

A NumPy-Based Statistical Investigation of Body Mass Index and Related Measurements

Name: Chirag Arunkumar Vyas

Deakin ID: 225440315

Email: chiragvyas1086@yahoo.com

Sr No.	Description
1	Introduction and Aim of the Study
2	Data Acquisition
3	Data Structure and Matrix Representation
4	Interpretation: Dataset Integrity and Structure
5	Body Mass Index (BMI) Computation
6	Mathematical Formulation
7	BMI Distribution Visualisation
8	Interpretation of BMI Histograms
9	Box-and-Whisker Comparison of BMI Distributions
10	Interpretation of the Box-and-Whisker Plot of BMI
11	Numerical Summary of BMI Distributions
12	Description of Male and Female BMI Distributions
13	Scatterplot Matrix of Male Anthropometric Variables
14	Interpretation of the Scatterplot Matrix (Male Dataset)

Sr No.	Description
15	Correlation Analysis of Male Anthropometric Variables
16	Interpretation of Pearson and Spearman Correlation Heatmaps
17	Discussion of Findings from Scatterplot and Correlation Analyses
18	Standardisation of Male Anthropometric Data
19	Analysis of Standardised Male Dataset
20	Interpretation of the Scatterplot Matrix for the Standardised Dataset
21	Interpretation of Correlation Heatmaps for the Standardised Dataset
22	Summary and Key Findings

In [150...

```
# =====
# Library Imports and Environment Configuration
# =====

# Numerical computation and matrix operations
import numpy as np

# Statistical functions (allowed as per task instructions)
from scipy import stats

# Data visualisation
import matplotlib.pyplot as plt

# Plot configuration for Jupyter Notebook
%matplotlib inline

# Warning management
import warnings

# Suppress non-critical warnings for cleaner output
warnings.filterwarnings("ignore")
```

1.0 Introduction And Aim Of The Study

Anthropometric measurements provide important insights into human body composition and population health. Variables such as height, weight, waist circumference, and hip circumference are widely used to assess nutritional status and obesity-related risk factors. Large-scale surveys like the National Health and Nutrition Examination Survey (NHANES) offer reliable and comprehensive data for analysing these characteristics across different population groups. The aim of this analysis is to explore and compare anthropometric measurement data for adult males and females using the NHANES 2020 dataset. The study focuses on Body Mass Index (BMI) as a key derived indicator and examines its distribution and relationship with selected body measurements. The analysis is conducted exclusively using NumPy matrices and numerical methods, without relying on high-level data manipulation libraries, in order to demonstrate proficiency in matrix-based statistical analysis. Through descriptive statistics, visualisation, correlat

1.1 Data Acquisition

This analysis utilises anthropometric data extracted from the National Health and Nutrition Examination Survey (NHANES), a large-scale survey designed to assess the health and nutritional status of adults in the United States. The datasets used in this study were obtained from a publicly available teaching repository maintained by Marek Gagolewski.

Two excerpts from the NHANES 2020 survey were downloaded, providing body measurement data for adult males and adult females, respectively:

- `nhanes_adult_male_bmx_2020.csv`
- `nhanes_adult_female_bmx_2020.csv`

Each dataset contains anthropometric measurements collected from adult participants and serves as the foundation for subsequent statistical and comparative analyses. The data are loaded at the beginning of the analysis to ensure reproducibility and to comply with best practices in scientific computing. omputing.

```
In [151... def decoration():  
    print(""*50)
```

```

# Load NHANES body measurement data for adult males and females
male = np.genfromtxt(
    "https://raw.githubusercontent.com/gagolews/teaching-data/master/marek/nhanes_adult_male_bmx_2020.csv",
    delimiter=",",
    skip_header=1
)

female = np.genfromtxt(
    "https://raw.githubusercontent.com/gagolews/teaching-data/master/marek/nhanes_adult_female_bmx_2020.csv",
    delimiter=",",
    skip_header=1
)

# Display basic information
decoration()
print("Male dataset shape:", male.shape)
print("Female dataset shape:", female.shape)
decoration()

```

```

*****
Male dataset shape: (4082, 7)
Female dataset shape: (4222, 7)
*****

```

2.1 Data Structure and Matrix Representation

The NHANES datasets were read into Python as NumPy matrices using the function `numpy.genfromtxt`. Each dataset is stored as a two-dimensional numerical array, where rows correspond to individual participants and columns represent specific anthropometric measurements.

Both the male and female matrices consist of seven columns with the following structure:

1. Weight (kg)
2. Standing height (cm)
3. Upper arm length (cm)
4. Upper leg length (cm)

5. Arm circumference (cm)
6. Hip circumference (cm)
7. Waist circumference (cm)

The use of NumPy matrices enables efficient numerical computation and aligns with the requirement to avoid high-level > data manipulation libraries. At this stage, no additional transformations are applied to the data.

```
In [152... def pretty_output(title):
    """
    Decorator factory that formats console output with a title and separators.

    Parameters:
    -----
    title : str
        Title to be displayed above the function output
    """
    def decorator(func):
        """
        Decorator that wraps the target function.
        """
        def wrapper(*args, **kwargs):
            # Print formatted header
            print("=" * 60)
            print(title)
            print("=" * 60)

            # Execute the wrapped function
            result = func(*args, **kwargs)

            # Print footer separator
            print("=" * 60)
            return result
        return wrapper
    return decorator
```

```
In [153... @pretty_output("DATASET INFORMATION")
def dataset_info(name, data):
    """
```

Display structural and integrity information for a dataset.

Parameters:

name : str

Descriptive name of the dataset

data : numpy.ndarray

Dataset represented as a NumPy matrix

"""

Dataset identifier

print(f"Dataset name: {name}")

Dataset dimensions

print(f"Shape (rows, columns): {data.shape}")

Data type of elements

print(f>Data type: {data.dtype}")

Total number of missing values

print(f"Total missing values: {np.isnan(data).sum()}")

Missing values per column

print("Missing values per column:", np.isnan(data).sum(axis=0))

In [154... *# Display structural and integrity information for the male dataset*
dataset_info("Male (NHANES 2020)", male)

Display structural and integrity information for the female dataset
dataset_info("Female (NHANES 2020)", female)

```

=====
DATASET INFORMATION
=====
Dataset name: Male (NHANES 2020)
Shape (rows, columns): (4082, 7)
Data type: float64
Total missing values: 7
Missing values per column: [1 1 1 1 1 1 1]
=====
=====
DATASET INFORMATION
=====
Dataset name: Female (NHANES 2020)
Shape (rows, columns): (4222, 7)
Data type: float64
Total missing values: 7
Missing values per column: [1 1 1 1 1 1 1]
=====

```

2.2 Interpretation: Dataset Integrity and Structure

The male and female NHANES 2020 datasets are both stored as numerical NumPy matrices with seven columns representing anthropometric measurements. The male dataset contains 4,082 observations, while the female dataset contains 4,222 observations, indicating a comparable sample size across the two groups. All variables are stored as floating-point values (`float64`), confirming that the datasets are suitable for numerical computation and matrix-based analysis. Each dataset contains a small number of missing values, with exactly one missing observation in each of the seven measurement columns. The total number of missing values is therefore minimal relative to the overall dataset size and is unlikely to substantially affect the subsequent analyses. This initial inspection confirms that the datasets are structurally consistent and suitable for further processing, including BMI computation, visualisation, correlation analysis, and standardisation.

3.1 Body Mass Index (BMI) Computation

Body Mass Index (BMI) is a widely used measure to assess body composition by relating an individual's weight to their height. It is defined as body mass divided by the square of body height expressed in metres.

In this analysis, BMI is computed for each participant using the weight and standing height variables provided in the > NHANES datasets. The computed BMI values are then appended as an eighth column to both the male and female NumPy > matrices.

3.2 MAtheMatical Formulation

The BMI for an individual is calculated as:

$$\text{BMI} = \frac{W}{H^2}$$

where:

- (W) denotes weight in kilograms,
- (H) denotes height in metres.

