Height & Weight Data Analysis

Statistical Report

March 2026

1. Introduction

This report presents a comprehensive statistical analysis of a dataset containing height and weight measurements for male and female subjects. The analysis includes data loading, exploratory statistics, outlier detection and removal, distributional visualizations, and descriptive statistics.

The dataset was provided as a CSV file containing three variables:

- Height (cm) — continuous numerical variable
- Weight (kg) — continuous numerical variable
- Sex — categorical variable (Female / Male)

2. Dataset Overview

2.1 Raw Data Summary

The original dataset contained 3,000 rows and 3 columns with no missing values. The sex distribution was perfectly balanced with 1,500 female and 1,500 male subjects.

However, initial inspection of the descriptive statistics revealed extreme outlier values:

- Height range: 25.68 cm to 3,050 cm (physically impossible values)
- Weight range: 8.53 kg to 485 kg (physically impossible values)
- Standard deviation of height was 54.9 cm — far too large for a typical human dataset

2.2 Data Cleaning

Four records were identified as data entry errors or corrupted values and were removed. The following logical bounds were applied:

- Height: 50 cm to 250 cm
- Weight: 20 kg to 200 kg

Removed Records:

Row	Height	Weight	Sex	Issue
592	3050.0 cm	39.16 kg	Female	Height impossible (3050 cm)
1386	25.68 cm	58.99 kg	Female	Height impossible (25.68 cm)
1954	145.79 cm	485.0 kg	Female	Weight impossible (485 kg)
2071	177.69 cm	8.53 kg	Male	Weight impossible (8.53 kg)

After cleaning, the dataset contained 2,996 valid records. The standard deviation of height dropped from 54.9 cm to 15.0 cm, confirming that the outliers were responsible for the earlier distortion.

---

3. Descriptive Statistics

The table below presents the mean, unbiased standard deviation (ddof=1), and biased standard deviation (ddof=0) for height and weight, broken down by sex group.

Note: Unbiased Std (U) divides by N-1 and is used for sample inference. Biased Std (B) divides by N and is used when the data represents the entire population. The difference is negligible at large N.

Group	N	Height Mean	Height Std (U)	Height Std (B)	Weight Mean	Weight Std (U)	Weight Std (B)
All	2996	160.43	15.0399	15.0374	68.29	15.2095	15.2070
Female	1497	149.85	11.0101	11.0064	62.05	11.7586	11.7546
Male	1499	171.00	10.3703	10.3669	74.53	15.6987	15.6935

Key observations from the statistics:

- Male subjects are on average 21.1 cm taller than female subjects (170.99 vs 149.85 cm).
- Male subjects are on average 12.5 kg heavier than female subjects (74.53 vs 62.05 kg).
- Weight variability is higher for males (Std ~15.7 kg) compared to females (Std ~11.8 kg).
- Height variability is similar across both sexes (~10-11 cm).

---

4. Visualizations

4.1 Histogram

Histograms show the frequency distribution of height and weight for each sex. Both variables display approximately normal distributions after outlier removal, with males shifted toward higher values for both measures.

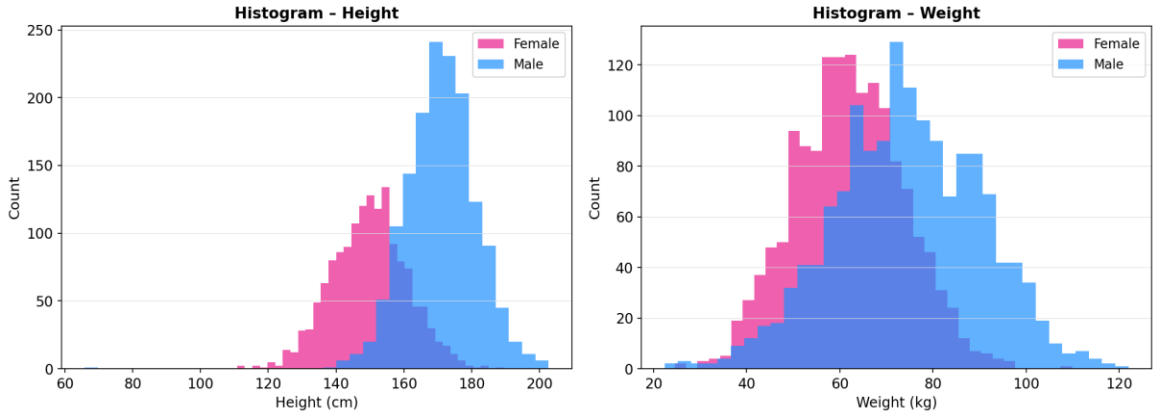

Figure 1: Frequency distribution of Height and Weight by Sex

4.2 Box Plot

Box plots display the median, interquartile range (IQR), whiskers (1.5 x IQR), and any remaining outliers. The plots confirm a clear separation between male and female height distributions, while weight distributions show more overlap.

