Statistical Intuitions and Applications

Assignment #1

library importation

In [ ]:
```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import random
import scipy.stats as stats
```

Task 1.

- As mentioned above, you will select a random sample of 100 individuals from the comapny's data set.
- You will then conduct analyzes on this random sample.
- Look at the code below. To select a random sample from the data, you should replace **Name** with your own name in the code.
- After you have done so run the code. The code will generate a csv file with a random sample of 100 participants. It will also be labeled with your name.
- REMEMBER: you need to add this csv file to a zip file along with your .ipynb. file when submitting your assignment.

data imporation

In [ ]:
```
try:
    df = pd.read_csv('eman alkhyeli.csv')
except FileNotFoundError:
    original_data = pd.read_csv("https://raw.githubusercontent.com/ZUCourses/SIA-Publi
    df1 = original_data.sample(100)
    df1.to_csv('eman alkhyeli.csv')
    df = pd.read_csv('eman alkhyeli.csv')
    df = pd.DataFrame(df)
    df.to_csv('eman alkhyeli.csv')

df.head()
```

Out[ ]:

	Unnamed: 0.1	Unnamed: 0	bia_di	bii_di	bit_di	che_de	che_di	elb_di	wri_di	kne_di	...	bic_gi	for_
0	485	235	40.1	26.4	32.0	21.8	30.2	15.8	12.4	20.7	...	38.5	3(
1	23	493	38.5	25.6	31.7	17.0	25.6	11.8	10.2	16.8	...	28.1	2:
2	435	6	43.3	27.0	31.5	19.6	31.3	14.0	11.5	18.8	...	33.0	2ε
3	257	275	39.6	28.6	31.2	18.0	27.2	12.4	10.1	19.3	...	27.4	2:
4	502	17	39.9	30.0	34.5	21.0	29.4	15.6	11.9	21.2	...	37.0	2ε

5 rows × 27 columns

Task 2.

- Now that you have your data set you are ready to start analyzing it!
- The first step is to explore your dataset.
- Look at the variables that make up the data set.
- Once you've done so, imagine you are writing a report for the fitness company that hired you.
- Start with a brief introduction to the research question you are exploring, then the dataset you are analysing (e.g., what is the sample you are analyzing? What are the variables?)
- Assume that your #Audience is the company's leadership. They will be with what you are reporting.

Add this brief introduction below.

Introduction

In this comprehensive analysis, we delve into a dataset encompassing the intricate body measurements of 507 physically active individuals, constituting 247 men and 260 women. The dataset encapsulates a diverse array of metrics, ranging from fundamental demographic details such as age, weight, and height to intricate body girth and skeletal diameter measurements. The richness of this dataset presents an invaluable opportunity for our organization's leadership to gain profound insights into the distinct physical characteristics prevalent among individuals actively engaged in physical pursuits.

Variables Under Analysis

The variables to be scrutinized are not merely data points but windows into the nuanced aspects of human physiology. These include:

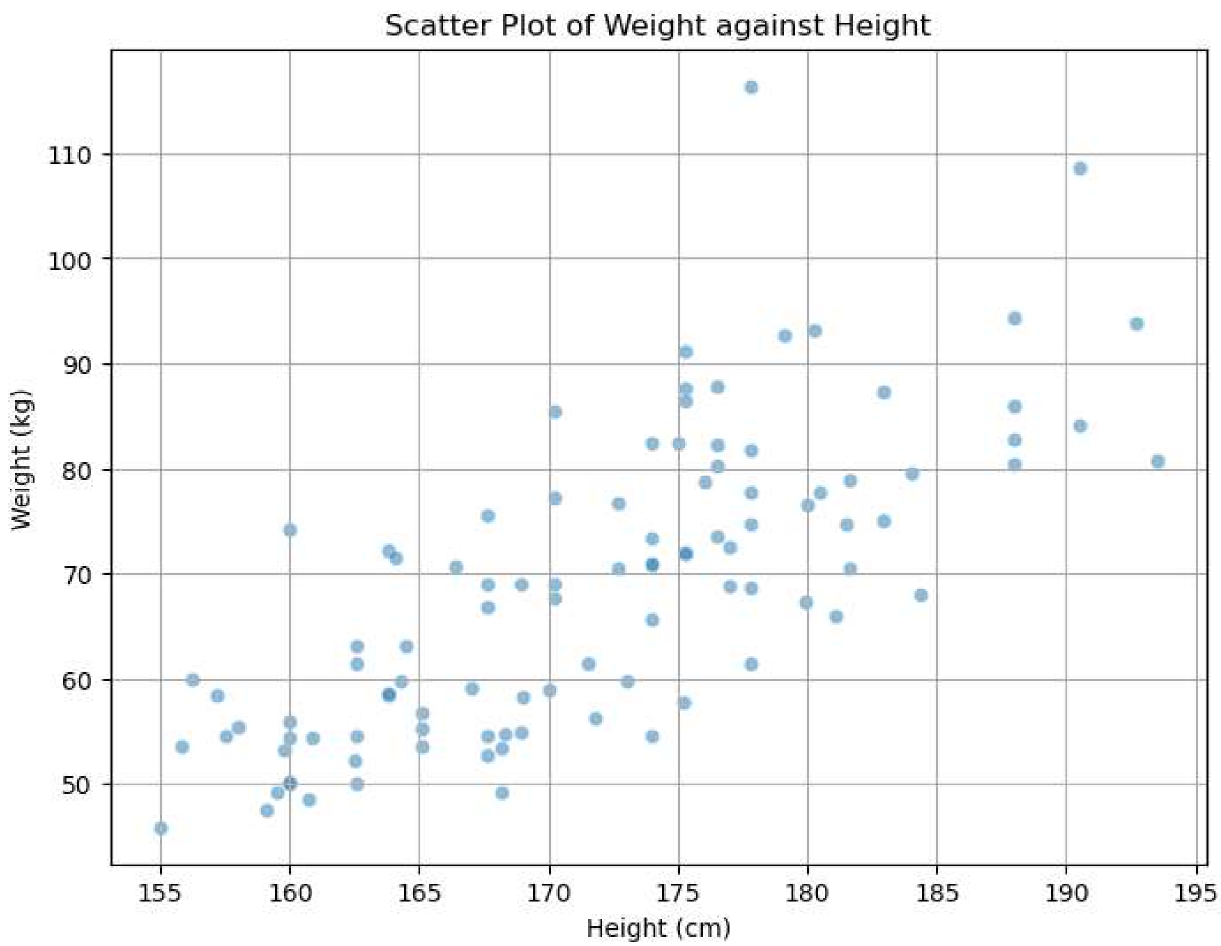
1. **Biacromial Diameter (bia_di):** The width of an individual's shoulders, a crucial metric in understanding upper body structure.
2. **Biiliac Diameter (bii_di):** The pelvic breadth, providing insights into the hip structure and pelvic dimensions.