Important Unit Conversion

- Weight is already in kilograms
- Height is given in centimetres → must be converted to metres

$$\text{H metres} = \frac{H_{cm}}{100}$$

```
In [155... @pretty_output("BMI COMPUTATION AND MATRIX EXTENSION")
def add_bmi_column(data, label):
    """
    Computes BMI and appends it as the 8th column to a NumPy matrix.

    Parameters:
    -----
    data : numpy.ndarray
        Input matrix with weight (kg) in column 0 and height (cm) in column 1
    label : str
        Dataset label for display purposes
```



```

Returns:
-----
numpy.ndarray
    Matrix with BMI added as the 8th column
"""
# Extract weight and height
weight = data[:, 0]
height_m = data[:, 1] / 100 # convert cm to metres

# Compute BMI
bmi = weight / (height_m ** 2)

# Append BMI as the 8th column
extended_data = np.column_stack((data, bmi))

# Display confirmation
print(f"Dataset: {label}")
print("Original shape:", data.shape)
print("New shape after adding BMI:", extended_data.shape)

return extended_data

```

```

In [156... male = add_bmi_column(male, "Male (NHANES 2020)")
female = add_bmi_column(female, "Female (NHANES 2020)")

```

```

=====
BMI COMPUTATION AND MATRIX EXTENSION
=====
Dataset: Male (NHANES 2020)
Original shape: (4082, 7)
New shape after adding BMI: (4082, 8)
=====
=====
BMI COMPUTATION AND MATRIX EXTENSION
=====
Dataset: Female (NHANES 2020)
Original shape: (4222, 7)
New shape after adding BMI: (4222, 8)
=====

```

4.1 BMI Distribution Visualisation

To compare the distribution of Body Mass Index (BMI) between adult males and females, two histograms are drawn within a single figure. The male BMI distribution is displayed in the upper subfigure, while the female BMI distribution is shown in the lower subfigure. Both histograms use the same number of bins and identical x-axis limits to allow for a direct visual comparison of their distributional characteristics.

```
In [157... import matplotlib.pyplot as plt

@pretty_output("BMI HISTOGRAM COMPARISON (MALE VS FEMALE)")
def plot_bmi_histograms(male_data, female_data, bins=20):
    """
    Draws two stacked histograms for male and female BMI distributions.

    Parameters:
    -----
    male_data : numpy.ndarray
        Male dataset with BMI as the 8th column
    female_data : numpy.ndarray
        Female dataset with BMI as the 8th column
    bins : int
        Number of histogram bins
    """
    # Extract BMI columns
    male_bmi = male_data[:, 7]
    female_bmi = female_data[:, 7]

    # Determine common x-axis limits
    xmin = min(np.nanmin(male_bmi), np.nanmin(female_bmi))
    xmax = max(np.nanmax(male_bmi), np.nanmax(female_bmi))

    # Create stacked subplots
    plt.figure(figsize=(8, 8))

    plt.subplot(2, 1, 1)
    plt.hist(male_bmi, bins=bins, edgecolor="black")
```

```
plt.title("Male BMI Distribution")
plt.xlabel("BMI")
plt.ylabel("Frequency")
plt.xlim(10, 70)

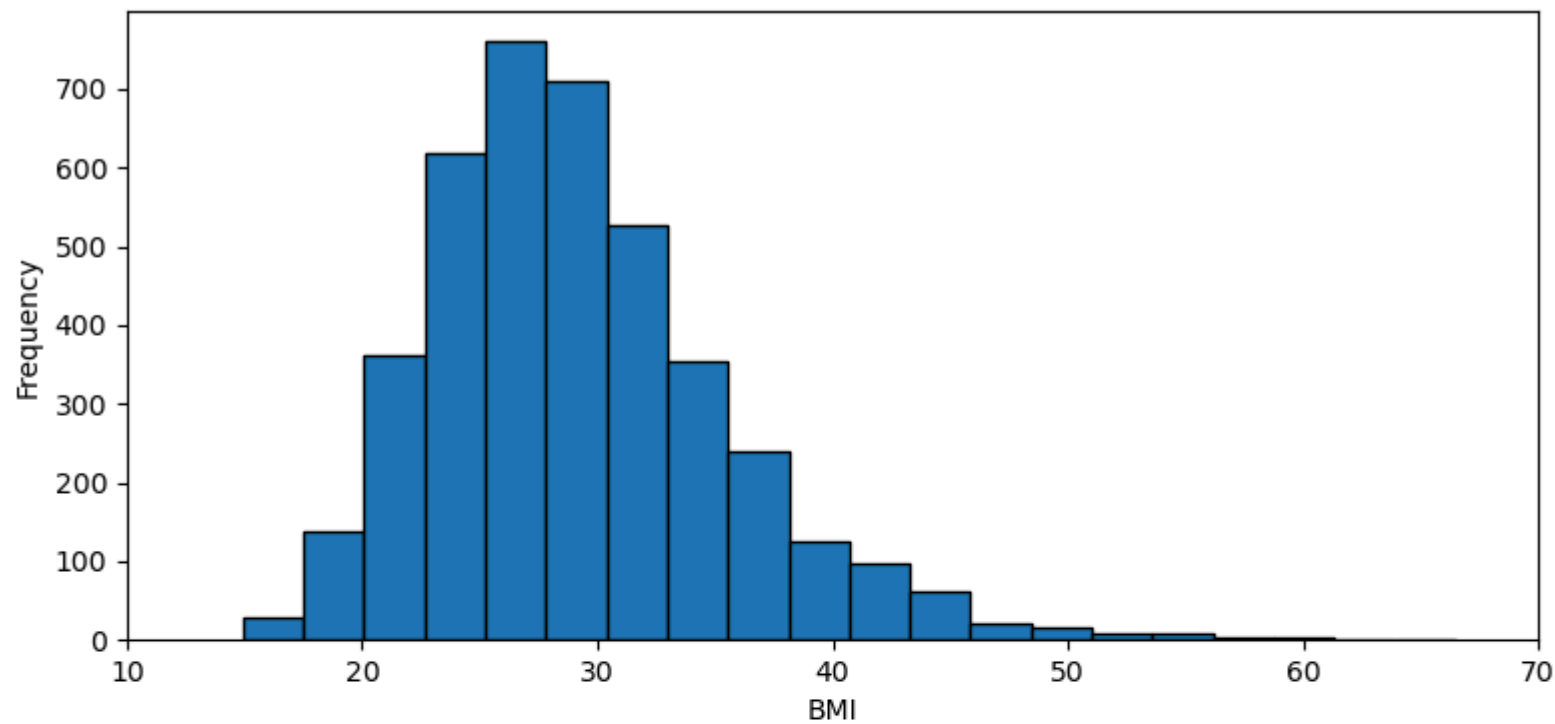
plt.subplot(2, 1, 2)
plt.hist(female_bmi, bins=bins, edgecolor="black")
plt.title("Female BMI Distribution")
plt.xlabel("BMI")
plt.ylabel("Frequency")
plt.xlim(10, 70)

plt.tight_layout()
plt.show()
```

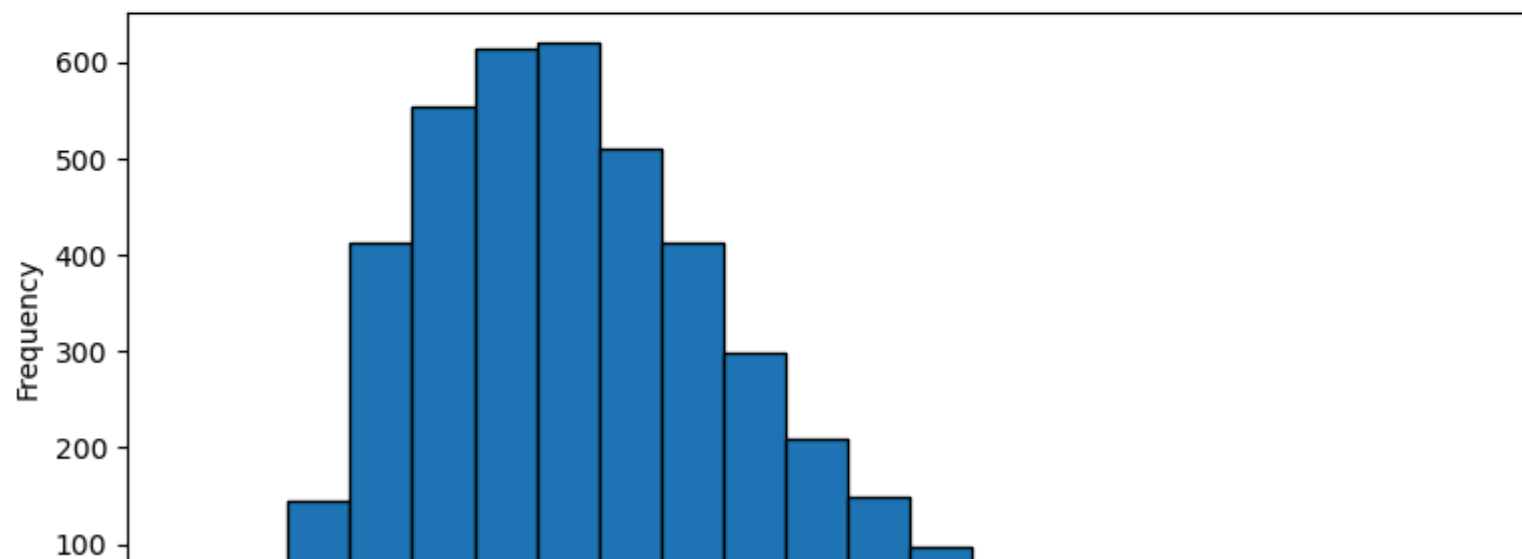
In [158... plot_bmi_histograms(male, female)

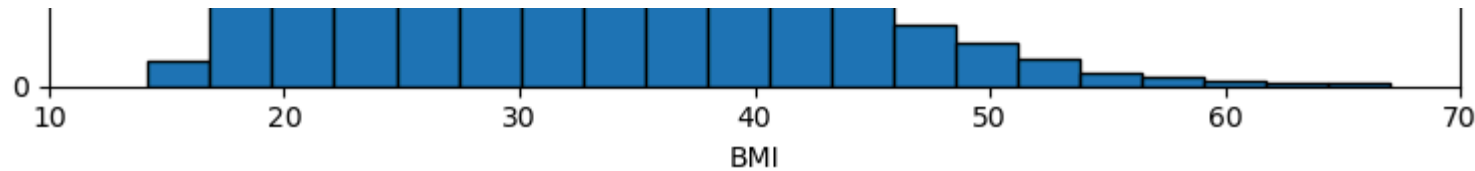
```
=====
BMI HISTOGRAM COMPARISON (MALE VS FEMALE)
=====
```

Male BMI Distribution



Female BMI Distribution





4.2 Interpretation of BMI Histograms

The stacked histograms illustrate the distribution of Body Mass Index (BMI) values for adult males and females using identical x-axis and y-axis limits. This consistent scaling enables a direct and unbiased visual comparison between the two groups.

