <dbl>             <dbl>       <dbl>
1 Adelie  Torgersen           39.1          18.7               181        3750
2 Adelie  Torgersen           39.5          17.4               186        3800
3 Adelie  Torgersen           40.3          18                 195        3250
4 Adelie  Torgersen           36.7          19.3               193        3450
5 Adelie  Torgersen           39.3          20.6               190        3650
6 Adelie  Torgersen           38.9          17.8               181        3625
# i 2 more variables: sex <chr>, year <dbl>
```

Methods

In this section, we perform exploratory data analysis (EDA) and prepare the data for modeling.

```
# Run 02_methods.R
model_data <- readr::read_csv("~/work/data/penguins_model.csv") %>%
  dplyr::mutate(species = as.factor(species))
```

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

```
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.
```

```
head(model_data)
```

```
# A tibble: 6 x 5
  species bill_length_mm bill_depth_mm flipper_length_mm body_mass_g
  <fct>            <dbl>         <dbl>             <dbl>       <dbl>
1 Adelie            39.1          18.7               181        3750
2 Adelie            39.5          17.4               186        3800
3 Adelie            40.3          18                 195        3250
4 Adelie            36.7          19.3               193        3450
5 Adelie            39.3          20.6               190        3650
6 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.

```
# Run 03_model.R
penguin_fit <- readr::read_rds("~/work/output/penguin_fit.RDS")
summary(penguin_fit)
```

```
        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.

```
# Run 04_results.R
conf_mat <- readr::read_rds("~/work/output/conf_mat.RDS")
conf_mat
```

```
$table
           Truth
Prediction  Adelie Chinstrap Gentoo
  Adelie        36         0      0
  Chinstrap      1        17      0
  Gentoo         0         0     30

attr(,"class")
[1] "conf_mat"
```

Package Installation

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

```
# Run 05_package.R
func_outputs <- readr::read_csv("~/work/output/func_outputs.csv")
```

```
Rows: 3 Columns: 2
-- Column specification --------------------------------------------------------
Delimiter: ","
chr (1): Function
dbl (1): Output

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.
```

```
func_outputs
```

```
# A tibble: 3 x 2
  Function                                    Output
  <chr>                                        <dbl>
1 regexcite20250416::is_leap(2000)                 1
2 regexcite20250416::is_leap(1900)                 0
3 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.