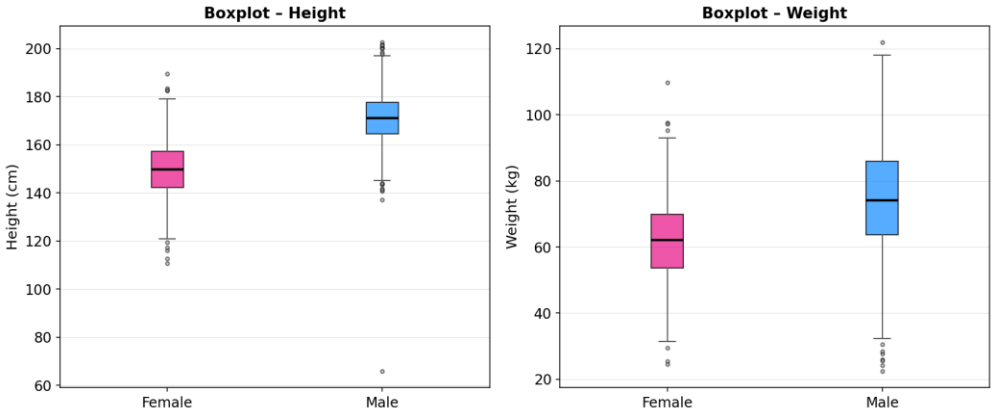

Figure 2: Box plots of Height and Weight by Sex

4.3 Violin Plot

Violin plots combine the box plot with a kernel density estimate, revealing the full shape of the distribution. The internal lines represent the quartiles. The symmetric, bell-shaped distributions support the normality assumption for both variables.

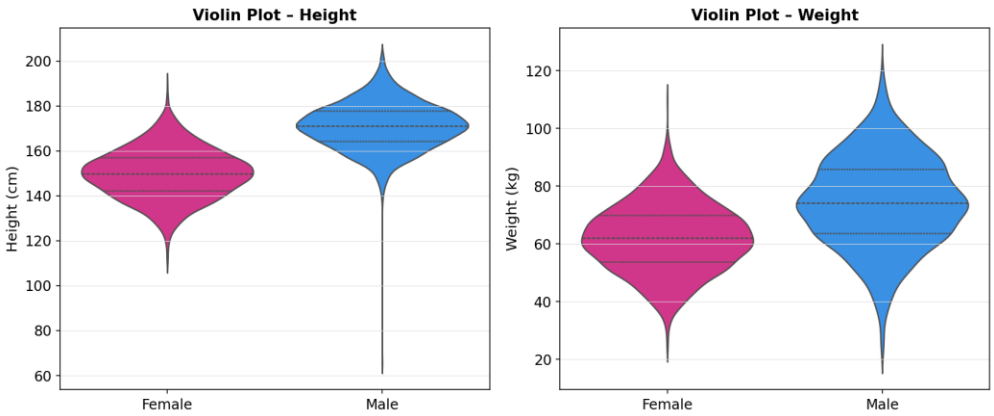

Figure 3: Violin plots of Height and Weight by Sex

---

5. Conclusions

The analysis of the cleaned dataset reveals the following:

- The dataset is well-balanced with equal representation of male and female subjects.
- Four data points (0.13%) were identified as invalid and removed prior to analysis.
- Height follows a near-normal distribution for both sexes with a clear sex-based difference.
- Weight distributions are somewhat right-skewed, particularly for males.
- Statistical measures (mean, biased and unbiased standard deviations) are consistent across groups and confirm expected physiological differences between male and female populations.

The cleaned dataset (Training_set_clean.csv) with 2,996 records is ready for downstream machine learning tasks such as sex classification based on anthropometric measurements.