Describing & Comparing Distributions

Mr. Merrick · September 28, 2025

SOCS Checklist

S — Shape: modality (uni/bi/multi), symmetry vs. skew, clusters/gaps. *ECDF tips:* steep = high density, flat = gap/tail, early rise = right-skew, late rise = left-skew.

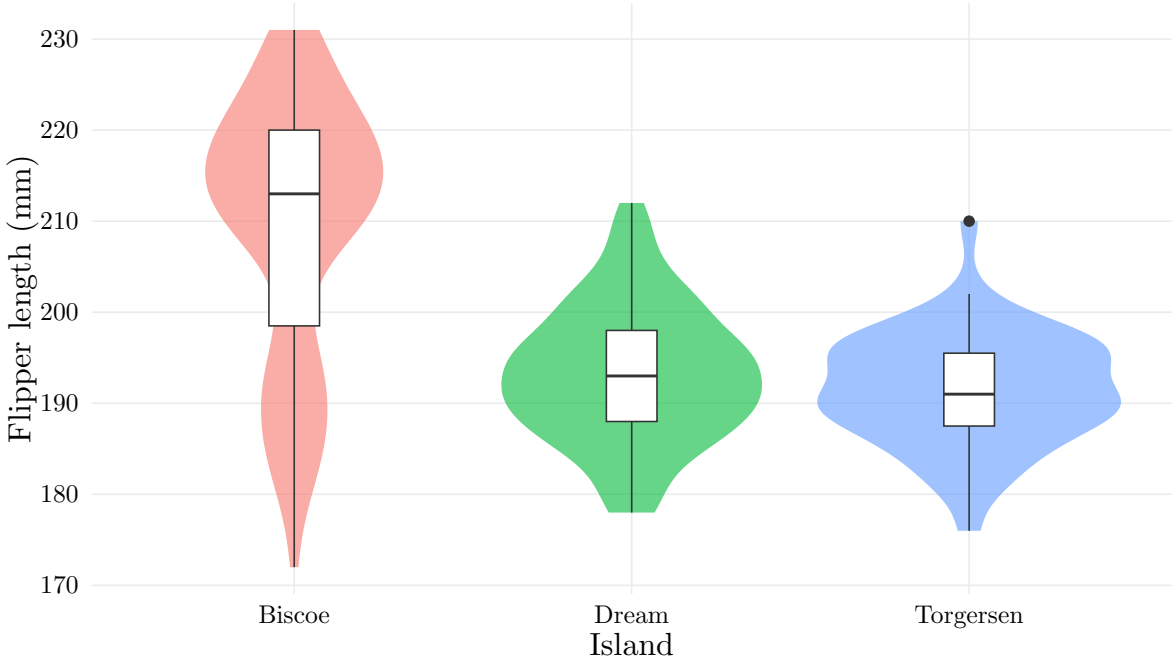
O — Outliers: unusual or extreme values, isolated points or small clusters. Gaps.

C — Center: Use median: ECDF at $F(x) = 0.5$, or boxplot/violin median.

S — Spread: Use range as a single number.

1. Flipper Length by Island (Penguins)

Task. Describe and compare the distributions of flipper length (mm) for the three islands.



Solution (SOCS): Context: Distribution of *penguin flipper length* for the islands *Biscoe*, *Dream*, and *Torgersen* (units: millimeters).

Shape: *Biscoe* shows **two peaks**, one around 185–195 mm and another around 210–220 mm, indicating two distinct clusters of penguins. *Dream* is unimodal and roughly symmetric, centered in the mid-190s. *Torgersen* is unimodal with a slight right tail.

Outliers: *Torgersen* has one high outlier near ∼210 mm; no clearly isolated points for *Biscoe* or *Dream*.