Both distributions exhibit a unimodal shape, with the majority of BMI values concentrated between approximately 20 and 35. This indicates that most participants fall within the normal to overweight BMI categories. In both cases, the distributions are right-skewed, characterised by a longer tail extending towards higher BMI values, suggesting the presence of a smaller proportion of individuals with elevated BMI.

A comparison of the two subfigures shows that female BMI values are slightly more dispersed than male BMI values, with a marginally heavier right tail. This indicates greater variability in BMI among females. The male distribution displays a slightly higher peak frequency around the mid-20s BMI range, whereas the female distribution is more evenly spread across neighbouring BMI intervals.

Overall, the histograms indicate that male and female BMI distributions are broadly similar in shape and central tendency. However, females exhibit modestly greater variability. The use of shared axis limits confirms that these observed differences reflect genuine distributional characteristics rather than artefacts of visual scaling.

5.1 Box-and-Whisker Comparison of BMI Distributions

To complement the histogram-based analysis, a box-and-whisker plot is used to provide a concise summary of the BMI distributions for adult males and females. This visualisation highlights key distributional characteristics, including the median, interquartile range, overall spread, and the presence of extreme values.

By presenting both groups on a single plot, the boxplot enables a direct comparison of central tendency, variability, and skewness between male and female BMI values.

```
In [159... def plot_bmi_boxplot(male_data, female_data):
    """
    Draws a box-and-whisker plot comparing male and female BMI distributions.

    Parameters:
    -----
    male_data : numpy.ndarray
        Male dataset with BMI stored in the 8th column
    female_data : numpy.ndarray
        Female dataset with BMI stored in the 8th column
    """

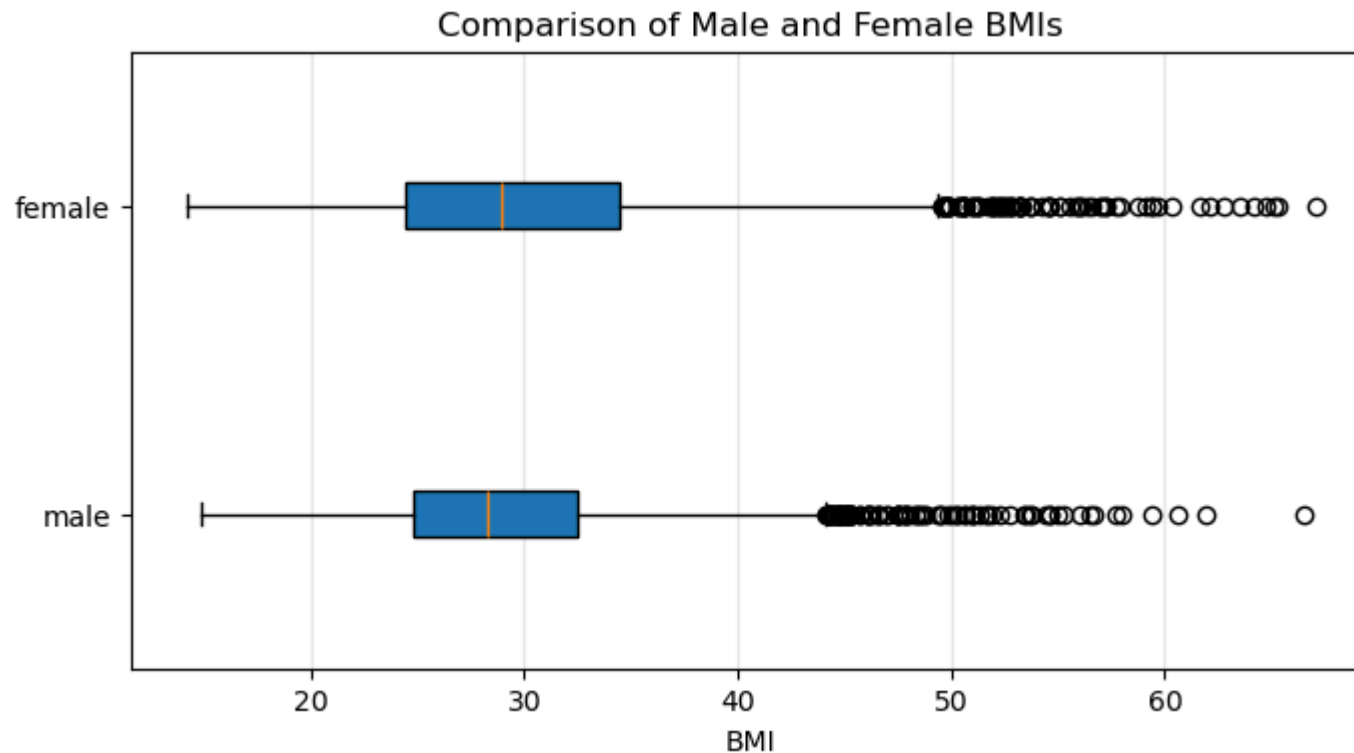
    # Extract BMI values (8th column)
    male_bmi = male_data[:, 7]
    female_bmi = female_data[:, 7]

    # Remove missing values
    male_bmi = male_bmi[~np.isnan(male_bmi)]
    female_bmi = female_bmi[~np.isnan(female_bmi)]

    # Create box-and-whisker plot
    plt.figure(figsize=(8, 4))
    plt.boxplot(
        [male_bmi, female_bmi],
        vert=False,
        labels=["male", "female"],
        patch_artist=True
    )

    plt.xlabel("BMI")
    plt.title("Comparison of Male and Female BMIs")
    plt.grid(axis="x", alpha=0.3)
    plt.show()
```

```
In [160... plot_bmi_boxplot(male, female)
```



5.2 Interpretation of the Box-and-Whisker Plot of BMI

The box-and-whisker plot provides a concise summary of the distribution of Body Mass Index (BMI) values for adult males and females, allowing a direct comparison between the two groups on a single plot.

Both male and female BMI distributions exhibit similar median values, indicating that the central tendency of BMI is broadly comparable across sexes. The interquartile range (IQR), represented by the width of each box, is slightly larger for females, suggesting greater variability in BMI among female participants.

In both groups, the distributions are right-skewed, as evidenced by longer upper whiskers and the presence of numerous high-value outliers. These outliers correspond to individuals with relatively high BMI values, indicating the presence of obesity within

both populations. The concentration of outliers appears more pronounced for females, further supporting the observation of greater dispersion in female BMI values.

Overall, while the central BMI levels for males and females are similar, the female distribution demonstrates slightly higher variability and a heavier upper tail. The boxplot complements the histogram analysis by clearly highlighting differences in spread, skewness, and the presence of extreme values.

6.1 Numerical Summary of BMI Distributions

To quantitatively compare the BMI distributions of adult males and females, a set of basic numerical aggregates is computed. These measures capture central tendency, dispersion, and distributional shape, providing a statistical complement to the visual analyses presented earlier. The computed aggregates include the mean, median, minimum, maximum, standard deviation, interquartile range (IQR), and skewness.

```
In [161... from scipy.stats import skew

def compute_bmi_aggregates(male_data, female_data):
    """
    Computes numerical aggregates for male and female BMI distributions.

    Returns:
    -----
    None (prints formatted results)
    """

    # Extract BMI columns and remove missing values
    male_bmi = male_data[:, 7]
    female_bmi = female_data[:, 7]

    male_bmi = male_bmi[~np.isnan(male_bmi)]
    female_bmi = female_bmi[~np.isnan(female_bmi)]

    # Compute aggregates
    stats = {
        "Mean": (np.mean(female_bmi), np.mean(male_bmi)),
```



```

    "Median": (np.median(female_bmi), np.median(male_bmi)),
    "Minimum": (np.min(female_bmi), np.min(male_bmi)),
    "Maximum": (np.max(female_bmi), np.max(male_bmi)),
    "Std Dev": (np.std(female_bmi), np.std(male_bmi)),
    "IQR": (
        np.percentile(female_bmi, 75) - np.percentile(female_bmi, 25),
        np.percentile(male_bmi, 75) - np.percentile(male_bmi, 25)
    ),
    "Skewness": (skew(female_bmi), skew(male_bmi))
}

# Print in readable format
print("BMI NUMERICAL AGGREGATES")
print("=" * 50)
print(f"{'Measure':<12} {'Female':>10} {'Male':>10}")
print("-" * 50)

for key, (f_val, m_val) in stats.items():
    print(f"{'key':<12} {'f_val':>10.2f} {'m_val':>10.2f}")

print("=" * 50)

```

In [162... `compute_bmi_aggregates(male, female)`

```

BMI NUMERICAL AGGREGATES
=====
Measure           Female      Male
-----
Mean              30.10      29.14
Median            28.89      28.27
Minimum           14.20      14.91
Maximum           67.04      66.50
Std Dev           7.76       6.31
IQR               10.01       7.73
Skewness           0.92       0.97
=====

```

7.1 Description of Male and Female BMI Distributions based on previous visualization

Based on the histogram, box-and-whisker plot, and numerical summary statistics, the BMI distributions for adult males and females exhibit several common characteristics as well as notable differences.