3. **Bitrochanteric Diameter (bit_di):** The width between the outer points of the hips, a key measure in understanding the lower body structure.
4. **Chest Depth (che_de):** The measurement between the spine and sternum at nipple level during mid-expiration, offering insights into chest morphology.
5. **Chest Diameter (che_di):** The chest circumference at nipple level during mid-expiration, indicative of overall chest girth.
6. **Elbow Diameter (elb_di):** The sum of two elbow diameters, providing information about upper limb skeletal structure.
7. **Wrist Diameter (wri_di):** The sum of two wrist diameters, reflecting the girth of the wrists.
8. **Knee Diameter (kne_di):** The sum of two knee diameters, indicative of lower limb skeletal dimensions.
9. **Ankle Diameter (ank_di):** The sum of two ankle diameters, offering insights into the lower extremity structure.
10. **Shoulder Girth (sho_gi):** The girth measured over deltoid muscles, a crucial metric for assessing shoulder development.
11. **Chest Girth (che_gi):** The girth measured at nipple level in males and just above breast tissue in females during mid-expiration.
12. **Waist Girth (wai_gi):** The girth at the narrowest part of the torso below the rib cage, providing insights into waistline dimensions.
13. **Navel (Abdominal) Girth (nav_gi):** The girth measured at the umbilicus and iliac crest using the iliac crest as a landmark.
14. **Hip Girth (hip_gi):** The girth measured at the level of bitrochanteric diameter, indicative of hip dimensions.
15. **Thigh Girth (thi_gi):** The girth measured below the gluteal fold, serving as an indicator of thigh development.
16. **Bicep Girth (bic_gi):** The girth measured when flexed, providing insights into the development of the biceps.
17. **Forearm Girth (for_gi):** The girth measured when extended with the palm up, indicative of forearm development.
18. **Knee Diameter (kne_gi):** Repeated for emphasis, the sum of two knee diameters, crucial in understanding lower limb dimensions.
19. **Calf Maximum Girth (cal_gi):** The average of right and left calf girths, offering insights into calf muscle development.
20. **Ankle Minimum Girth (ank_gi):** The average of right and left ankle girths, providing nuanced details about ankle dimensions.
21. **Wrist Minimum Girth (wri_gi):** The average of right and left wrist girths, capturing the slender dimensions of the wrists.
22. **Age (age):** The chronological age of the individuals.
23. **Weight (wgt):** The weight of the individuals in kilograms.
24. **Height (hgt):** The height of the individuals in centimeters.
25. **Gender (sex):** A categorical variable indicating gender, with 1 denoting male and 0 denoting female.

Research Question

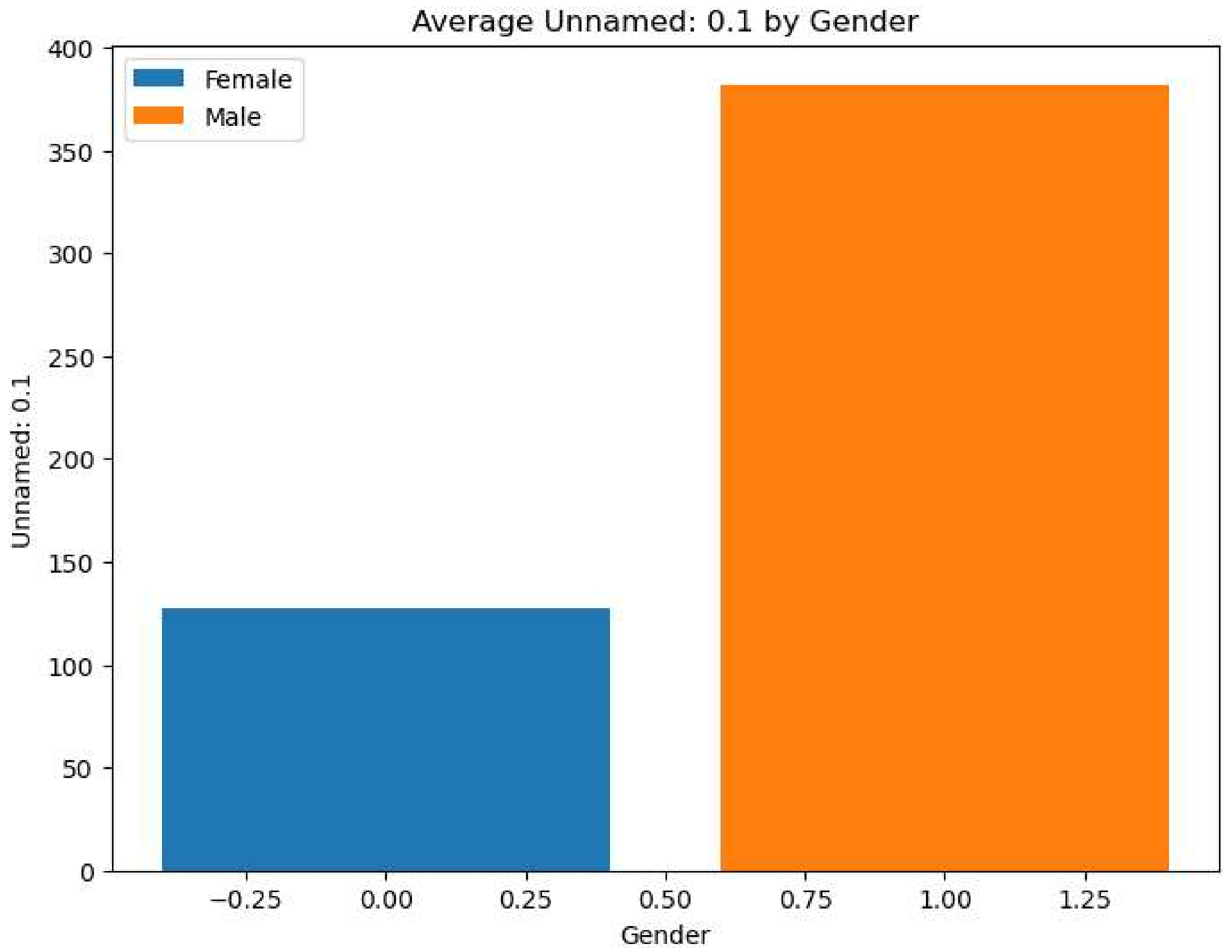
Our focal inquiry revolves around unraveling the intricate interplay of these variables, particularly discerning how body measurements vary between genders among physically active individuals. This analysis seeks to elucidate whether there exist significant distinctions in body girth measurements and skeletal diameter measurements between men and women within the context of an active lifestyle. The findings promise not only a deeper understanding of physical diversity but also actionable insights for our leadership to tailor strategies and initiatives that resonate with the unique physiological characteristics of our target audience.

