Describing & Comparing Distributions

Mr. Merrick · September 22, 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.

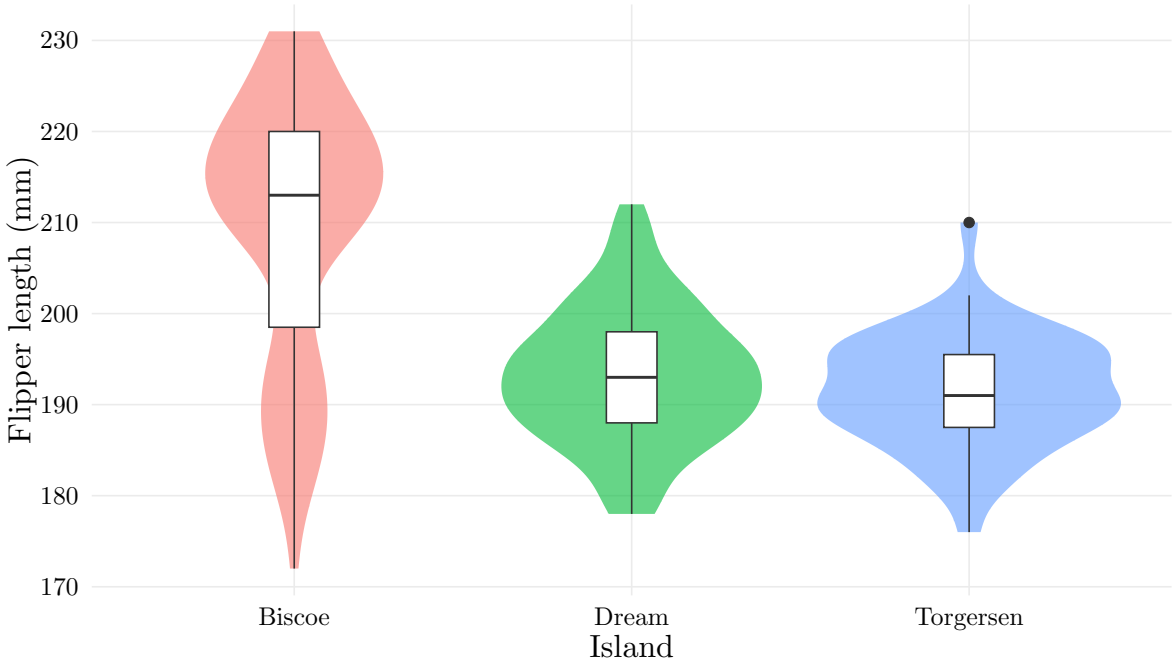
O — Outliers: unusual or extreme values, isolated points or small clusters. Gaps.

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: All three are roughly unimodal. *Biscoe* shows a main cluster around 210–215 mm with a secondary bump near ∼190 mm and a longer left tail, making it slightly left-skewed. *Dream* is unimodal with a mild right tail. *Torgersen* is unimodal, roughly symmetric and similar in shape to Dream.

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

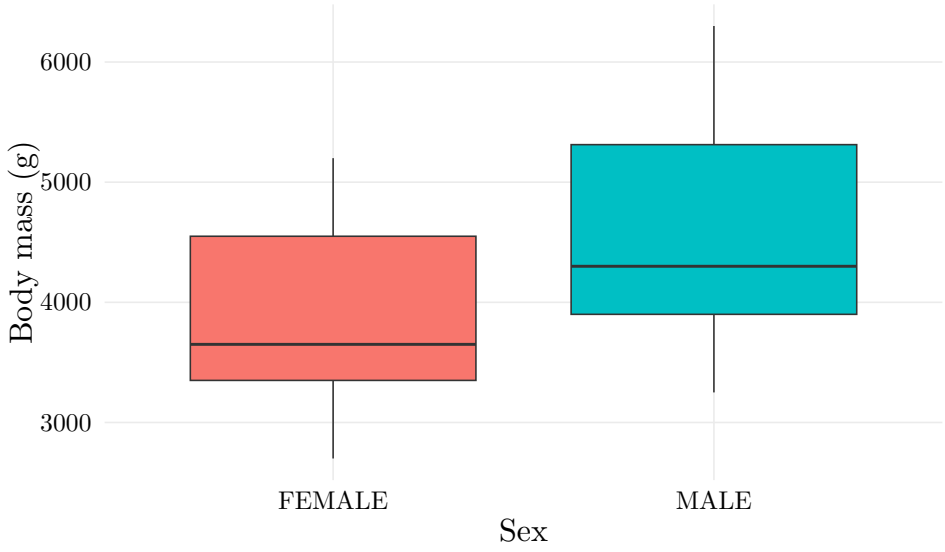
Center: median(Biscoe) ≈ 213–215 mm > median(Dream) ≈ 193–195 mm > median(Torgersen) ≈ 190–192 mm.