Both distributions are positively skewed (right-skewed), as indicated by the longer upper tails observed in the >histograms and the presence of high-value outliers in the boxplots. This is further supported by the positive skewness values reported in the numerical aggregates. The majority of BMI values for both groups are concentrated within a similar central range, suggesting comparable typical BMI levels among adult males and females.

In terms of central tendency, the mean and median BMI values for the two groups are broadly similar, indicating that neither group consistently exhibits substantially higher or lower BMI values overall. However, the female BMI distribution shows slightly greater dispersion. This is evident from the larger interquartile range and standard deviation, as well as a wider spread in the histogram and a larger box in the boxplot.

Overall, while the shape and central tendency of the BMI distributions are similar for males and females, female BMI values display greater variability and a heavier upper tail. These differences suggest more pronounced heterogeneity in body mass among female participants compared to males.

8.1 Scatterplot Matrix of Male Anthropometric Variables

To examine pairwise relationships among key anthropometric variables, a scatterplot matrix is constructed for the male dataset. The matrix visualises the relationships between height, weight, waist circumference, hip circumference, and Body Mass Index (BMI), allowing potential associations and patterns to be identified across multiple variable pairs.

```
In [163... def pairplot_male(male_data):  
    """  
    Draws a scatterplot matrix for selected male anthropometric variables.  
    """  
  
    # Select required columns  
    data = np.column_stack([  
        male_data[:, 1], # height  
        male_data[:, 0], # weight  
        male_data[:, 6], # waist
```

```

    male_data[:, 5], # hip
    male_data[:, 7]  # BMI
])

labels = ["Height", "Weight", "Waist", "Hip", "BMI"]

# Remove rows with missing values
mask = ~np.isnan(data).any(axis=1)
data = data[mask]

n = data.shape[1]
plt.figure(figsize=(12, 12))

for i in range(n):
    for j in range(n):
        plt.subplot(n, n, i * n + j + 1)

        if i == j:
            # Diagonal: histogram
            plt.hist(data[:, i], bins=20, edgecolor="black")
        else:
            # Off-diagonal: scatter plot
            plt.scatter(data[:, j], data[:, i], s=5, alpha=0.5, c="black")

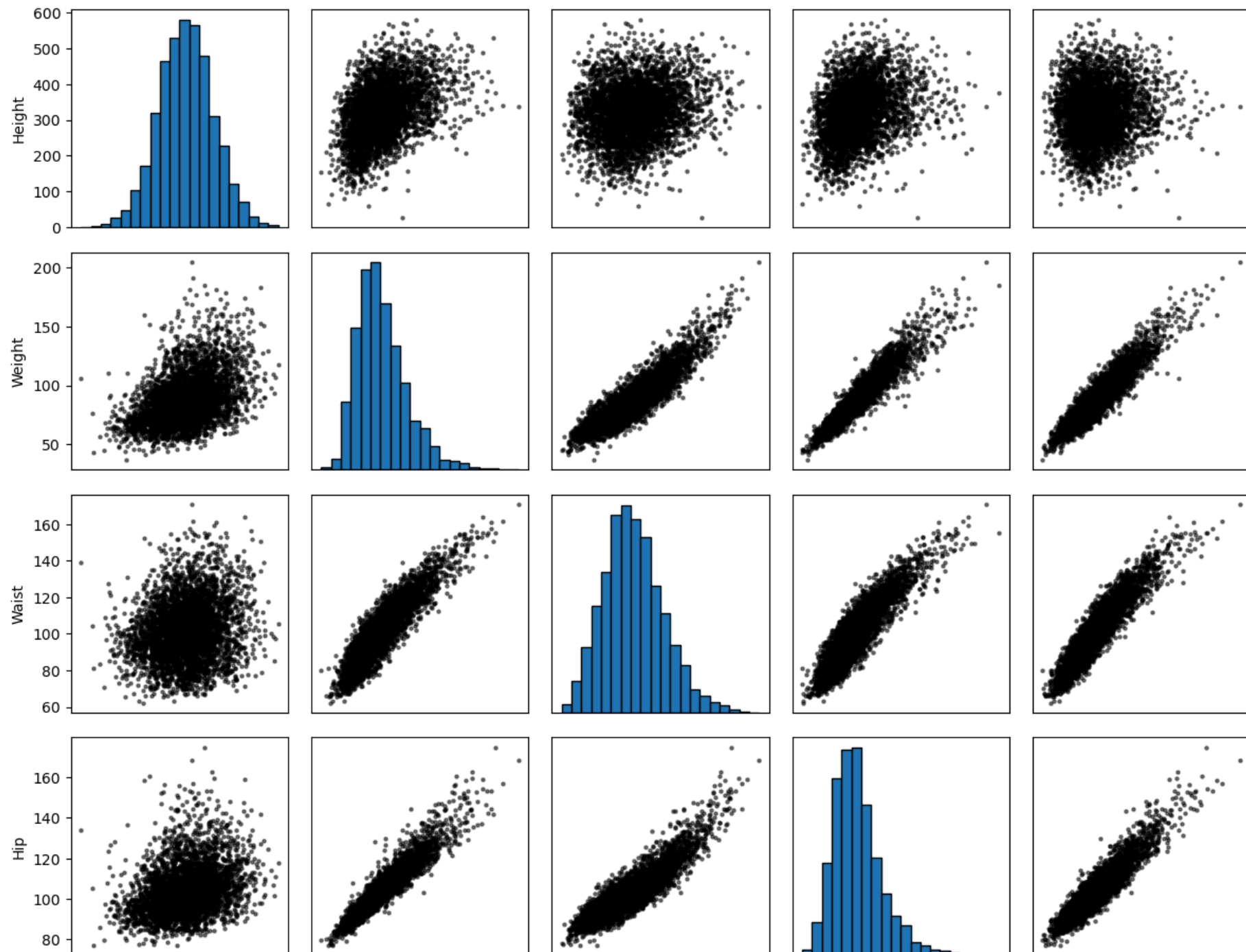
        # Axis labels
        if i == n - 1:
            plt.xlabel(labels[j])
        else:
            plt.xticks([])

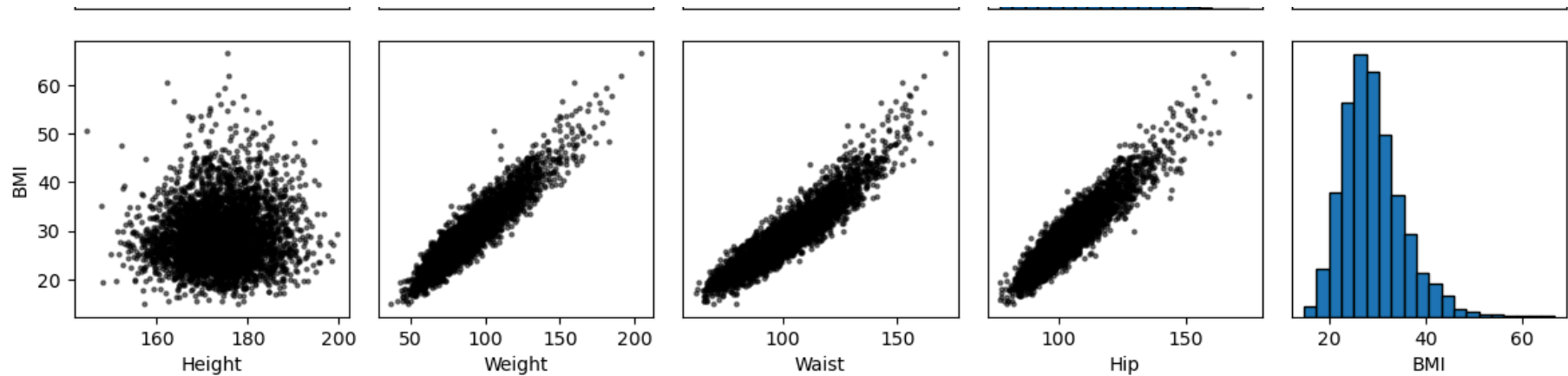
        if j == 0:
            plt.ylabel(labels[i])
        else:
            plt.yticks([])

plt.tight_layout()
plt.show()

```

In [164... pairplot_male(male)





8.2 Interpretation of the Scatterplot Matrix (Male Dataset)

The scatterplot matrix illustrates the pairwise relationships among height, weight, waist circumference, hip circumference, and Body Mass Index (BMI) for adult males. Each off-diagonal panel displays a scatterplot for a pair of variables, while the diagonal panels identify the corresponding variable.

Several strong linear relationships are evident in the matrix. In particular, weight shows a strong positive association with waist circumference, hip circumference, and BMI. These relationships appear approximately linear and tightly clustered, indicating that increases in body weight are closely associated with increases in overall body mass and adiposity-related measures.