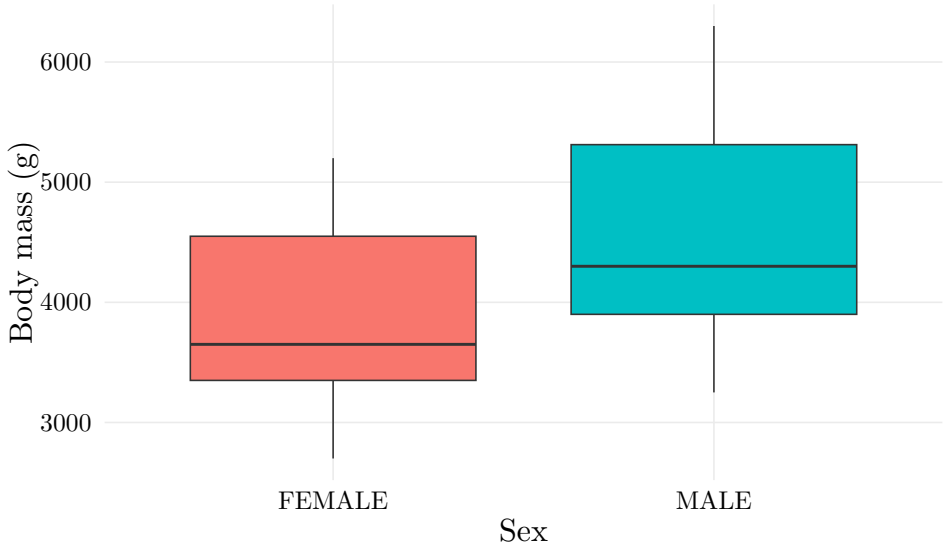
Center: median(Biscoe) ≈ 213–215 mm > median(Dream) ≈ 193–195 mm > median(Torgersen) ≈ 190–192 mm.

Spread: *Biscoe* varies most: IQR ≈ 15–20 mm; range ≈ 45 mm. *Dream* is intermediate: IQR ≈ 10 mm; range ≈ 20–25 mm. *Torgersen* is tightest: IQR ≈ 8–10 mm; range ≈ 25–30 mm including its outlier.

Conclusion: Compared to Dream and Torgersen, *Biscoe* penguins tend to have the **longest and most variable** flipper lengths, with clear evidence of two peaks. *Dream* penguins are shorter and more consistent. *Torgersen* penguins are the **shortest overall** with the least variability, aside from a single high outlier.

2. Body Mass by Sex (Penguins)

Task. Compare male vs. female body mass distributions.



Solution (SOCS): Context: Distribution of *penguin body mass* by sex (units: grams).

Shape: Because only a boxplot is shown, we cannot determine detailed shape (e.g., unimodal, skewed). We can only compare centers, spreads, and possible outliers.

Outliers: No extreme outlier points are plotted for either group.

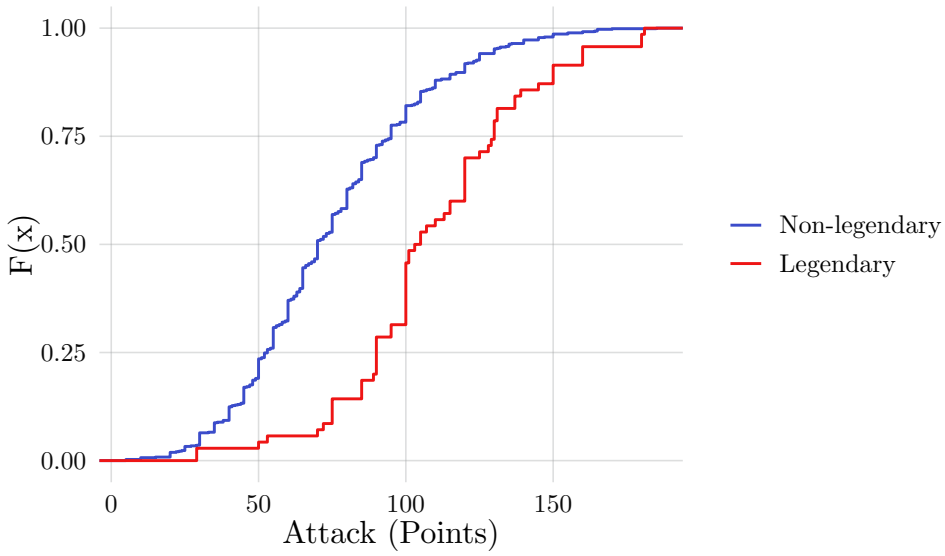
Center: The median male body mass ($\approx$ 4300–4400 g) is greater than the median female body mass ($\approx$ 3600–3700 g).