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

Conclusion: Compared to Dream and Torgersen, *Biscoe* penguins tend to have longer and more variable flippers (in mm), with evidence of two clusters; *Torgersen* are shortest with least variability.

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: A boxplot does not reveal the detailed shape (unimodal, skewed, etc.), so we cannot comment on shape from this display alone.

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

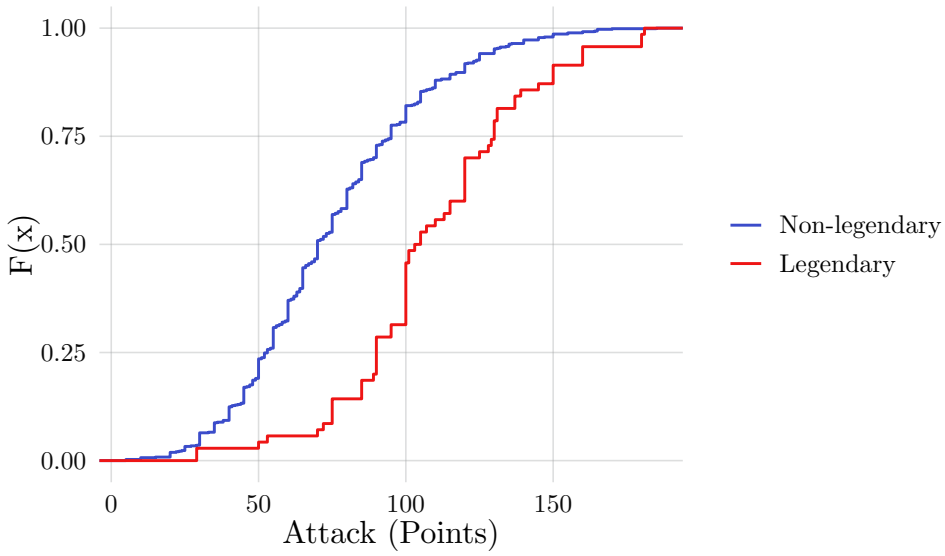
Center: The median male body mass ($\approx$ 4300–4400 g) is greater than the median female body mass ($\approx$ 3600–3700 g).

Spread: The male IQR is larger ($\approx$ 1400 g) than the female IQR ($\approx$ 1100 g). The overall range for males is about $6200 - 3100 = 3100$ g; for females about $5200 - 2800 = 2400$ g. There is overlap: the upper quartile of females overlaps the lower quartile of males.

Conclusion: In context, male penguins tend to be heavier and more variable in body mass than females, but some females are heavier than lighter males due to the overlap.

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 ECDF $F(x)$ allows us to infer distribution shape: regions of steep rise correspond to peaks in the underlying histogram, while flatter stretches indicate tails or gaps.

Shape: The *Non-legendary* ECDF increases fairly smoothly and symmetrically around its middle values (steepest rise around 70–110 points), suggesting a roughly symmetric unimodal distribution centered in that range. The *Legendary* ECDF shows two distinct steep increases, one near 90–100 and another around 120–130 points, which indicates two modes. Its right tail also extends farther, showing mild right skew.

Outliers: Neither group has isolated jumps far from the bulk, but the flatter upper tails (above ∼150 for Non-legendary and ∼160 for Legendary) indicate a few unusually high Attack values.