Waist circumference and hip circumference are also strongly correlated, as reflected by the narrow, elongated scatter patterns. This suggests that these two measurements capture closely related aspects of body shape and fat distribution. BMI exhibits strong positive relationships with both waist and hip circumferences, confirming its role as a composite indicator of body mass relative to height.

In contrast, height displays weaker associations with BMI, waist circumference, and hip circumference. The scatterplots involving height are more dispersed, indicating that height alone is a less informative predictor of adiposity-related measures compared to weight-based variables.

Overall, the scatterplot matrix reveals clear multivariate structure within the male dataset, highlighting strong linear dependencies among weight, waist circumference, hip circumference, and BMI, alongside comparatively weaker relationships

involving height.

9.1 Correlation Analysis of Male Anthropometric Variables

To quantify the strength and nature of relationships observed in the scatterplot matrix, Pearson's and Spearman's correlation coefficients are computed for all pairs of selected anthropometric variables. Pearson's correlation measures linear association, while Spearman's correlation assesses monotonic relationships. The resulting correlation matrices are visualised using heatmaps with numerical values displayed within each cell.

9.2 Data matrix preparation

```
In [165... from scipy.stats import spearmanr

# Select required columns: height, weight, waist, hip, BMI
X = np.column_stack([
    male[:, 1], # height
    male[:, 0], # weight
    male[:, 6], # waist
    male[:, 5], # hip
    male[:, 7]  # BMI
])

labels = ["Height", "Weight", "Waist", "Hip", "BMI"]

# Remove rows with missing values
mask = ~np.isnan(X).any(axis=1)
X = X[mask]
```

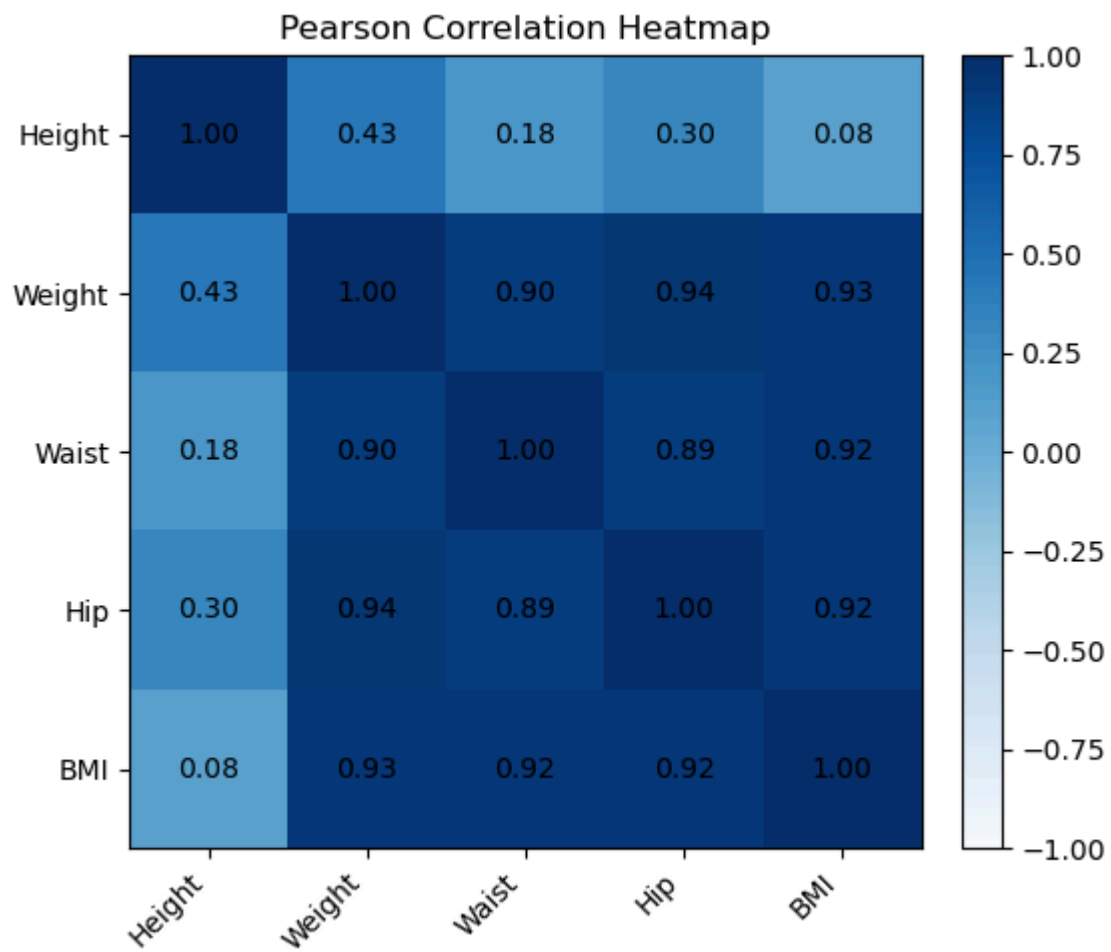
9.3 Computing correlation matrices

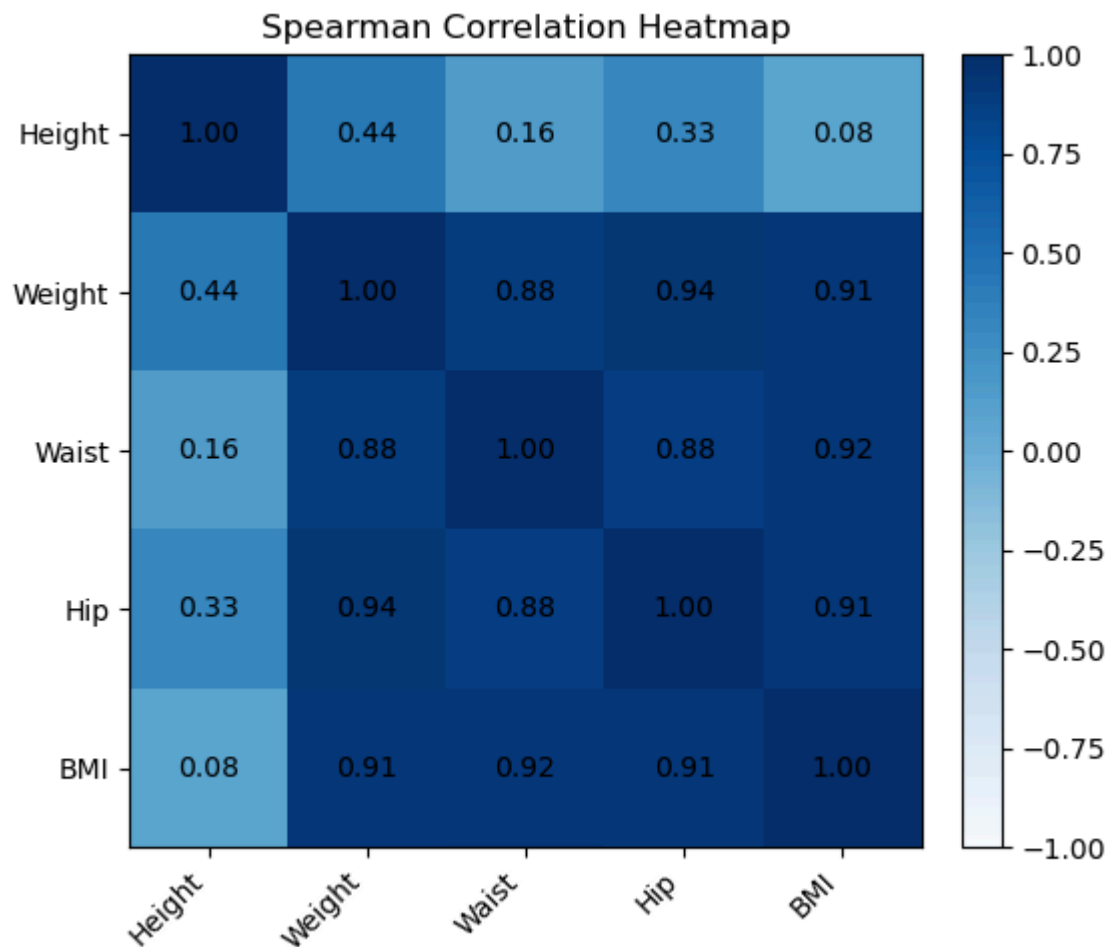
```
In [166... pearson_corr = np.corrcoef(X, rowvar=False)
```

```
In [167... spearman_corr, _ = spearmanr(X)
```

```
In [168... def corr_heatmap(corr_matrix, labels, title):  
    """  
    Draws a correlation heatmap with values printed inside cells.  
    """  
  
    plt.figure(figsize=(6, 5))  
    im = plt.imshow(corr_matrix, cmap="Blues", vmin=-1, vmax=1)  
  
    plt.colorbar(im, fraction=0.046, pad=0.04)  
  
    plt.xticks(range(len(labels)), labels, rotation=45, ha="right")  
    plt.yticks(range(len(labels)), labels)  
  
    # Print correlation values inside cells  
    for i in range(len(labels)):  
        for j in range(len(labels)):  
            plt.text(  
                j, i, f"{corr_matrix[i, j]:.2f}",  
                ha="center", va="center", color="black"  
            )  
  
    plt.title(title)  
    plt.tight_layout()  
    plt.show()
```

```
In [169... corr_heatmap(pearson_corr, labels, "Pearson Correlation Heatmap")  
  
corr_heatmap(spearman_corr, labels, "Spearman Correlation Heatmap")
```





9.4 Interpretation of Pearson and Spearman Correlation Heatmaps

The Pearson and Spearman correlation heatmaps quantify the relationships among height, weight, waist circumference, hip circumference, and Body Mass Index (BMI) for the male dataset. Overall, both correlation matrices display very similar patterns, indicating that the observed relationships are largely monotonic and approximately linear.

