Exploring Data with ggplot2

Minh Long

2024-11-10

Introduction

In this guide, we'll explore using `ggplot2` for data visualization, focusing on data from the **Palmer Penguins** and **Diamonds** datasets. We'll use several plotting techniques to bring out meaningful patterns in our data.

Setting Up the Environment

Before we begin, let's load the required libraries. Here, we'll use the `tidyverse` package for plotting and data manipulation, and the `palmerpenguins` package for accessing the Palmer Penguins dataset.

```
library(tidyverse)
```

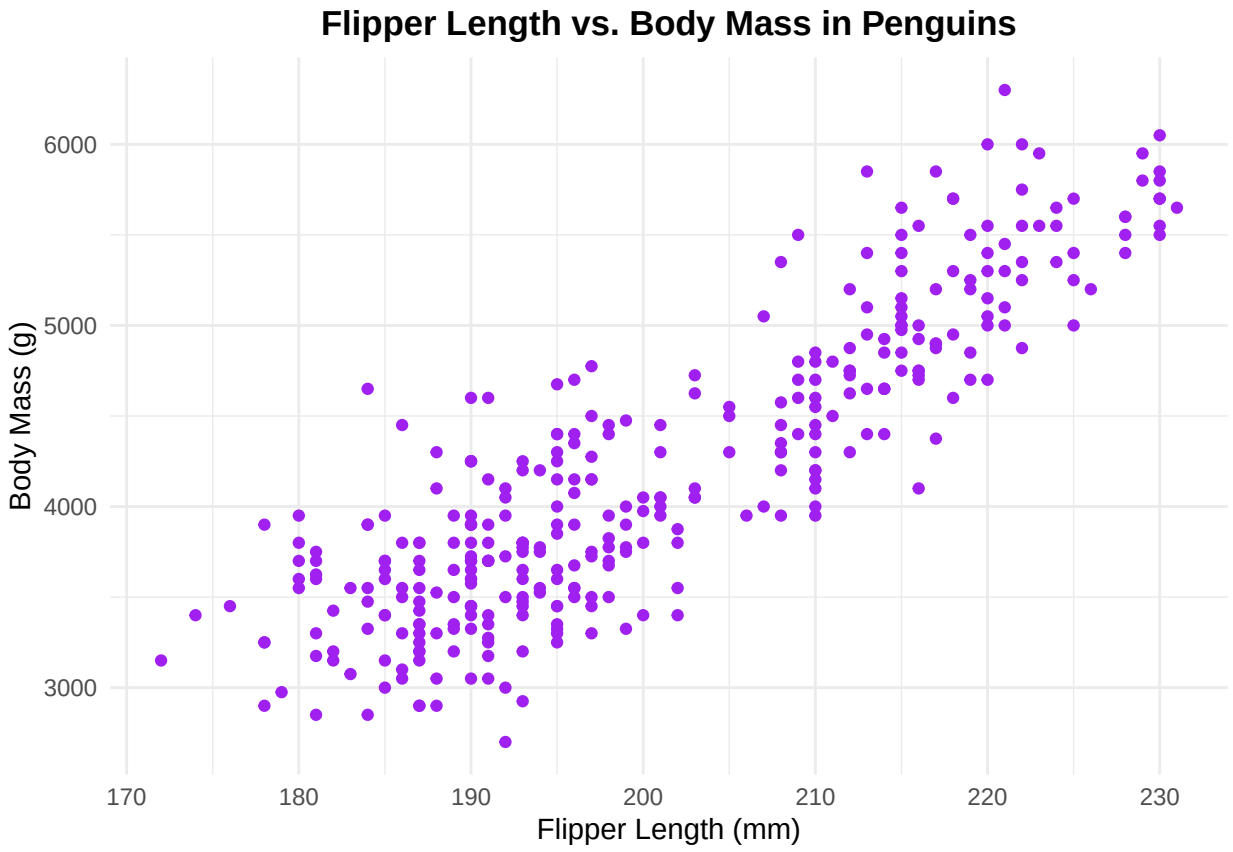
```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.5
## v forcats   1.0.0     v stringr   1.5.1
## v ggplot2   3.5.1     v tibble    3.2.1
## v lubridate 1.9.3     v tidyr     1.3.1
## v purrr     1.0.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
library(palmerpenguins)
```