Spread: Male body mass is more variable: IQR $\approx$ 1400 g compared to IQR $\approx$ 1100 g for females. Overall range for males is about 3100–6200 g ($\approx$ 3100 g), while for females it is about 2800–5200 g ($\approx$ 2400 g). There is overlap: the upper quartile of females extends into the lower quartile of males.

Conclusion: Male penguins tend to be **heavier** and show **more variability** in body mass than females. However, the overlap means some females are heavier than lighter males.

3. Empirical CDF of Attack (Pokémon)

Task. Using the ECDF, describe and compare the distributions of Attack for Legendary vs. Non-legendary Pokémon.



Solution (SOCS): Context: Distribution of *Pokémon Attack scores* by Legendary status (units: points). The empirical CDF $F(x)$ shows how the distributions accumulate: regions of steep rise correspond to concentrations of values, while flatter regions indicate tails or sparse areas.

Shape: The *Non-legendary* ECDF rises fairly smoothly, with its steepest increase between about 70–110 points, suggesting a roughly symmetric unimodal distribution centered in this range. The *Legendary* ECDF has steeper rises near 90–100 and again near 120–130 points, suggesting concentrations of values in those regions. Its right tail also extends farther, showing mild right skew.

Outliers: Neither group shows isolated jumps far from the bulk, but the gradual flattening in the upper tails (above ∼150 for Non-legendary and ∼160 for Legendary) indicates a few unusually high Attack scores.

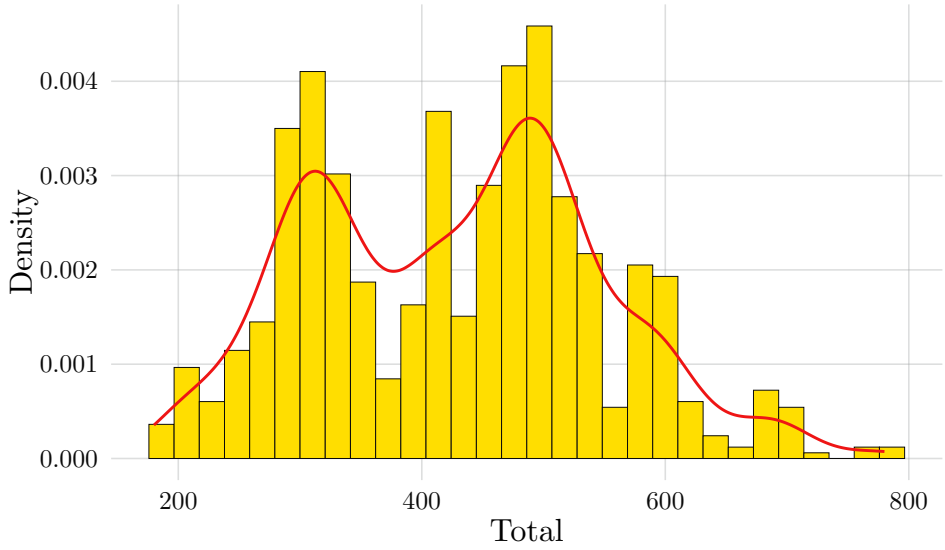
Center: At $F(x) = 0.5$, the median for Non-legendary Pokémon is about 95 points, while the median for Legendary Pokémon is higher, about 118 points.

Spread: Non-legendary Pokémon are more variable overall. Their interquartile range (IQR) is about $110 - 70 = 40$ points, with total range about $170 - 10 = 160$ points. For Legendaries, the IQR is about $130 - 100 = 30$ points, with range about $185 - 55 = 130$ points.

Conclusion: Legendary Pokémon tend to have **higher Attack scores** (points) than Non-legendary Pokémon, with evidence of concentrated values around 90–100 and 120–130. Non-legendary Pokémon are more variable overall, while Legendaries extend farther to the right but are less spread out in the middle.

5. Histogram of Total (Pokémon) Score with KDE

Task. Describe the distribution of the overall `Total` Pokémon scores.