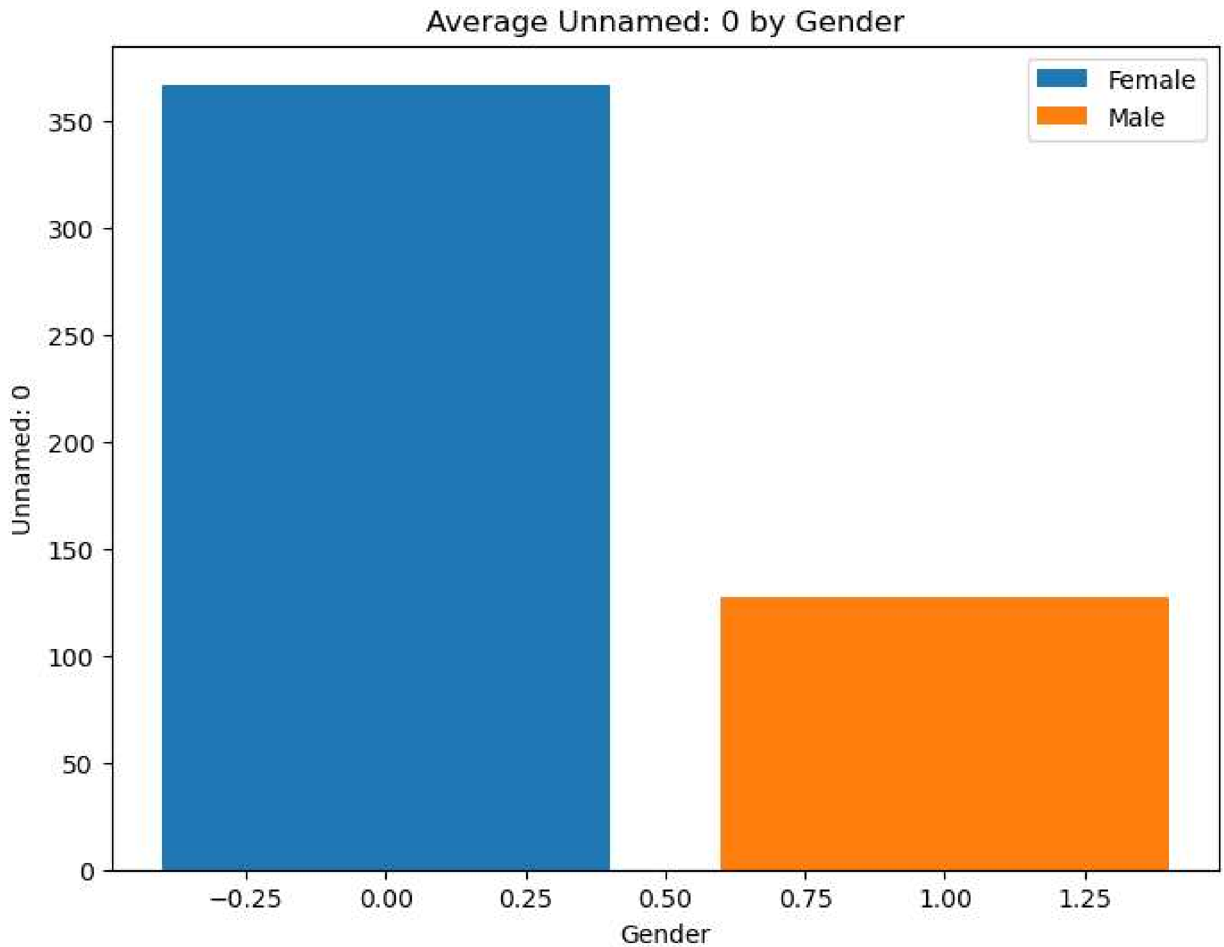
In [ ]:
```
import seaborn as sns

# Scatter plot of weight against height using Seaborn
plt.figure(figsize=(8, 6))
sns.scatterplot(data=df, x='hgt', y='wgt', alpha=0.5)
plt.title('Scatter Plot of Weight against Height')
plt.xlabel('Height (cm)')
plt.ylabel('Weight (kg)')
plt.grid(True)
plt.show()
```