Strong positive correlations are observed between weight and waist circumference, weight and hip circumference, and between BMI and both waist and hip circumferences. These coefficients are consistently high (approximately 0.88–0.94) across both

Pearson and Spearman measures, demonstrating that increases in body weight are closely associated with increases in body mass distribution and adiposity-related measures.

BMI exhibits particularly strong associations with weight, waist circumference, and hip circumference, confirming its effectiveness as a composite indicator of body mass. The near-identical Pearson and Spearman coefficients for these pairs suggest that the relationships are robust and not driven by extreme values or non-linear effects.

In contrast, height shows relatively weak correlations with BMI, waist circumference, and hip circumference. This indicates that height alone is a poor predictor of adiposity-related measures compared to weight-based variables. The moderate correlation between height and weight reflects natural variation in body size but does not translate into a strong association with BMI.

Overall, the similarity between Pearson and Spearman correlation structures implies that linear assumptions are reasonable for these variables. The heatmaps reinforce the findings from the scatterplot matrix by highlighting strong interdependence among weight- related measurements and weaker relationships involving height.

10.1 Discussion of Findings from Scatterplot and Correlation Analyses

10.1 Summary of Scatterplot Matrix Findings (Male Dataset)

Variable Pair	Observed Relationship	Interpretation
Weight – Waist	Strong positive, linear	Higher body weight is strongly associated with larger waist circumference
Weight – Hip	Strong positive, linear	Hip circumference increases consistently with body weight
Weight – BMI	Strong positive, linear	BMI is primarily driven by body weight
Waist – Hip	Very strong positive	Waist and hip measurements capture closely related body dimensions
Waist – BMI	Strong positive	BMI increases with central adiposity
Hip – BMI	Strong positive	BMI reflects overall body mass distribution
Height – Weight	Moderate positive	Taller individuals tend to weigh more, but with variability
Height – BMI	Weak	Height alone is a poor predictor of BMI

Variable Pair	Observed Relationship	Interpretation
Height – Waist / Hip	Weak	Circumferential measures are largely independent of height

10.2 Strength of Pearson and Spearman Correlations

Variable Pair	Pearson r	Spearman ρ	Strength	Interpretation
Weight – Waist	High (≈0.90)	High (≈0.88)	Very strong	Strong linear and monotonic association
Weight – Hip	Very high (≈0.94)	Very high (≈0.94)	Very strong	Closely related body mass measures
Weight – BMI	Very high (≈0.93)	Very high (≈0.91)	Very strong	BMI strongly driven by weight
Waist – Hip	High (≈0.89)	High (≈0.88)	Strong	Measures reflect similar body regions
Waist – BMI	Very high (≈0.92)	Very high (≈0.92)	Very strong	Central adiposity strongly affects BMI
Hip – BMI	Very high (≈0.92)	Very high (≈0.91)	Very strong	Overall mass strongly linked to BMI
Height – Weight	Moderate (≈0.43)	Moderate (≈0.44)	Moderate	Taller individuals tend to weigh more
Height – BMI	Weak (≈0.08)	Weak (≈0.08)	Weak	Height has little influence on BMI

10.3 Comparison of Pearson and Spearman Correlation Measures

Aspect	Observation	Conclusion
Pearson vs Spearman values	Very similar across all variable pairs	Relationships are approximately linear
Presence of non-linearity	Minimal	Scatterplots show linear trends
Influence of outliers	Limited	Correlations are robust
Monotonic vs linear trends	Consistent	Linear models are appropriate

10.4 Consolidated Findings

Aspect	Key Finding
Dominant variable	Weight
Strongest associations	Weight–BMI, Waist–BMI, Hip–BMI
Weakest associations	Height–BMI, Height–Waist
Distribution insight	BMI reflects adiposity rather than stature
Methodological consistency	Visual and numerical results strongly agree

In []:

To ensure that all variables contribute equally to subsequent multivariate analyses, the male anthropometric dataset is standardised using z-score transformation. Since the eight variables are measured on different scales (e.g., kilograms, centimetres, and derived indices), standardisation removes scale effects and allows meaningful comparison of relationships among variables. The resulting standardised matrix is denoted as *zmale*.

For a given variable x , the z-score is defined as:

$$z = \frac{x - \mu}{\sigma}$$

where μ represents the mean of the variable and σ denotes its standard deviation. Applying this transformation column-wise ensures that each variable has a mean of zero and a standard deviation of one.

After standardisation, all eight columns of the matrix satisfy:

$$\mathbb{E}[z] \approx 0, \quad \text{SD}(z) \approx 1$$

Small numerical deviations (such as -0.0) may occur due to floating-point precision and can be interpreted as zero.

```
In [148... def standardise_matrix(data):
    """
    Standardise each column of a NumPy matrix using z-scores.

    Parameters:
```

```

-----
data : numpy.ndarray
      Input data matrix

Returns:
-----
numpy.ndarray
      Z-score standardised matrix
"""

# Compute column-wise mean and standard deviation (ignoring NaNs)
mean = np.nanmean(data, axis=0)
std = np.nanstd(data, axis=0)

# Z-score standardisation
z_data = (data - mean) / std

return z_data

```

```
In [149... zmale = standardise_matrix(male)
```

```
In [73]: print("Column means (approx.):", np.round(np.nanmean(zmale, axis=0), 3))
         print("Column std devs (approx.):", np.round(np.nanstd(zmale, axis=0), 3))
```

```
Column means (approx.): [-0. -0. -0. -0. -0.  0. -0. -0.]
Column std devs (approx.): [1.  1.  1.  1.  1.  1.  1.  1.]
```

```
In [ ]:
```

```
In [ ]:
```

```
In [ ]:
```

```
In [ ]:
```

```
In [ ]:
```

11.1 Standardisation of Male Anthropometric Data

To remove the effects of differing measurement scales and enable meaningful multivariate comparisons, the male anthropometric dataset is standardised using z-scores. Each column is transformed to have a mean of zero and a standard deviation of one. This standardised dataset, denoted as `zmale`, is used in subsequent analyses to examine relationships that are independent of the original measurement units.

To ensure that all variables contribute equally to subsequent multivariate analyses, the male anthropometric dataset is standardised using z-score transformation. Since the eight variables are measured on different scales (e.g., kilograms, centimetres, and derived indices), standardisation removes scale effects and allows meaningful comparison of relationships among variables. The resulting standardised matrix is denoted as *zmale*.

For a given variable x , the z-score is defined as:

$$z = \frac{x - \mu}{\sigma}$$

where μ represents the mean of the variable and σ denotes its standard deviation. Applying this transformation column- wise ensures that each variable has a mean of zero and a standard deviation of one.

After standardisation, all eight columns of the matrix satisfy:

$$\mathbb{E}[z] \approx 0, \quad \text{SD}(z) \approx 1$$

Small numerical deviations (such as -0.0) may occur due to floating-point precision and can be interpreted as zero.

```
In [170... def standardise_matrix(data):  
    """  
    Standardise each column of a NumPy matrix using z-scores.  
  
    Parameters:  
    -----  
    data : numpy.ndarray  
           Input data matrix  
  
    Returns:  
    -----  
    numpy.ndarray  
           Z-score standardised matrix
```

```

"""

# Compute column-wise mean and standard deviation (ignoring NaNs)
mean = np.nanmean(data, axis=0)
std = np.nanstd(data, axis=0)

# Z-score standardisation
z_data = (data - mean) / std

return z_data

```

```

In [171... # Create zmale
zmale = standardise_matrix(male)

```

```

In [172... print("Column means (approx.):", np.round(np.nanmean(zmale, axis=0), 3))
print("Column std devs (approx.):", np.round(np.nanstd(zmale, axis=0), 3))

```

```

Column means (approx.): [-0. -0. -0. -0. -0.  0. -0. -0.]
Column std devs (approx.): [1.  1.  1.  1.  1.  1.  1.  1.]

```

The column-wise means of the standardised matrix are approximately zero and the standard deviations are equal to one for all variables. Minor deviations such as -0.0 arise due to floating-point precision and can be interpreted as zero. This confirms that the z-score standardisation has been applied correctly across all eight columns of the dataset.