Solution (SOCS): Context: Distribution of *Pokémon Total score* (units: points) shown with a histogram and an overlaid kernel density curve.

Shape: The distribution is **bimodal**. One mode occurs near 320–340 points and a second, larger mode occurs near 500–520 points. A noticeable trough appears around 380–430 points. A thinner right tail extends beyond 650 up to about 780, indicating a slight right skew overall.

Outliers: A histogram does not identify individual outliers, but the very sparse bars above 700 suggest a few unusually high totals.

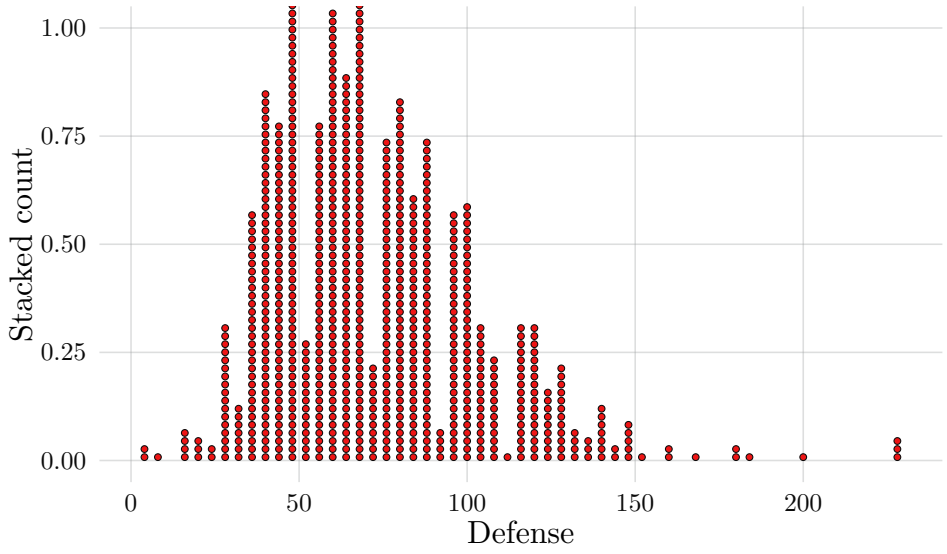
Center: Because the distribution is bimodal, a single "typical" value is less representative. The overall median lies in the mid- to high-400s, around 470–490 points.

Spread: The distribution spans from roughly 180 up to 780 points, giving a total range of about 600 points. The middle 50% of values appear to lie between about 360 and 560 points, for an estimated IQR $\approx$ 200 points.

Conclusion: Total scores reveal two distinct ability tiers—one around the low 300s and another around the low 500s—with a modest right tail that includes a few exceptionally strong Pokémon.

6. Dotplot of Defense (Pokémon)

Task. Describe the distribution for Pokemon defense scores.



Solution (SOCS): Context: Distribution of *Pokémon Defense* values (units: points), shown with a stacked dotplot.

Shape: The distribution is **unimodal**, with the bulk of values concentrated between about 50 and 90 points. The left side rises steeply from near 0, while the right side declines more gradually, indicating a **right-skewed** distribution. The discreteness of the scoring (integers) produces visible vertical stacks.

Outliers and Unusual Features: A handful of Pokémon have exceptionally high Defense scores above 150, with the maximum near 230. These are isolated relative to the main cluster. There are also **gaps** in the upper range (e.g., few or no values between about 180 and 200), which stand out compared to the dense central cluster.