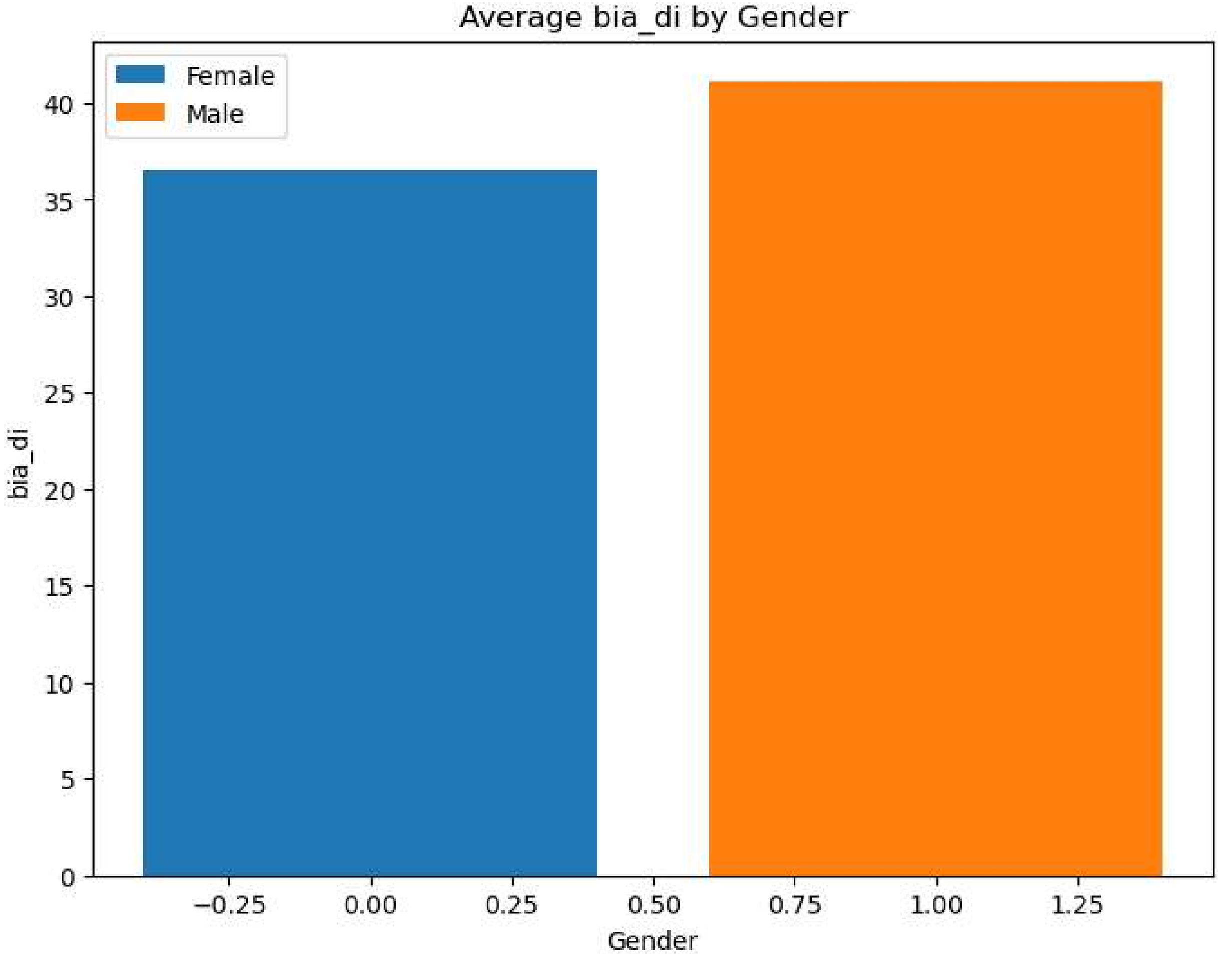


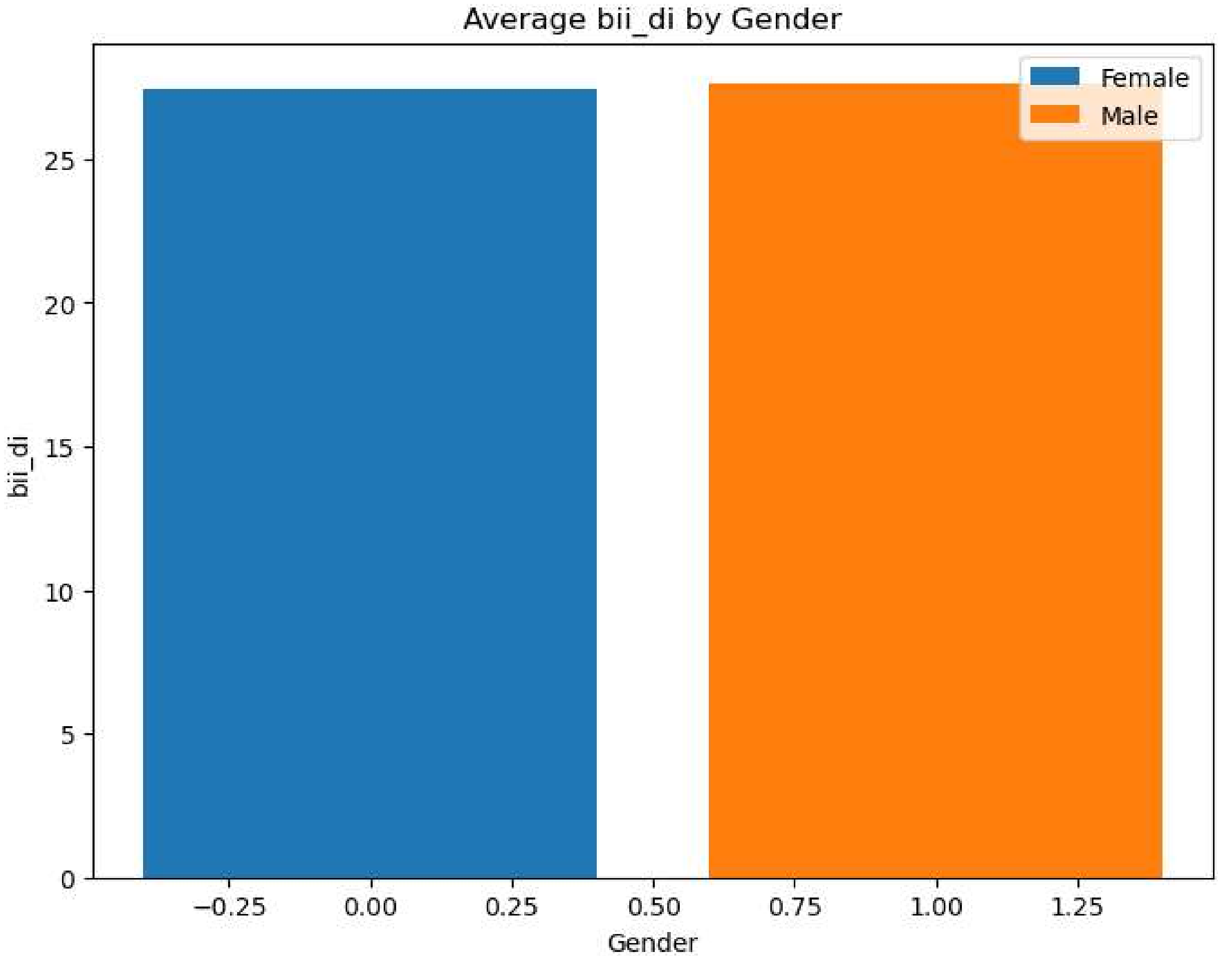
In [ ]:
```
import matplotlib.pyplot as plt

# Group the DataFrame by gender
grouped = df.groupby('sex')

# Iterate through each column and create a bar plot
for column in df.columns[:-1]:  # Exclude the 'sex' column
```

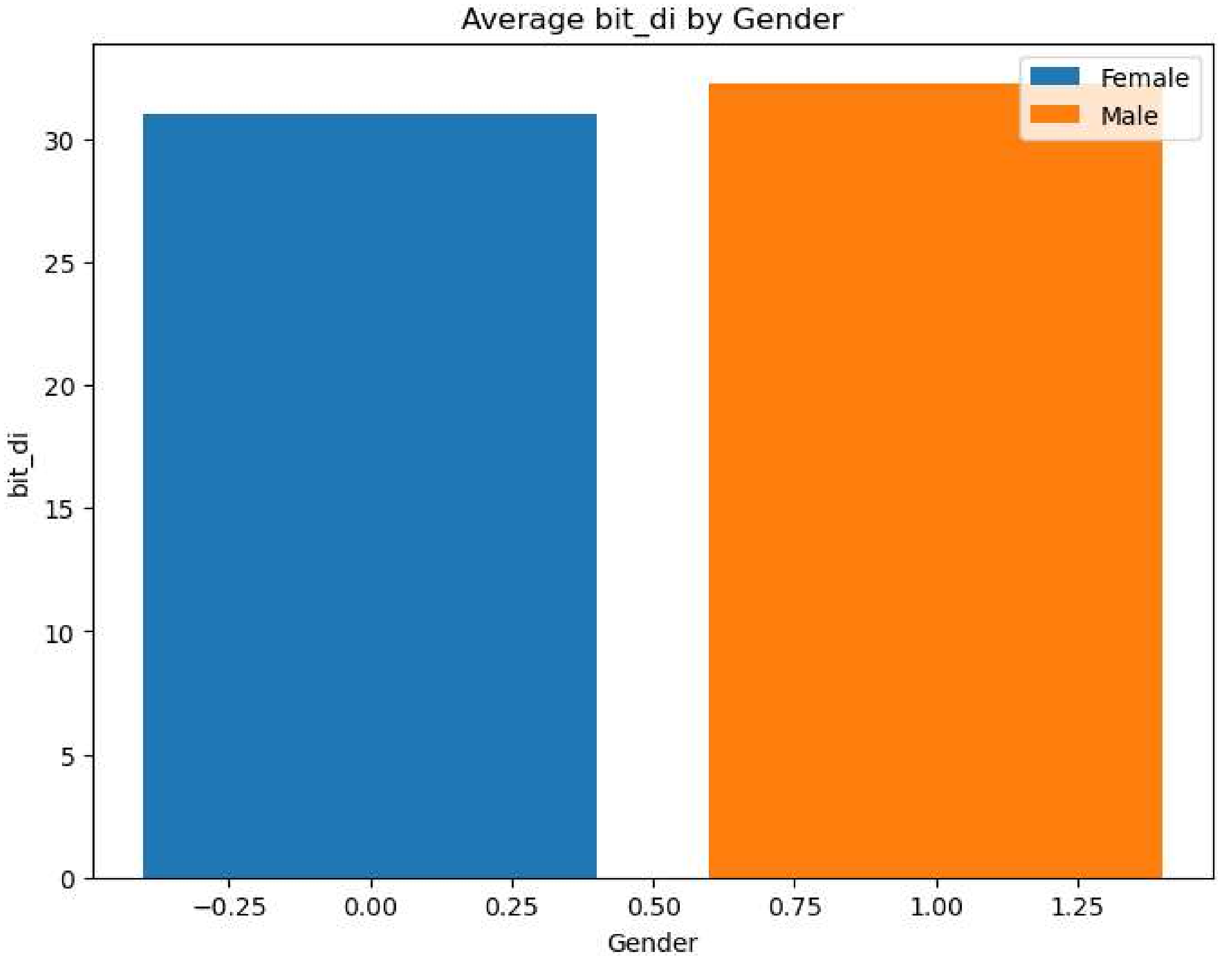
```
    plt.figure(figsize=(8, 6))
    for gender, data in grouped:
        plt.bar(gender, data[column].mean(), label='Male' if gender == 1 else 'Female'
    plt.title(f'Average {column} by Gender')
    plt.xlabel('Gender')
    plt.ylabel(column)
    plt.legend()
    plt.show()
```