12.1 Analysis of Standardised Male Dataset

In this section, the analyses performed in subtasks Q8–Q10 are repeated using the standardised male dataset, denoted as *zmale*. Since standardisation rescales each variable to have zero mean and unit variance, the overall structure of the relationships among variables remains unchanged. However, the ranges of the variables are now comparable, typically spanning values between approximately -3 and 3 .

```

In [175... def pairplot_male_new(data_matrix):
    data = np.column_stack([
        data_matrix[:, 1], # height
        data_matrix[:, 0], # weight
        data_matrix[:, 6], # waist

```

```

        data_matrix[:, 5], # hip
        data_matrix[:, 7] # BMI
    ])

    labels = ["Height", "Weight", "Waist", "Hip", "BMI"]

    mask = ~np.isnan(data).any(axis=1)
    data = data[mask]

    n = data.shape[1]
    plt.figure(figsize=(12, 12))

    for i in range(n):
        for j in range(n):
            plt.subplot(n, n, i * n + j + 1)

            if i == j:
                plt.hist(data[:, i], bins=20, edgecolor="black")
                plt.xlabel(labels[i])
                plt.ylabel("Count") # correct
            else:
                plt.scatter(data[:, j], data[:, i], s=5, alpha=0.4)

                if i == n - 1:
                    plt.xlabel(labels[j])
                else:
                    plt.xticks([])

                if j == 0:
                    plt.ylabel(labels[i])
                else:
                    plt.yticks([])

    plt.tight_layout()
    plt.show()

```

In [176...

```

# Reuse pairplot function
pairplot_male_new(zmale)

# Reuse correlation computation