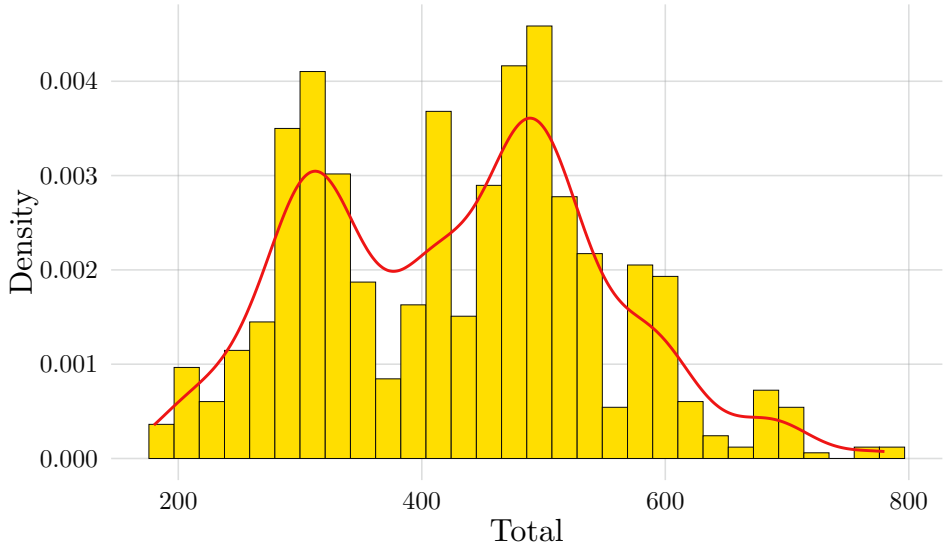
Center: At $F(x) = 0.5$, the median for Non-legendary Pokémon is about **90–95** points, while the median for Legendary Pokémon is higher, about **115–120** points.

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

Conclusion: In context, Legendary Pokémon tend to have *higher Attack scores (points)* than Non-legendary Pokémon, with evidence of bimodality among Legendaries. Non-legendary Attack scores are more variable but roughly symmetric around their median.

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: Clearly *bimodal*. One mode is near $\sim$ 320–340 points and a second, larger mode near $\sim$ 500–520 points. There is a visible dip/gap in the counts around $\sim$ 380–430 points. A thinner tail extends to the right beyond $\sim$ 650 toward $\sim$ 780, indicating a slight *right skew* overall.

Outliers: A histogram does not mark individual outliers, but the very sparse bars beyond $\sim$ 700 suggest a few unusually high totals.

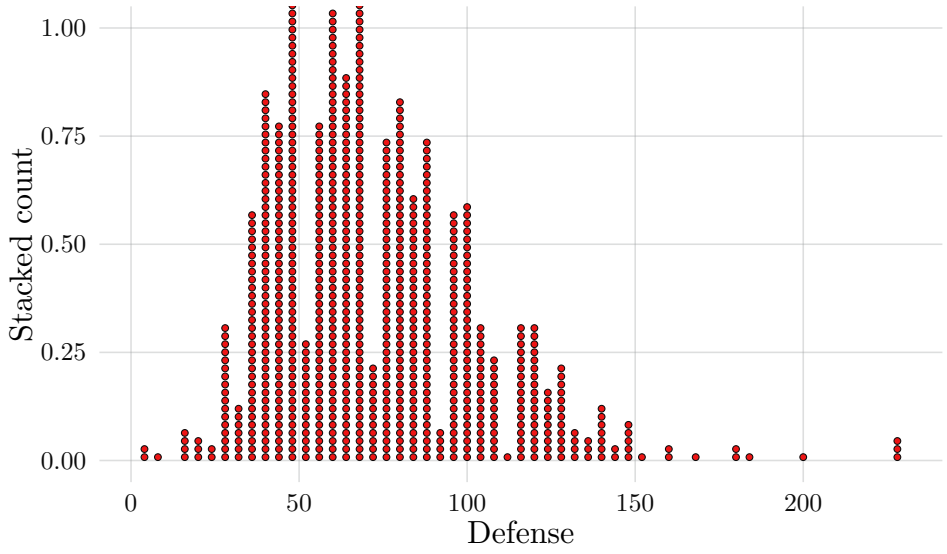
Center: Because the distribution is bimodal, a single "typical" value is less representative. By area, the *overall median* appears in the mid– to high–400s (roughly where the two humps balance, near $\sim$ 470–490 points).

Spread: The visible support runs from roughly $\sim$ 180 up to $\sim$ 780 $\Rightarrow$ **range** $\approx$ 600 points. The middle half of the data appears to lie roughly from $\sim$ 360 to $\sim$ 560 points $\Rightarrow$ IQR $\approx$ 200 points (visual estimate).

Conclusion: Total scores form two clear ability tiers—one around the low 300s and another around the low 500s—with a modest right tail containing a few very 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 discrete scoring (integers) creates visible vertical stacks at common Defense values.