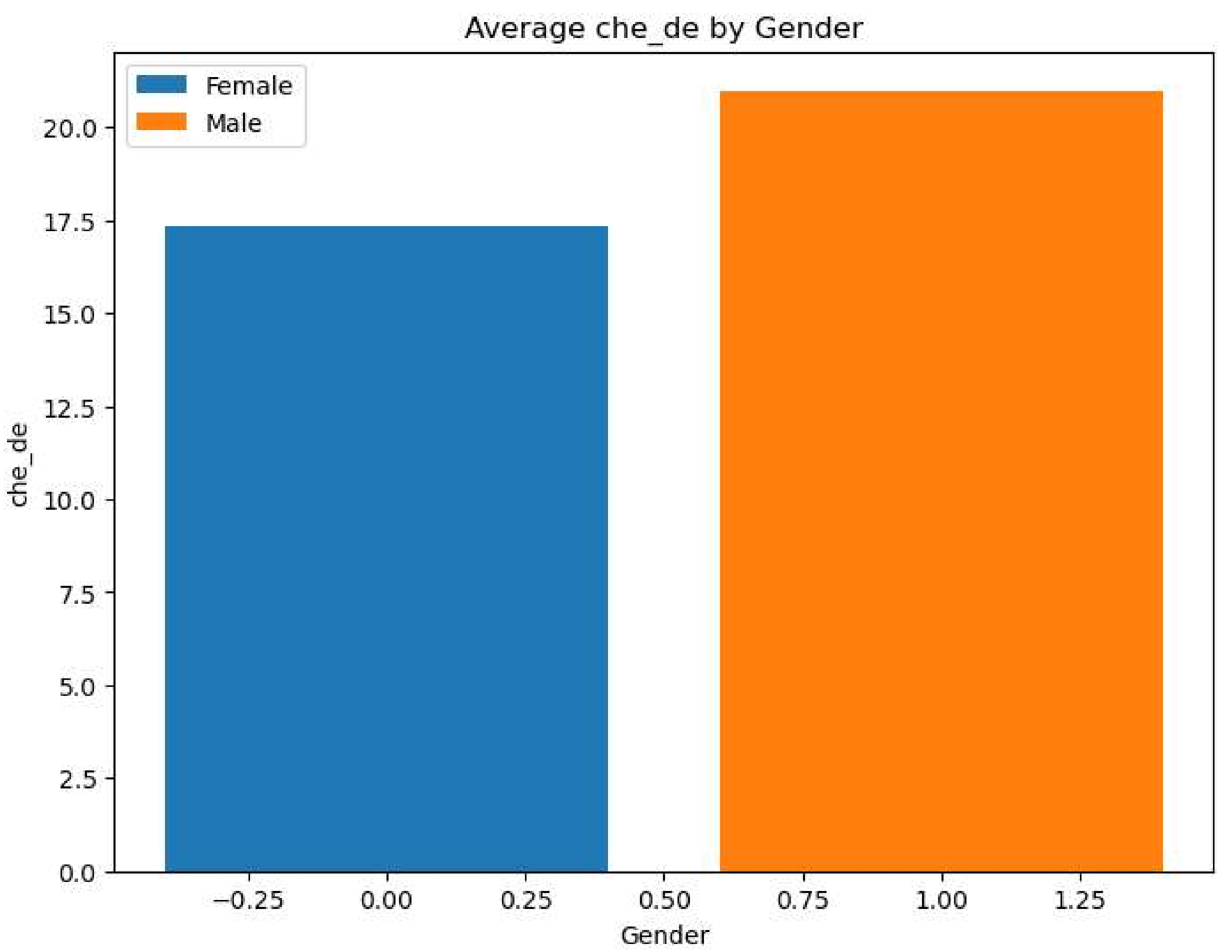


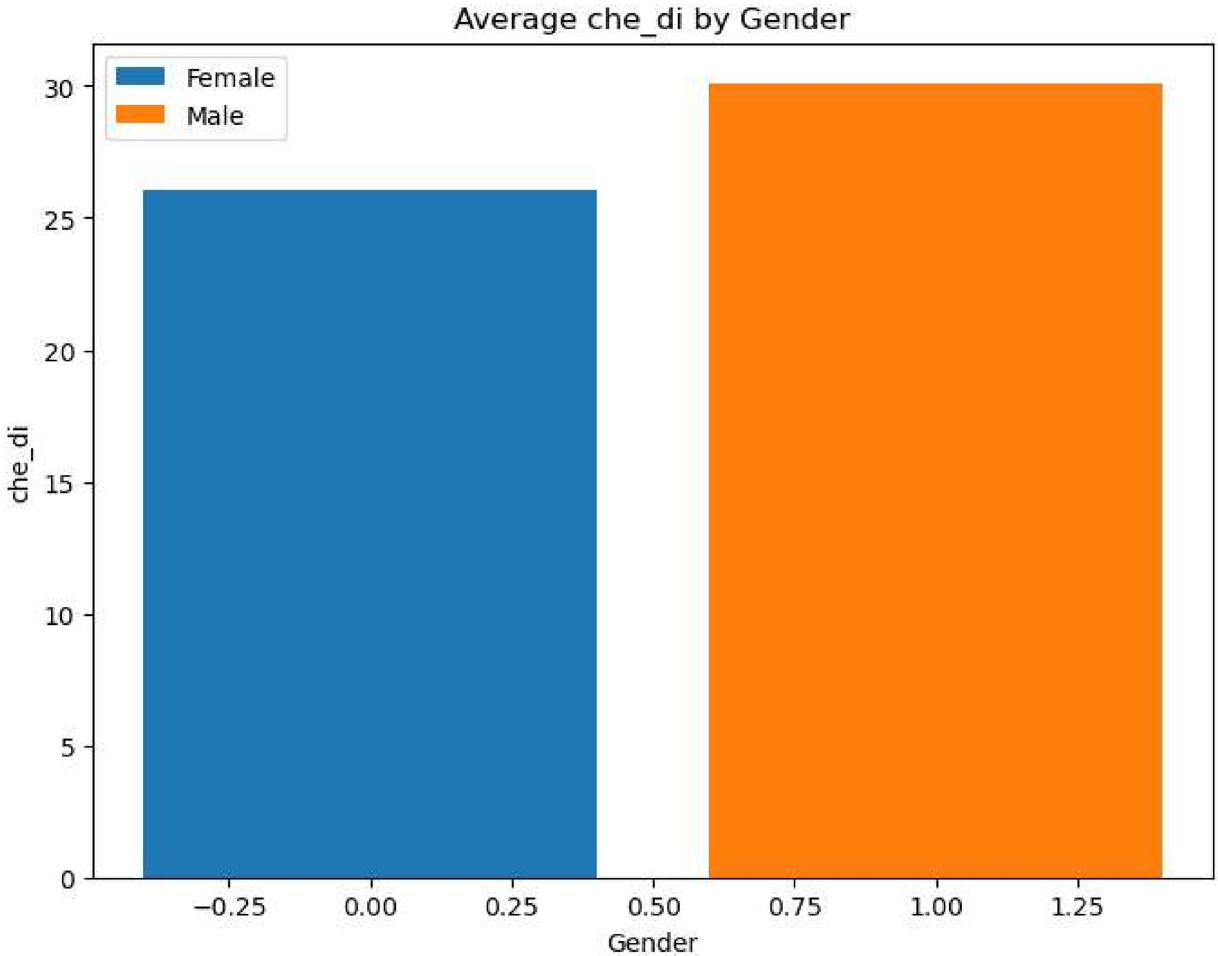

Average Unnamed: 0 by Gender
Female
Male
Unnamed: 0
0
50
100
150
200
250
300
350
−0.25
0.00
0.25
0.50
0.75
1.00
1.25
Gender

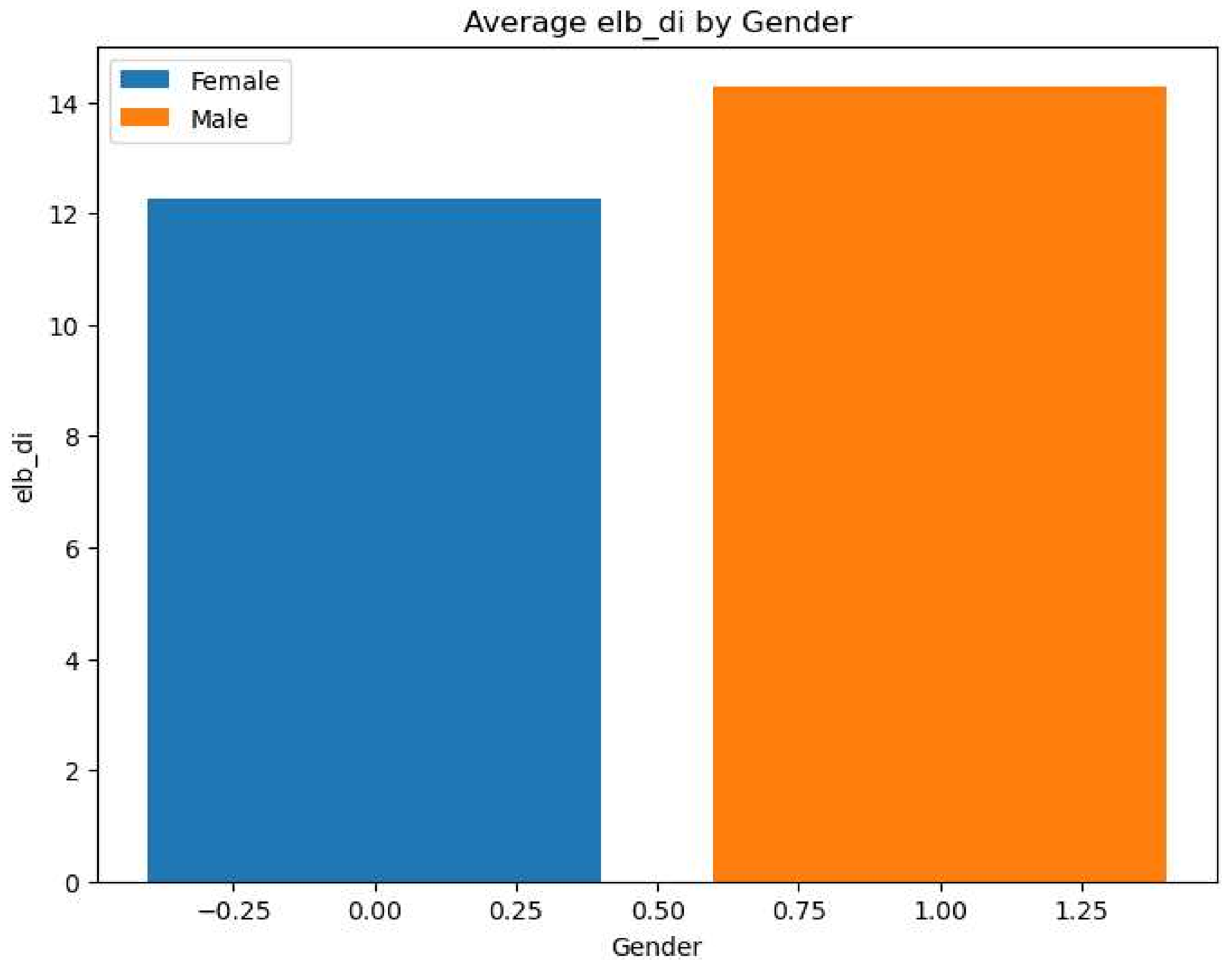

Average bia_di by Gender
Female
Male
40
35
30
25
20
15
10
5
0
bia_di
−0.25
0.00
0.25
0.50
0.75
1.00
1.25
Gender

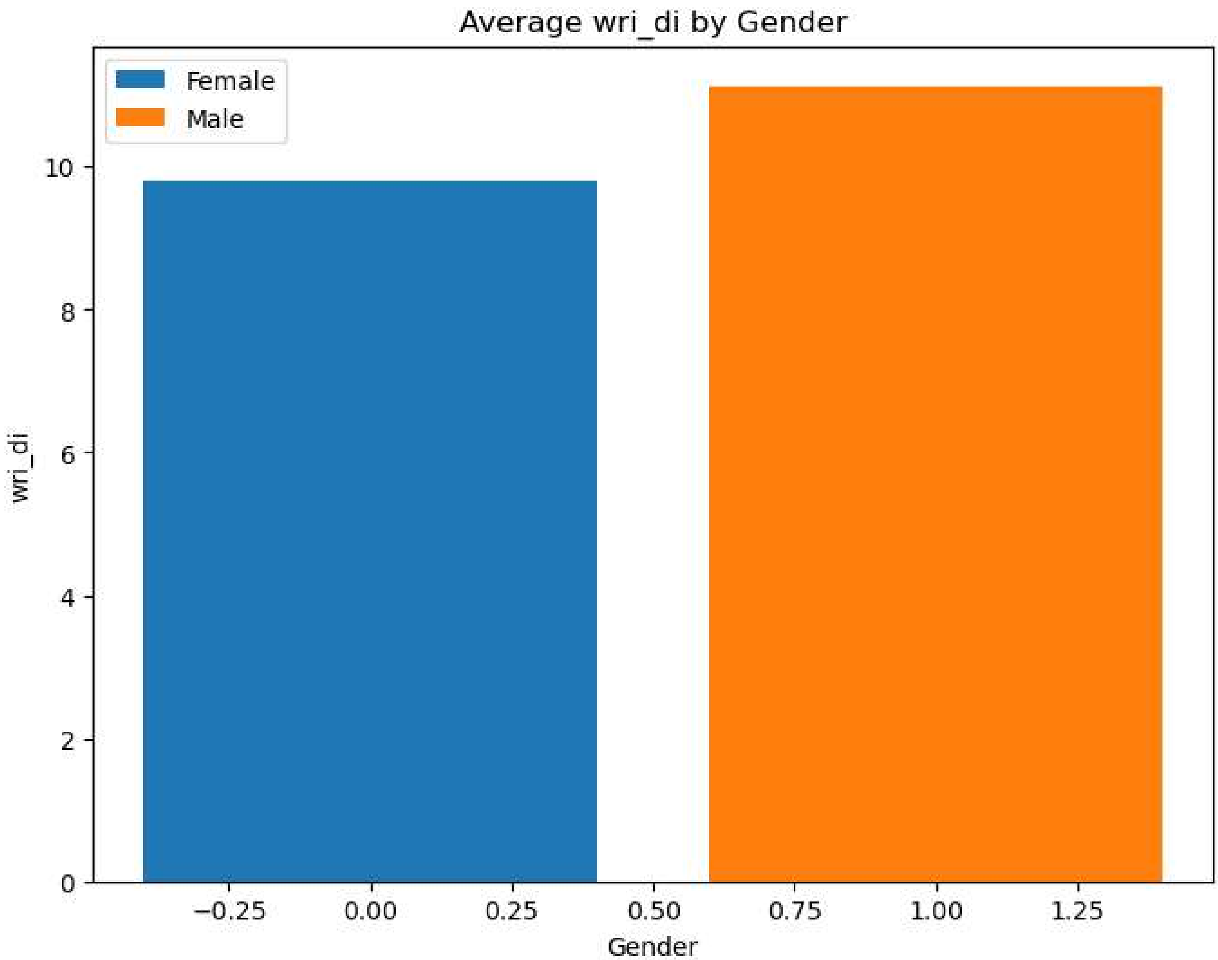

Average bii_di by Gender
Female
Male
bii_di
0
5
10
15
20
25
−0.25
0.00
0.25
0.50
0.75
1.00
1.25
Gender