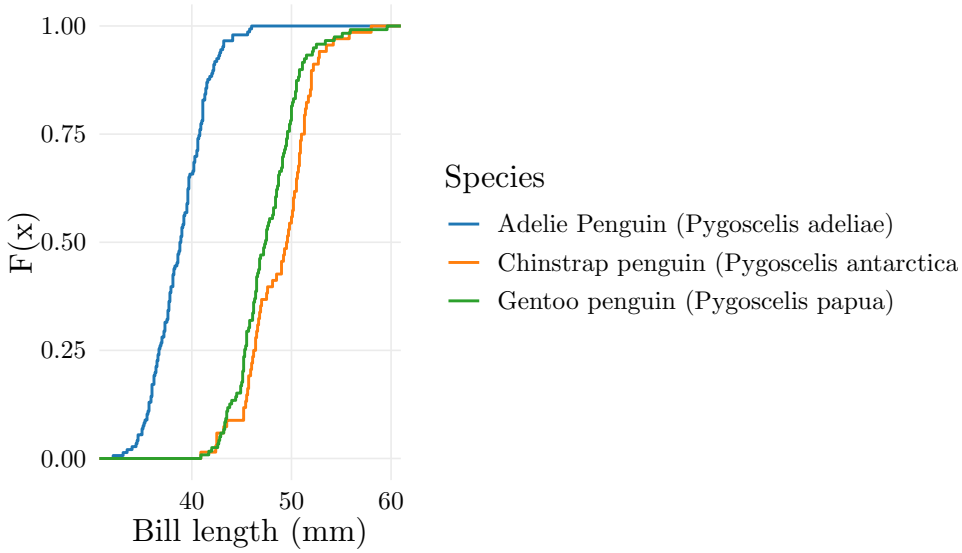
Center: The median Defense is about 65 points, lying near the middle of the dense cluster.

Spread: The distribution spans from about 5 up to 230 points, giving a total **range** of roughly 225 points. However, most observations fall within about 40 to 100, indicating that while the range is wide, the majority of Pokémon are concentrated in a narrower band.

Conclusion: Most Pokémon have moderate Defense values (50–90 points), but the distribution is **right-skewed**, with several unusually high values above 150 and visible gaps in the upper range.

7. ECDF of Bill Length (Penguins) by Species

Task. Use the ECDF to describe and compare the distributions of bill length across species.



Solution (SOCS): Context: Distribution of *penguin bill length* by species—Adélie, Chinstrap, and Gentoo—measured in millimeters and shown as ECDFs. Steeper ECDF segments indicate many values in a narrow range (higher density), while flatter stretches indicate tails or sparser regions.

Shape: The *Adélie* (blue) and *Gentoo* (green) curves are nearly parallel, suggesting similar shapes (roughly symmetric and unimodal). Gentoo is essentially a **right-shifted** version of Adélie, reflecting longer bills. The *Chinstrap* (orange) curve shows two noticeably steep rises (around the low 40s and near 50 mm), suggesting concentrations of values and possible clustering.

Outliers / Unusual Features: ECDFs do not mark outliers individually. However, the flat extreme tails (very small mass below 35 mm or above 55 mm) indicate that only a few penguins in each species have unusually short or long bills.

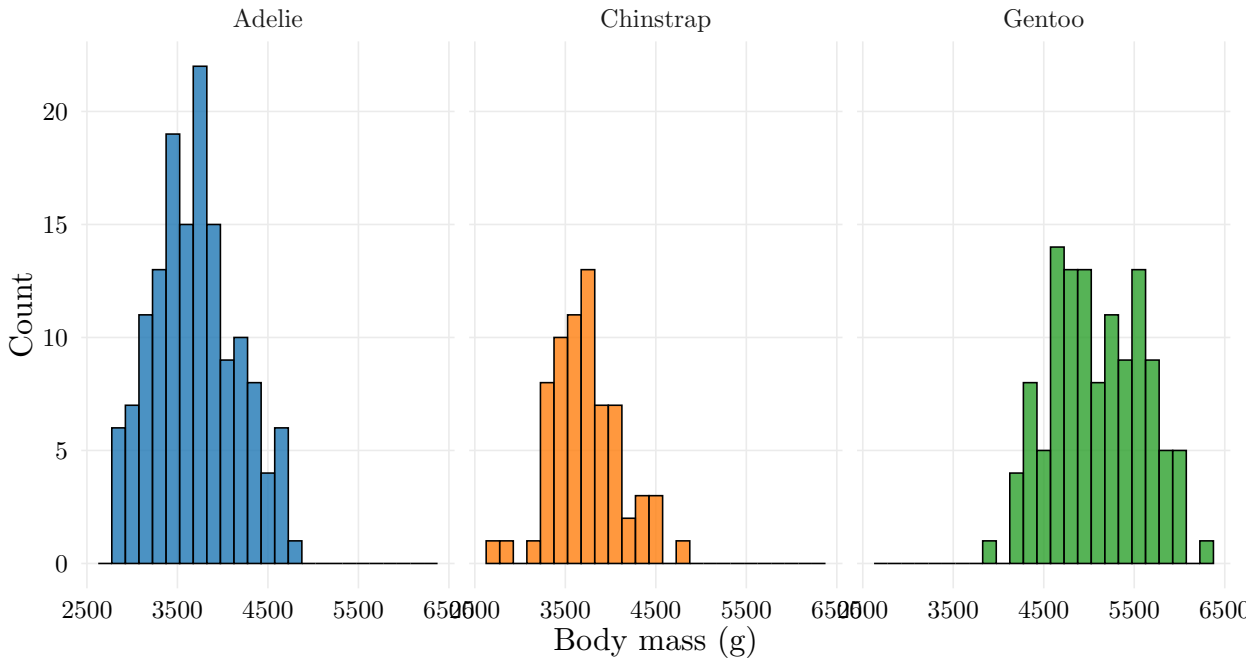
Center: From $F(x) = 0.5$, the medians are about 39 mm for Adélie, 47–48 mm for Gentoo, and 49–50 mm for Chinstrap. Thus, Chinstrap > Gentoo > Adélie in terms of typical bill length.