```

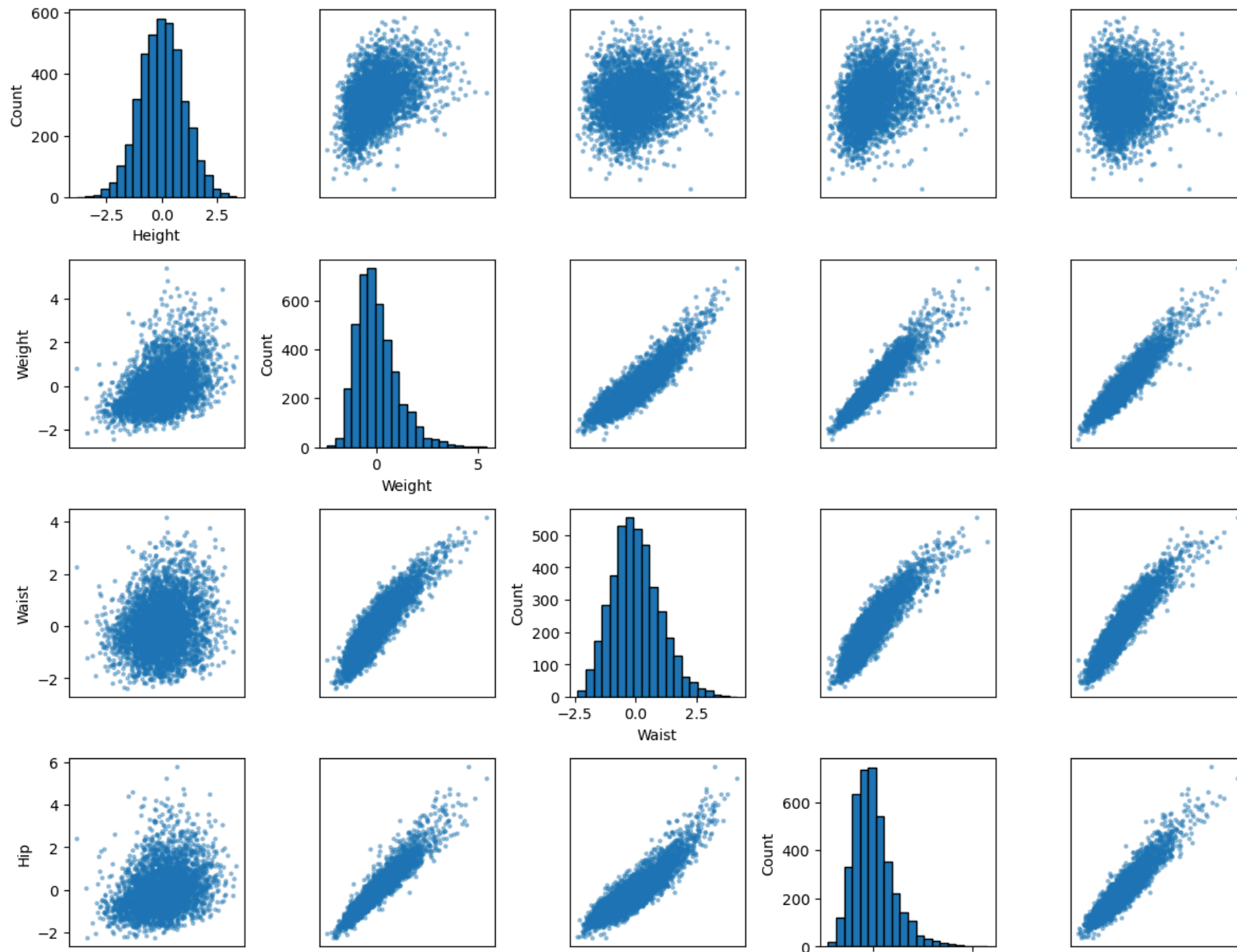


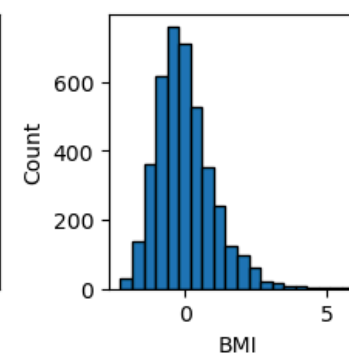
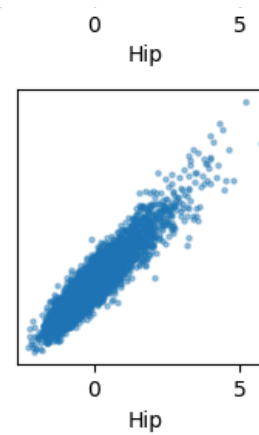
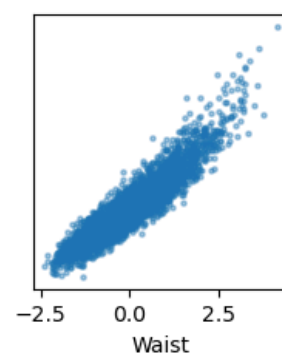
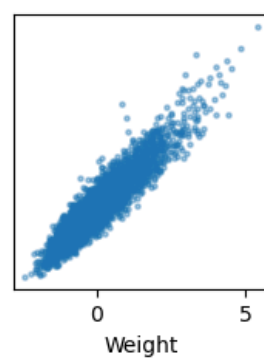
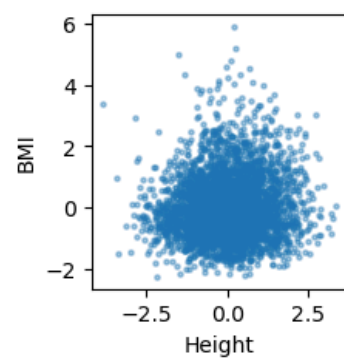
```
Z = np.column_stack([
    zmale[:, 1], # height
    zmale[:, 0], # weight
    zmale[:, 6], # waist
    zmale[:, 5], # hip
    zmale[:, 7]  # BMI
])

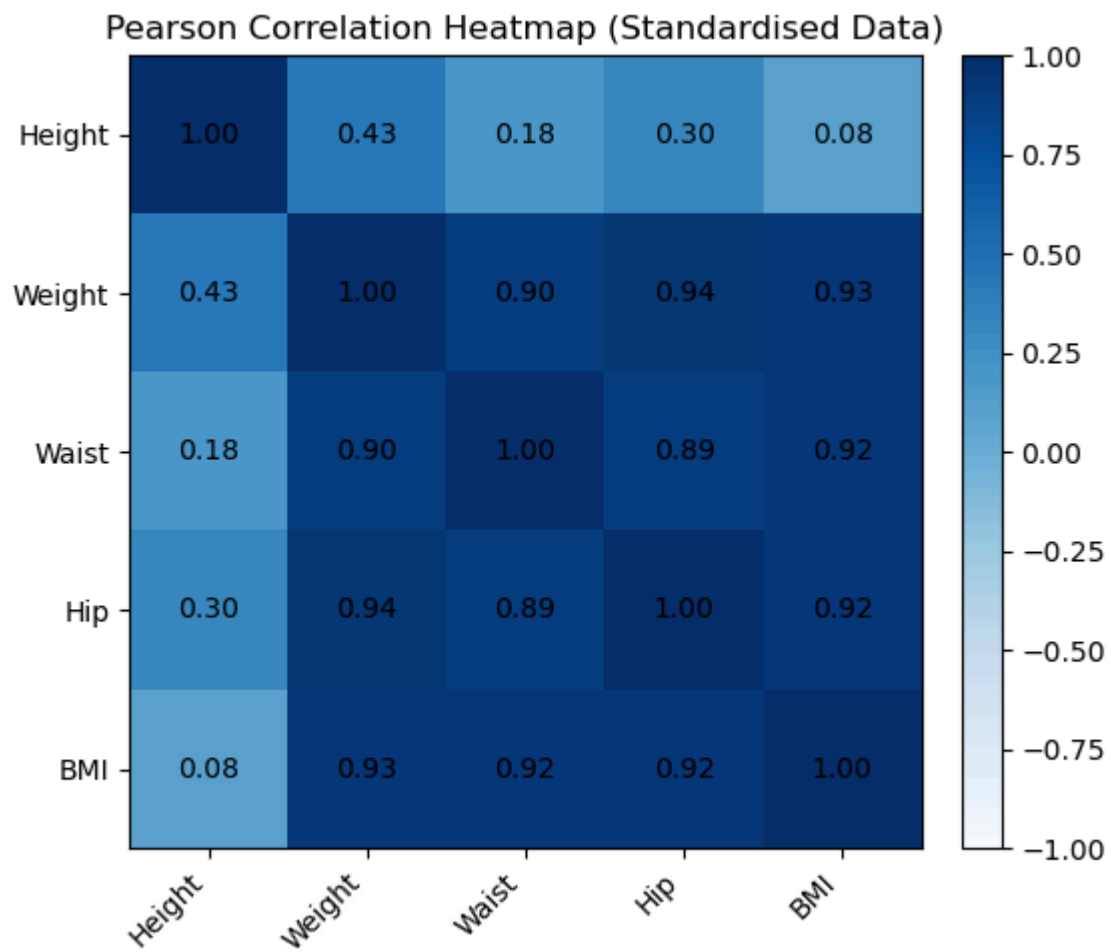
Z = Z[~np.isnan(Z).any(axis=1)]

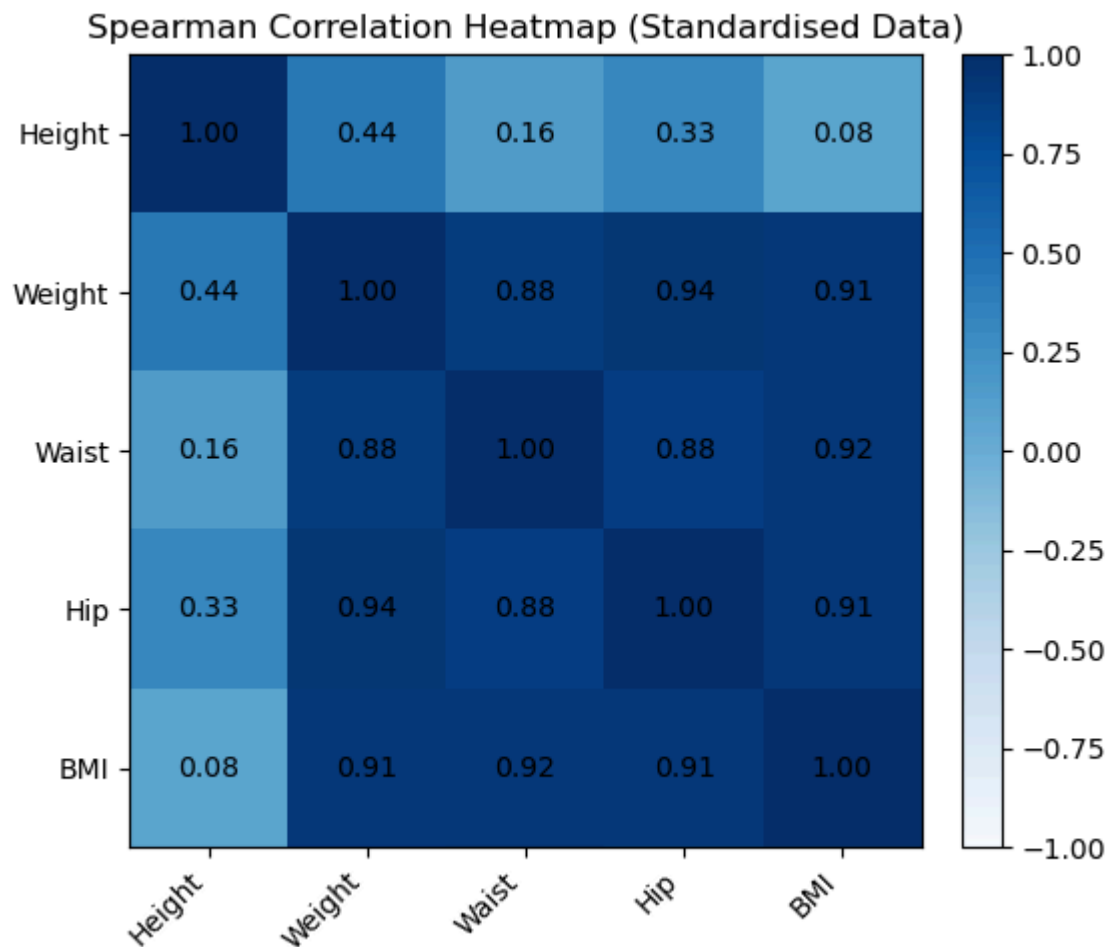
pearson_corr_z = np.corrcoef(Z, rowvar=False)
spearman_corr_z, _ = spearmanr(Z)

corr_heatmap(pearson_corr_z, labels, "Pearson Correlation Heatmap (Standardised Data)")
corr_heatmap(spearman_corr_z, labels, "Spearman Correlation Heatmap (Standardised Data)")
```









12.2 Interpretation of the Scatterplot Matrix for the Standardised Dataset

The scatterplot matrix based on the standardised male dataset (*zmale*) exhibits the same overall structure and relationships as observed in the original data. The key difference lies in the scale of the axes: all variables are now expressed in standard deviation units, with values typically ranging between approximately -3 and 3 and centred around zero.

Strong linear relationships remain clearly visible among weight, waist circumference, hip circumference, and BMI. These associations appear as narrow, elongated point clouds, indicating that the underlying relationships between these variables are preserved under standardisation. In particular, BMI continues to show strong positive associations with weight, waist, and hip circumference.

In contrast, height again displays weaker relationships with BMI and circumferential measurements, as reflected by the more dispersed scatter patterns. This confirms that height plays a comparatively limited role in explaining variation in adiposity-related measures, even after removing scale effects.

Overall, the standardised scatterplot matrix confirms that standardisation alters only the scale of the data and not the strength or nature of the relationships among variables. The similarity between this plot and the original scatterplot matrix supports the theoretical expectation that linear transformations preserve correlation structure.

12.3 Interpretation of Correlation Heatmaps for the Standardised Dataset

The Pearson and Spearman correlation heatmaps computed using the standardised male dataset (*zmale*) are numerically identical to those obtained from the original, unstandardised data. This outcome is expected and confirms the theoretical property that correlation coefficients are invariant under linear transformations such as shifting and scaling.

Strong positive correlations persist among weight, waist circumference, hip circumference, and BMI in both heatmaps. Correlation coefficients remain high (generally above 0.88), indicating that these variables continue to capture closely related aspects of body mass and adiposity, even after all variables have been expressed in standard deviation units.

Height again exhibits comparatively weak correlations with BMI, waist circumference, and hip circumference. The low correlation values demonstrate that standardisation does not artificially strengthen relationships; instead, it preserves the original dependence structure. The moderate correlation between height and weight remains unchanged, reflecting natural variation in body size rather than adiposity.

The close agreement between Pearson's and Spearman's coefficients further suggests that the observed relationships are largely linear and monotonic. Overall, the heatmaps confirm that standardisation alters only the scale of the variables and not the

strength or direction of their associations.

13.0 Summary and Key Findings

This study analysed adult anthropometric measurements derived from the National Health and Nutrition Examination Survey (NHANES 2020) to explore body composition characteristics and relationships among key physical variables. The analysis was conducted exclusively using NumPy matrices and numerical methods, ensuring transparent and low-level control over data handling and computation.

The datasets for adult males and females were obtained from a publicly available teaching repository and contained seven core anthropometric variables, including weight, standing height, limb lengths, and body circumferences. Preliminary inspection confirmed consistent data structure across sexes and only a small number of missing observations, which were handled appropriately during analysis.

Body Mass Index (BMI) was computed and appended as an additional derived variable for both datasets. This measure served as a central indicator of body composition and was used extensively in subsequent descriptive, visual, and correlational analyses.

Exploratory visualisations revealed that BMI distributions for both males and females are right-skewed, indicating a longer upper tail corresponding to higher BMI values. Box-and-whisker plots highlighted slight differences in central tendency and dispersion between the two groups, with females exhibiting marginally higher variability.

Numerical summary statistics supported these observations by quantifying measures of location, dispersion, and shape. Females displayed slightly higher spread in BMI values, while both groups exhibited positive skewness, consistent with known population-level BMI patterns.

Multivariate relationships among selected male anthropometric variables were explored using scatterplot matrices. Strong linear associations were observed between weight, waist circumference, hip circumference, and BMI, whereas height showed relatively weak relationships with adiposity-related measures.

Correlation analysis using both Pearson's and Spearman's coefficients confirmed these patterns. High correlations were observed among mass- and circumference-based variables, while height exhibited low correlation with BMI. The agreement between Pearson and Spearman measures indicates that these relationships are both linear and monotonic.

To facilitate scale-independent comparison, the male dataset was standardised using z-score transformation. Numerical validation confirmed that each standardised variable had a mean close to zero and a standard deviation close to one. Repeating the multivariate analyses on the standardised data produced identical correlation coefficients, demonstrating that standardisation affects scale but preserves relational structure.

Overall, the analysis provides a comprehensive examination of adult anthropometric characteristics and their interrelationships. It demonstrates effective use of NumPy for numerical data analysis, statistical summarisation, visualisation, and standardisation, while yielding meaningful insights into body composition patterns within th

===== END OF TASK2P =====

In []: