Penguins Data Analysis

Yousuf Ahmed Nahian

10/10/2025

Introduction

This report analyzes the Palmer Penguins dataset using the tidyverse in R. The goal is to explore, summarize, and visualize key patterns in penguin species through data manipulation and graphical analysis.

```
head(penguins)
```

```
## # A tibble: 6 x 8
##   species island    bill_length_mm bill_depth_mm flipper_length_mm body_mass_g
##   <fct>   <fct>              <dbl>         <dbl>             <int>       <int>
## 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           NA            NA                  NA          NA
## 5 Adelie  Torgersen           36.7          19.3               193        3450
## 6 Adelie  Torgersen           39.3          20.6               190        3650
## # i 2 more variables: sex <fct>, year <int>
```

```
summary(penguins)
```

```
##       species          island    bill_length_mm  bill_depth_mm
##  Adelie   :152   Biscoe   :168   Min.   :32.10   Min.   :13.10
##  Chinstrap: 68   Dream    :124   1st Qu.:39.23   1st Qu.:15.60
##  Gentoo   :124   Torgersen: 52   Median :44.45   Median :17.30
##                                  Mean   :43.92   Mean   :17.15
##                                  3rd Qu.:48.50   3rd Qu.:18.70
##                                  Max.   :59.60   Max.   :21.50
##                                  NA's   :2       NA's   :2
##  flipper_length_mm  body_mass_g       sex           year
##  Min.   :172.0     Min.   :2700   female:165   Min.   :2007
##  1st Qu.:190.0     1st Qu.:3550   male  :168   1st Qu.:2007
##  Median :197.0     Median :4050   NA's  : 11   Median :2008
##  Mean   :200.9     Mean   :4202                Mean   :2008
##  3rd Qu.:213.0     3rd Qu.:4750                3rd Qu.:2009
##  Max.   :231.0     Max.   :6300                Max.   :2009
##  NA's   :2         NA's   :2
```

This gives a quick overview of the data and gives summary statistics like mean, median, missing values

```r
penguins_clean <- penguins %>% drop_na
penguins_clean
```

```
## # A tibble: 333 x 8
##    species island    bill_length_mm bill_depth_mm flipper_length_mm body_mass_g
##    <fct>   <fct>              <dbl>         <dbl>             <int>       <int>
##  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
##  7 Adelie  Torgersen           39.2          19.6               195        4675
##  8 Adelie  Torgersen           41.1          17.6               182        3200
##  9 Adelie  Torgersen           38.6          21.2               191        3800
## 10 Adelie  Torgersen           34.6          21.1               198        4400
## # i 323 more rows
## # i 2 more variables: sex <fct>, year <int>
```

This cleans our data and removes any rows with missing values.

```r
adelie_penguins <- penguins_clean %>% filter(species == "Adelie")
adelie_penguins
```

```
## # A tibble: 146 x 8
##    species island    bill_length_mm bill_depth_mm flipper_length_mm body_mass_g
##    <fct>   <fct>              <dbl>         <dbl>             <int>       <int>
##  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
##  7 Adelie  Torgersen           39.2          19.6               195        4675
##  8 Adelie  Torgersen           41.1          17.6               182        3200
##  9 Adelie  Torgersen           38.6          21.2               191        3800
## 10 Adelie  Torgersen           34.6          21.1               198        4400
## # i 136 more rows
## # i 2 more variables: sex <fct>, year <int>
```

This line selects only Adelie penguins from the cleaned dataset so we can analyze them separately.

```r
adelie_data <- adelie_penguins %>% select(species, flipper_length_mm, body_mass_g)
adelie_data
```

```
## # A tibble: 146 x 3
##    species flipper_length_mm body_mass_g
##    <fct>               <int>       <int>
##  1 Adelie                181        3750
##  2 Adelie                186        3800
##  3 Adelie                195        3250
```

```
##  4 Adelie                193        3450
##  5 Adelie                190        3650
##  6 Adelie                181        3625
##  7 Adelie                195        4675
##  8 Adelie                182        3200
##  9 Adelie                191        3800
## 10 Adelie                198        4400
## # i 136 more rows
```

We do this to keep only the columns we need, making the dataset simpler and easier to work with.

```
adelie_data <- adelie_data %>% mutate(body_mass_kg = body_mass_g / 1000)
adelie_data
```

```
## # A tibble: 146 x 4
##    species flipper_length_mm body_mass_g body_mass_kg
##    <fct>               <int>       <int>        <dbl>
##  1 Adelie                181        3750         3.75
##  2 Adelie                186        3800         3.8
##  3 Adelie                195        3250         3.25
##  4 Adelie                193        3450         3.45
##  5 Adelie                190        3650         3.65
##  6 Adelie                181        3625         3.62
##  7 Adelie                195        4675         4.68
##  8 Adelie                182        3200         3.2
##  9 Adelie                191        3800         3.8
## 10 Adelie                198        4400         4.4
## # i 136 more rows
```

We convert the body mass to kg from grams.

```
penguins_summary <- penguins_clean %>%
  group_by(species) %>%
  summarise(
    avg_body_mass = mean(body_mass_g),
    avg_flipper_length = mean(flipper_length_mm),
    count = n()
  )
penguins_summary
```

```
## # A tibble: 3 x 4
##   species   avg_body_mass avg_flipper_length count
##   <fct>             <dbl>              <dbl> <int>
## 1 Adelie            3706.               190.   146
## 2 Chinstrap         3733.               196.    68
## 3 Gentoo            5092.               217.   119
```

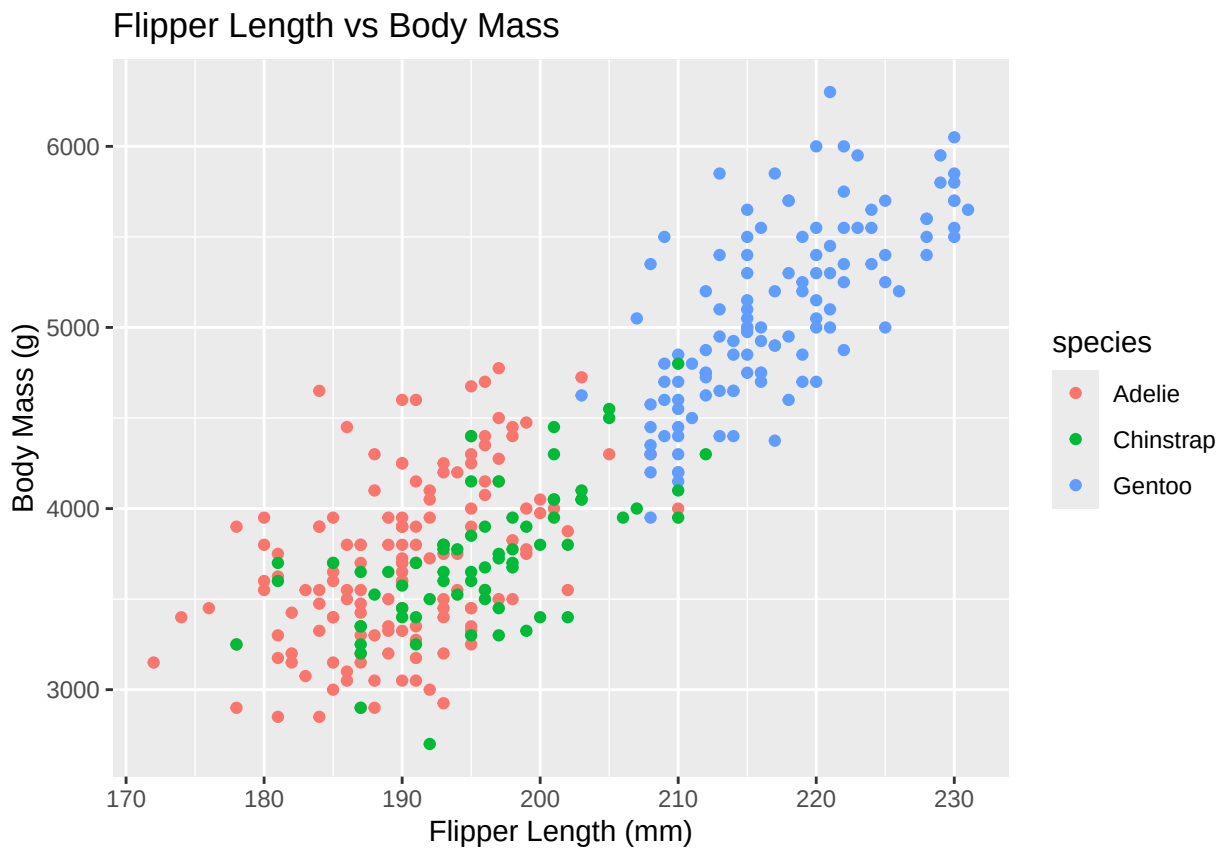
This calculates the average body mass, average flipper length, and number of penguins and saves it as a small summary table.

```r
penguins_summary <- penguins_summary %>%
  arrange(desc(avg_body_mass))
penguins_summary
```

```
## # A tibble: 3 x 4
##   species   avg_body_mass avg_flipper_length count
##   <fct>             <dbl>              <dbl> <int>
## 1 Gentoo            5092.               217.   119
## 2 Chinstrap         3733.               196.    68
## 3 Adelie            3706.               190.   146
```