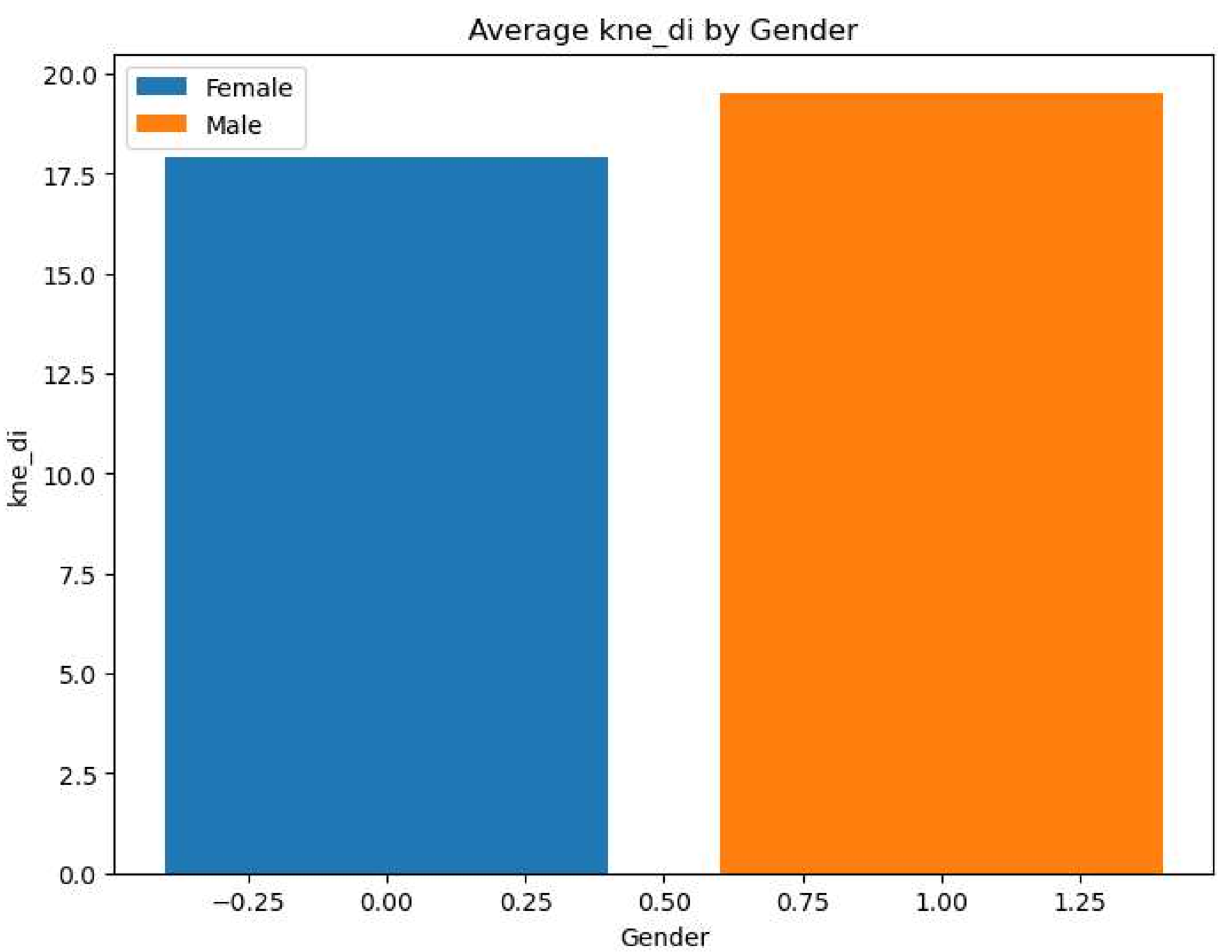

Average bit_di by Gender
Female
Male
bit_di
Gender
0
5
10
15
20
25
30
−0.25
0.00
0.25
0.50
0.75
1.00
1.25

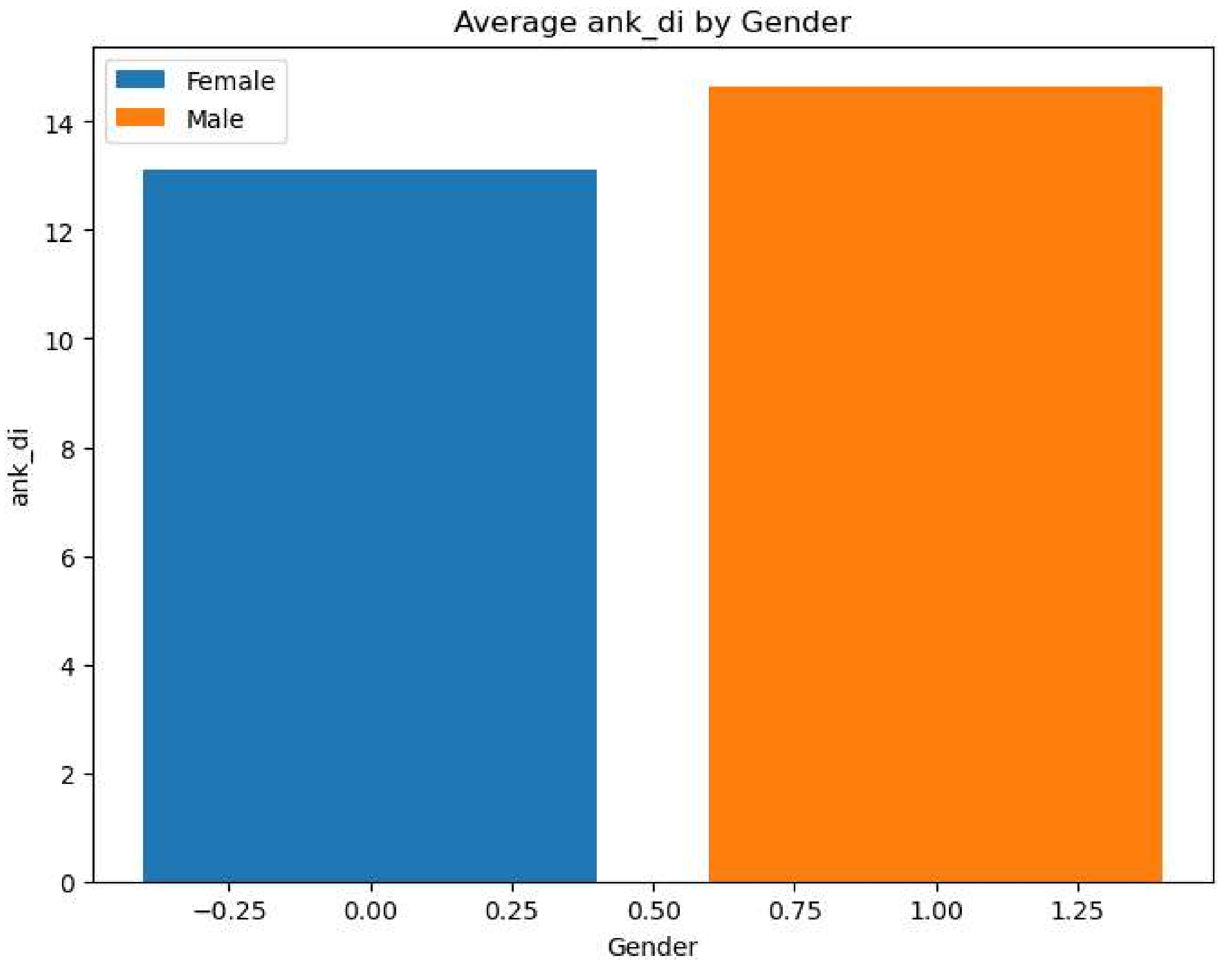

Average che_de by Gender
Female
Male
che_de
Gender
0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5
20.0
−0.25
0.00
0.25
0.50
0.75
1.00
1.25