Outliers: A handful of Pokémon have exceptionally high Defense values, above 150 points, with the maximum near 230. These are isolated relative to the main cluster.

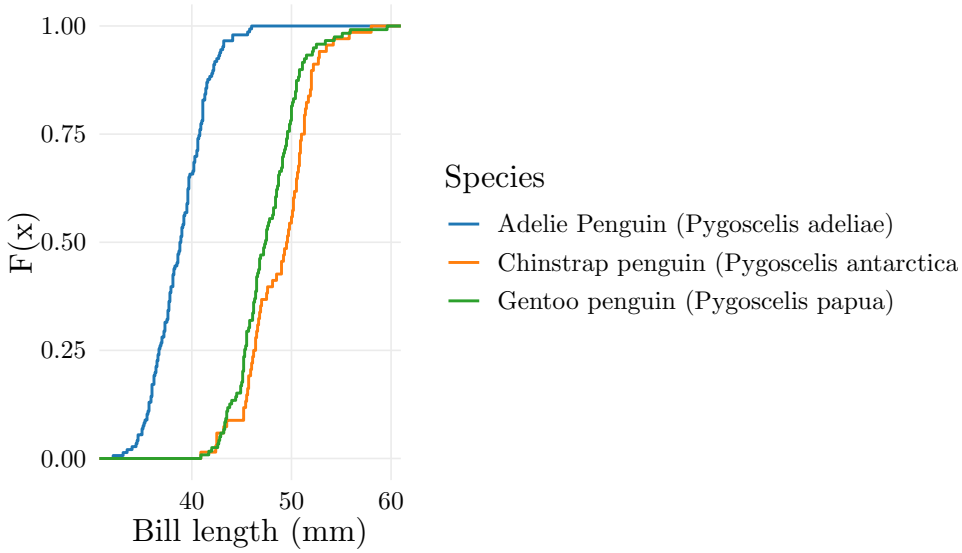
Center: The median appears to lie around 65 points, roughly the middle of the dense cluster.

Spread: The data span from a minimum near 5 up to about 230 points, giving a **range** of roughly 225 points. Most observations fall in the interval 40–100, showing that while the range is wide, the bulk of Pokémon are concentrated in a much narrower band.

Conclusion: Most Pokémon have moderate Defense (50–90 points), but the distribution is right-skewed with a small number of very high-Defense Pokémon standing out as outliers.

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 higher density; flatter stretches indicate gaps/tails.)

Shape: The *Adélie* (blue) and *Gentoo* (green) curves have very similar profiles—nearly parallel over their steep middle portions—suggesting similar underlying shapes (roughly unimodal and fairly symmetric), with Gentoo essentially a *right-shifted* version of Adélie. The *Chinstrap* (orange) curve shows two noticeably steep bands separated by a gentler section, indicating *irregularity/possible bimodality* (peaks with a small gap between clusters).

Outliers: ECDFs do not display individual outliers. Short, flat tail segments at the extremes (very small mass far left/right) suggest only a few very short or very long bills for each species.

Center: From $F(x) = 0.5$, the medians follow

$$\tilde{x}_{\text{Chinstrap}} \gtrsim \tilde{x}_{\text{Gentoo}} \gg \tilde{x}_{\text{Adélie}},$$