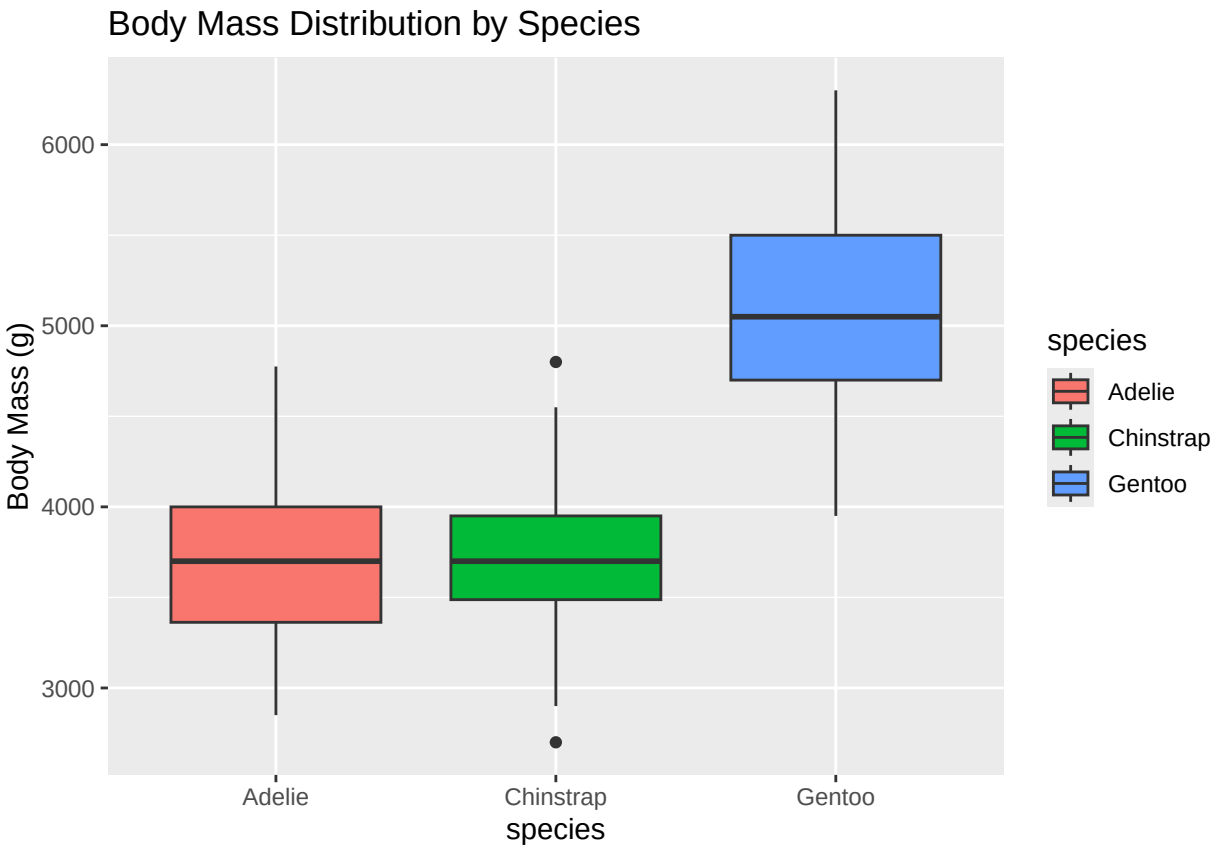
This sorts our data from heaviest to lightest.

```r
ggplot(penguins_clean, aes(x = flipper_length_mm, y = body_mass_g, color = species)) +
  geom_point() +
  labs(title = "Flipper Length vs Body Mass", x = "Flipper Length (mm)", y = "Body Mass (g)")
```



We create a Scatter plot which helps us to see relationships between Flipper Length and Body Mass.

```r
ggplot(penguins_clean, aes(x = species, y = body_mass_g, fill = species)) +
  geom_boxplot() +
  labs(title = "Body Mass Distribution by Species", y = "Body Mass (g)")
```



We now created a boxplot which helps us to see median, quartiles, and outliers which helps us to compare body mass across species.

Summary

Adelie penguins are generally smaller than Gentoo and Chinstrap penguins.

Flipper length is positively correlated with body mass.

Gentoo penguins are the heaviest species on average.