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

```
library(readr)
penguins <- read_csv("data/penguins.csv")
```

```
Rows: 333 Columns: 7
-- Column specification --------------------------------------------------------
Delimiter: ","
chr (3): species, island, sex
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(penguins)
```

```
# A tibble: 6 x 7
  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 1 more variable: sex <chr>
```

Methods

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

Summary statistics

```
library(readr)
sum_stats <- read_csv("results/tables/summary_stats.csv")
```

```
Rows: 1 Columns: 2
-- Column specification --------------------------------------------------------
Delimiter: ","
dbl (2): mean_bill_length, mean_bill_depth

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(sum_stats)
```

```
# A tibble: 1 x 2
  mean_bill_length mean_bill_depth
             <dbl>           <dbl>
1             44.0            17.2
```

Summary plot

Model

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

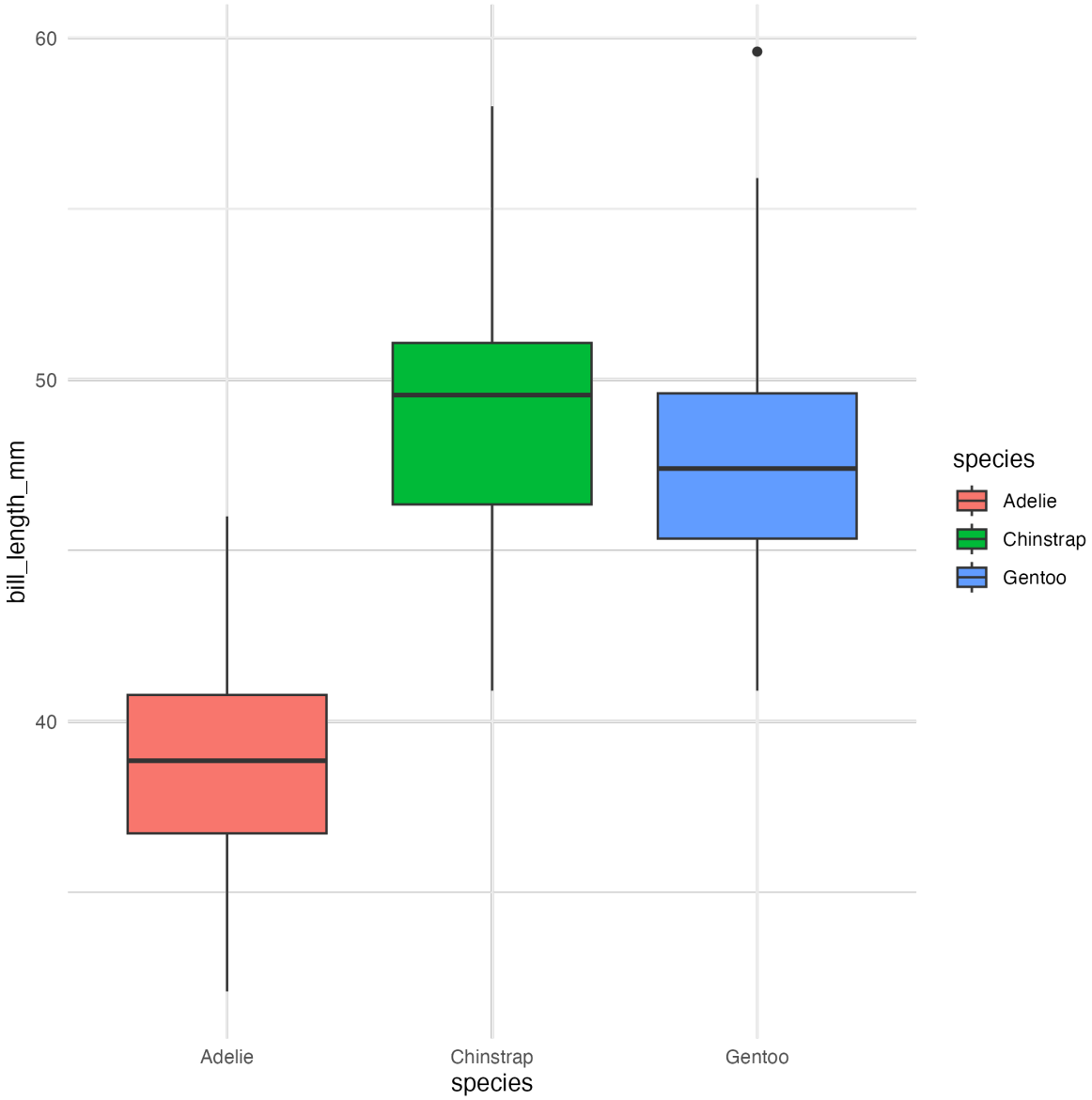

Figure 1: Bill Length by species

```r
train <- read_csv("data/train.csv")
```

```
Rows: 249 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.
```

```r
test <- read_csv("data/test.csv")
```

```
Rows: 84 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.
```

```r
head(train)
```

```
# A tibble: 6 x 5
  species bill_length_mm bill_depth_mm flipper_length_mm body_mass_g
  <chr>            <dbl>         <dbl>             <dbl>       <dbl>
1 Adelie            40.3          18                 195        3250
2 Adelie            36.7          19.3               193        3450
3 Adelie            38.9          17.8               181        3625
4 Adelie            39.2          19.6               195        4675
5 Adelie            41.1          17.6               182        3200
6 Adelie            38.6          21.2               191        3800
```

```r
head(test)
```

```
# A tibble: 6 x 5
  species bill_length_mm bill_depth_mm flipper_length_mm body_mass_g
  <chr>            <dbl>         <dbl>             <dbl>       <dbl>
```

```
1 Adelie            39.1          18.7               181        3750
2 Adelie            39.5          17.4               186        3800
3 Adelie            39.3          20.6               190        3650
4 Adelie            36.6          17.8               185        3700
5 Adelie            35.9          19.2               189        3800
6 Adelie            38.2          18.1               185        3950
```

Results

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

```
preds <- read_csv("results/tables/predictions.csv")
```

```
Rows: 84 Columns: 6
-- Column specification --------------------------------------------------------
Delimiter: ","
chr (2): .pred_class, 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.
```

```
conf <- read_csv("results/tables/confusion_matrix.csv")
```

```
Rows: 9 Columns: 3
-- Column specification --------------------------------------------------------
Delimiter: ","
chr (2): Prediction, Truth
dbl (1): n

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(preds)
```

```
# A tibble: 6 x 6
  .pred_class species bill_length_mm bill_depth_mm flipper_length_mm body_mass_g
  <chr>       <chr>            <dbl>         <dbl>             <dbl>       <dbl>
1 Adelie      Adelie            39.1          18.7               181        3750
```

```
2 Adelie      Adelie            39.5          17.4                186         3800
3 Adelie      Adelie            39.3          20.6                190         3650
4 Adelie      Adelie            36.6          17.8                185         3700
5 Adelie      Adelie            35.9          19.2                189         3800
6 Adelie      Adelie            38.2          18.1                185         3950
```

```
head(conf)
```

```
# A tibble: 6 x 3
  Prediction Truth         n
  <chr>      <chr>     <dbl>
1 Adelie     Adelie       36
2 Chinstrap  Adelie        1
3 Gentoo     Adelie        0
4 Adelie     Chinstrap     0
5 Chinstrap  Chinstrap    17
6 Gentoo     Chinstrap     0
```

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.