Exploring Flipper Length vs. Body Mass in Penguins

In this section, we will visualize and the relationship between flipper length and body mass in penguins. We will use purple color points to emphasize our data.

```
ggplot(data = penguins) +
  geom_point(mapping = aes(x = flipper_length_mm, y = body_mass_g), color = "purple") +
  labs(
    title = "Flipper Length vs. Body Mass in Penguins",
    x = "Flipper Length (mm)",
    y = "Body Mass (g)"
  ) +
  theme_minimal() +
  theme(plot.title = element_text(hjust = 0.5, face = "bold"))
```



Interpretation

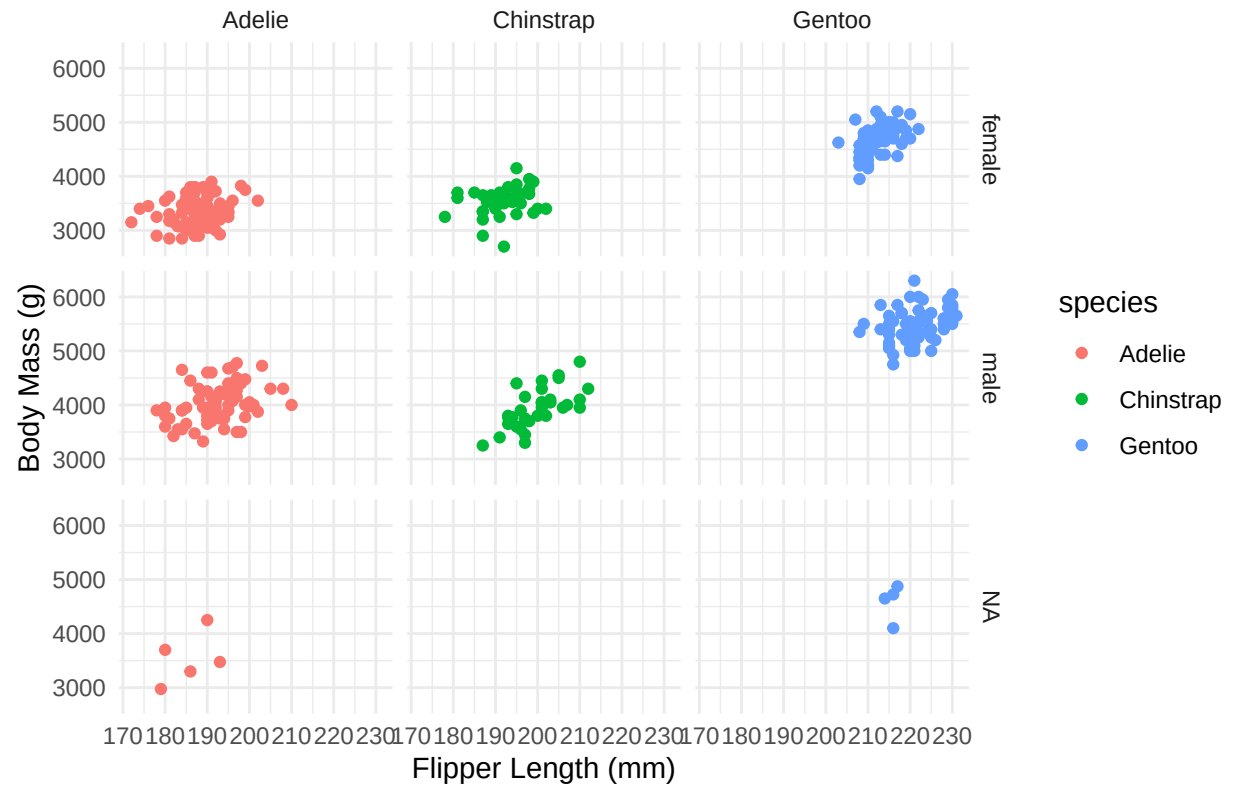
This scatter plot allows us to see if there is a correlation between a penguins's flipper length and its body mass.

Comparing Penguin Species by Sex

To dive deeper, we will create a face grid examine flipper length and body mass across different species and sexes. This comparison highlights variations within and across species.

```
ggplot(data = penguins) +
  geom_point(mapping = aes(x = flipper_length_mm, y = body_mass_g, color = species)) +
  facet_grid(sex ~ species) +
  labs(
    title = "Comparison of Flipper Length and Body Mass by Species and Sex",
    x = "Flipper Length (mm)",
    y = "Body Mass (g)"
  ) +
  theme_minimal() +
  theme(plot.title = element_text(hjust = 0.5, face = "bold"))
```

```
## Warning: Removed 2 rows containing missing values or values outside the scale range
## (`geom_point()`).
```



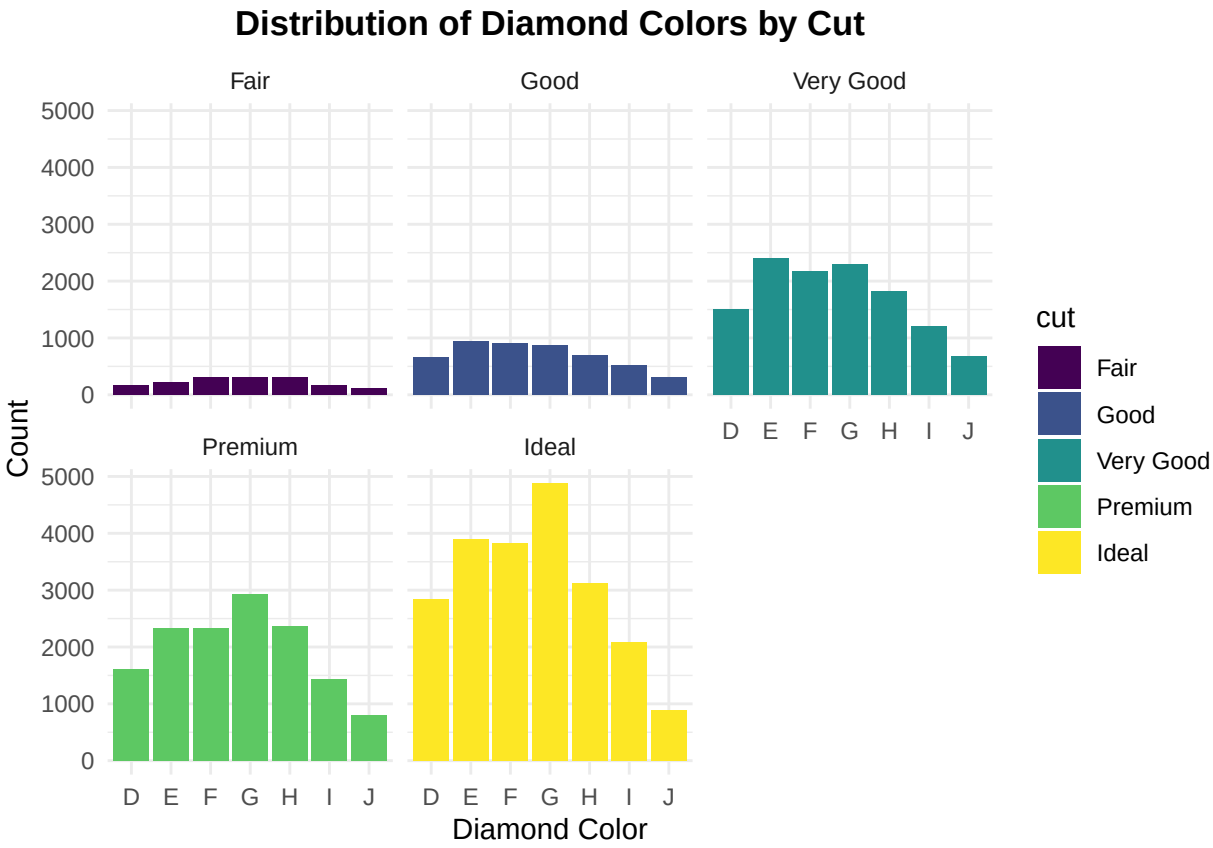
Interpretation

This faceted plot reveals differences in body mass and flipper length between male and female penguins of each species.

Analyzing Diamond Color and Cut

Now let's switch to the Diamonds dataset to explore the distribution of diamond colors by their cut quality. This bar plot provides insight into how color distribution varies across cuts.

```
ggplot(data = diamonds) +
  geom_bar(mapping = aes(x = color, fill = cut), position = "dodge") +
  facet_wrap(~ cut) +
  labs(
    title = "Distribution of Diamond Colors by Cut",
    x = "Diamond Color",
    y = "Count"
  ) +
  theme_minimal() +
  theme(plot.title = element_text(hjust = 0.5, face = "bold"))
```



Interpretation

Each facet represents a different cut type, allowing us to see the variation in color distribution for each cut quality.

Conclusion

Explanation of Changes:

1. **Title and Introduction**: Added an introductory section for context.
2. **Visualization Titles**: Each visualization has a title and improved axis labels for clarity.
3. **Interpretation Sections**: Added short explanations after each plot to guide the reader's interpretation.
4. **Themes**: Used `theme_minimal()` and centered plot titles for a cleaner look.
5. **Consistency in Code**: Improved code formatting and used `labs()` to define titles and labels for readability.