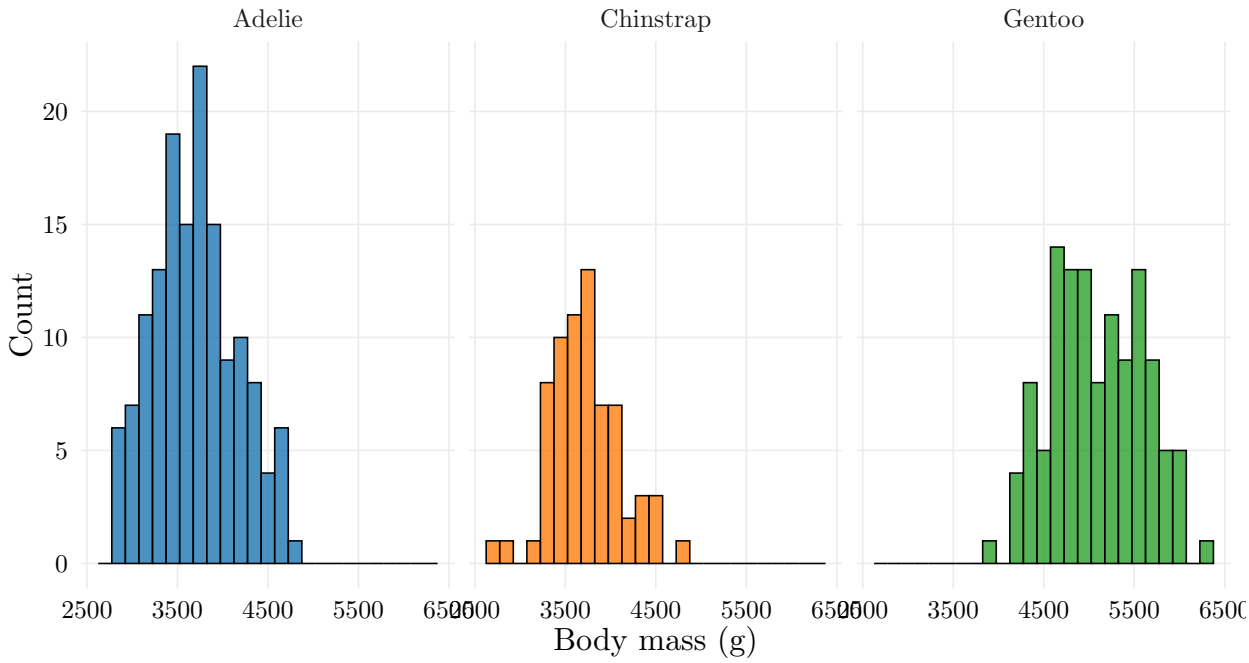
with approximate values: Adélie ≈ 39 mm; Gentoo ≈ 47–48 mm; Chinstrap ≈ 49–50 mm.

Spread: IQR corresponds to the horizontal width between $F = 0.25$ and $F = 0.75$. The IQRs are all *similar* (by eye, about 4–6 mm for each). Using visible support for a single-number range: Adélie $\approx 46 - 33 = \mathbf{13}$ mm, Gentoo $\approx 60 - 41 = \mathbf{19}$ mm, Chinstrap $\approx 55 - 43 = \mathbf{12}$ mm. Thus Gentoo shows the *widest range*, while IQRs are comparable across species.

Conclusion: Adélie and Gentoo appear to share a similar (roughly symmetric, unimodal) *shape*, with Gentoo shifted to larger bill lengths. Chinstrap bills are centered slightly higher than Gentoo and show *irregularities* consistent with multiple peaks. Overall, medians order as Chinstrap > Gentoo $\gg$ Adélie, with comparable IQRs and Gentoo having the largest range (mm).

9. Body Mass Histograms by Species (Facets)

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



Solution (SOCS): Context: The plots show the distribution of *penguin body mass* for Adélie, Chinstrap, and Gentoo species, measured in **grams**.

Shape: Adélie looks roughly *bell-shaped* and fairly symmetric about its center. Chinstrap is unimodal with a *slight right skew*. Gentoo is also roughly bell-shaped with a mild right tail.

Outliers / unusual features: Each species has some unusually heavy individuals at the far right (Adélie $\gtrsim$ 4600 g, Chinstrap $\gtrsim$ 4900 g, Gentoo $\gtrsim$ 6000 g). There are some gaps in the Chinstrap distribution (a low gap at around 3250g and a high gap at around 4600g), and the Gentoo distribution (a low gap at around 4000g and a high cap aroung 6100g) : Chinstrap shows a thinner region around the low 4000 g; Gentoo shows a dip near the mid–5200 g.

Center (medians): The *median* body mass is approximately Adélie: **3700 g**, Chinstrap: **3800 g** (very close to Adélie), Gentoo: **5100 g**. Thus Gentoo has a much larger median than the other two, which are similar.

Spread: From the visible supports: Adélie $\sim$ 2900–4700 g $\Rightarrow$ **range** $\approx$ 1800 g; Chinstrap $\sim$ 3200–5000 g $\Rightarrow$ **range** $\approx$ 1800 g; Gentoo $\sim$ 4100–6300 g $\Rightarrow$ **range** $\approx$ 2200 g. Gentoo varies most; Adélie and Chinstrap have smaller, similar spreads.

Conclusion: Gentoo penguins are the *heaviest and most variable*. Adélie and Chinstrap are lighter with *similar medians and ranges*. Adélie and Gentoo look roughly bell-shaped, while Chinstrap is slightly right-skewed, and both Chinstrap and Gentoo show minor gaps.

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.