Spread: Using the horizontal distance between $F = 0.25$ and $F = 0.75$, the IQRs are similar across species: Adélie $\approx$ 13 mm, Gentoo $\approx$ 19 mm (widest), and Chinstrap $\approx$ 12 mm. Overall ranges (by eye) are about 33–46 mm (Adélie), 43–55 mm (Chinstrap), and 43–60 mm (Gentoo).

Conclusion: Chinstrap penguins tend to have the **longest bills**, slightly longer on average than Gentoo, while Adélie penguins have the **shortest bills**. Gentoo and Adélie share a similar distributional shape, with Gentoo shifted to larger values. Gentoo also shows the widest spread. Chinstrap exhibits some irregular clustering but overall has bill lengths comparable in variability to the others.

9. Body Mass Histograms by Species (Facets)

Task. Compare the distributions of body mass for *Adélie*, *Chinstrap*, and *Gentoo* penguins.



Solution (SOCS): Context: Distribution of *penguin body mass* by species—Adélie, Chinstrap, and Gentoo—measured in grams and displayed with histograms.

Shape: *Adélie* is roughly bell-shaped and fairly symmetric around its center. *Chinstrap* is unimodal, roughly symmetric with a mild right skew. *Gentoo* is also roughly bell-shaped, with a slightly longer right tail.

Outliers / Unusual Features: All three species have a few unusually heavy individuals in their far right tails (Adélie above ∼4600 g, Chinstrap above ∼4900 g, Gentoo above ∼6000 g). There are also minor **gaps** visible: Adélie near 3250 g, Gentoo around 4000 g and near 5200 g.

Center: The medians are approximately Adélie ∼3700 g, Chinstrap ∼3800 g, and Gentoo ∼5100 g. Thus Gentoo penguins have much higher typical body mass than the other two, which are quite similar.

Spread: From the visible supports: Adélie spans about 2900–4700 g (range ≈1800 g), Chinstrap about 3200–5000 g (range ≈1800 g), and Gentoo about 4100–6300 g (range ≈2200 g). Gentoo shows the greatest variability; Adélie and Chinstrap have similar, smaller spreads.

Conclusion: Gentoo penguins are the **heaviest and most variable** in body mass. Adélie and Chinstrap are lighter with similar centers and spreads. All three species are roughly unimodal, with Gentoo and Chinstrap showing minor gaps and right tails.

Data Sources

- **Pokémon with stats** (Kaggle): https://www.kaggle.com/datasets/abcsds/pokemon.
- **Palmer Archipelago (Antarctica) penguin data** (Kaggle mirror of the `palmerpenguins` dataset): https://www.kaggle.com/datasets/parulpandey/palmer-archipelago-antarctica-penguin-data.