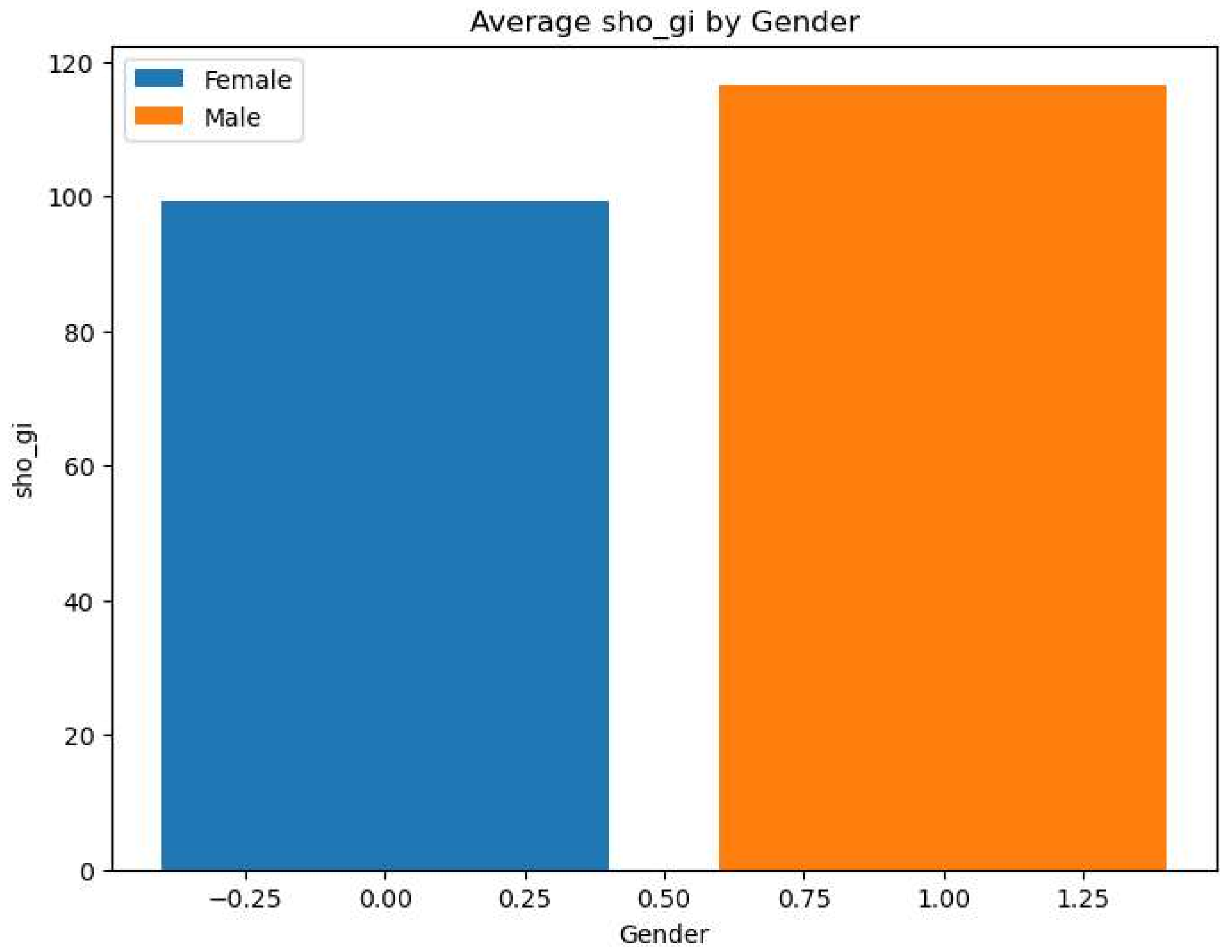

Average che_di by Gender
Female
Male
che_di
30
25
20
15
10
5
0
−0.25
0.00
0.25
0.50
0.75
1.00
1.25
Gender

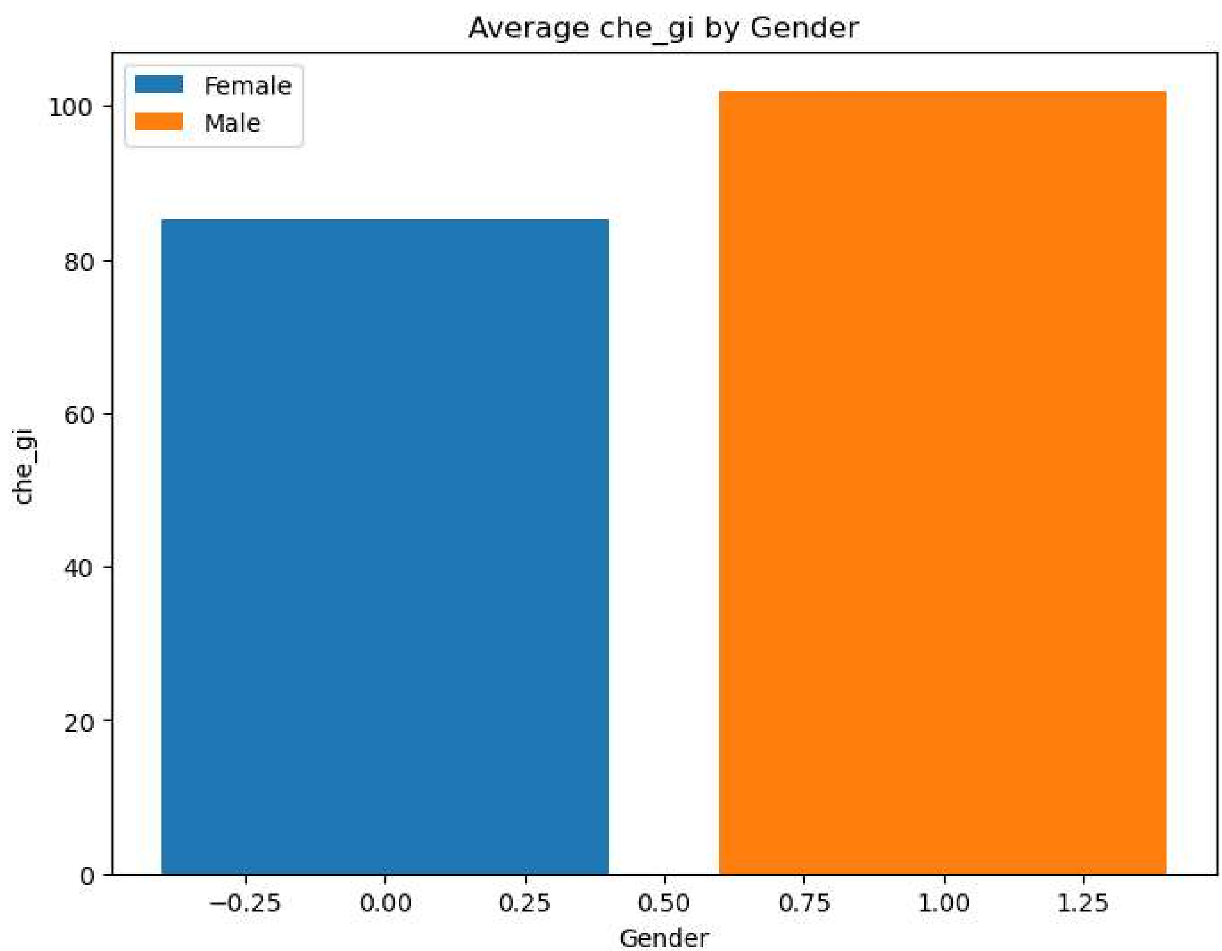

Average elb_di by Gender
Female
Male
14
12
10
8
6
4
2
0
elb_di
−0.25
0.00
0.25
0.50
0.75
1.00
1.25
Gender

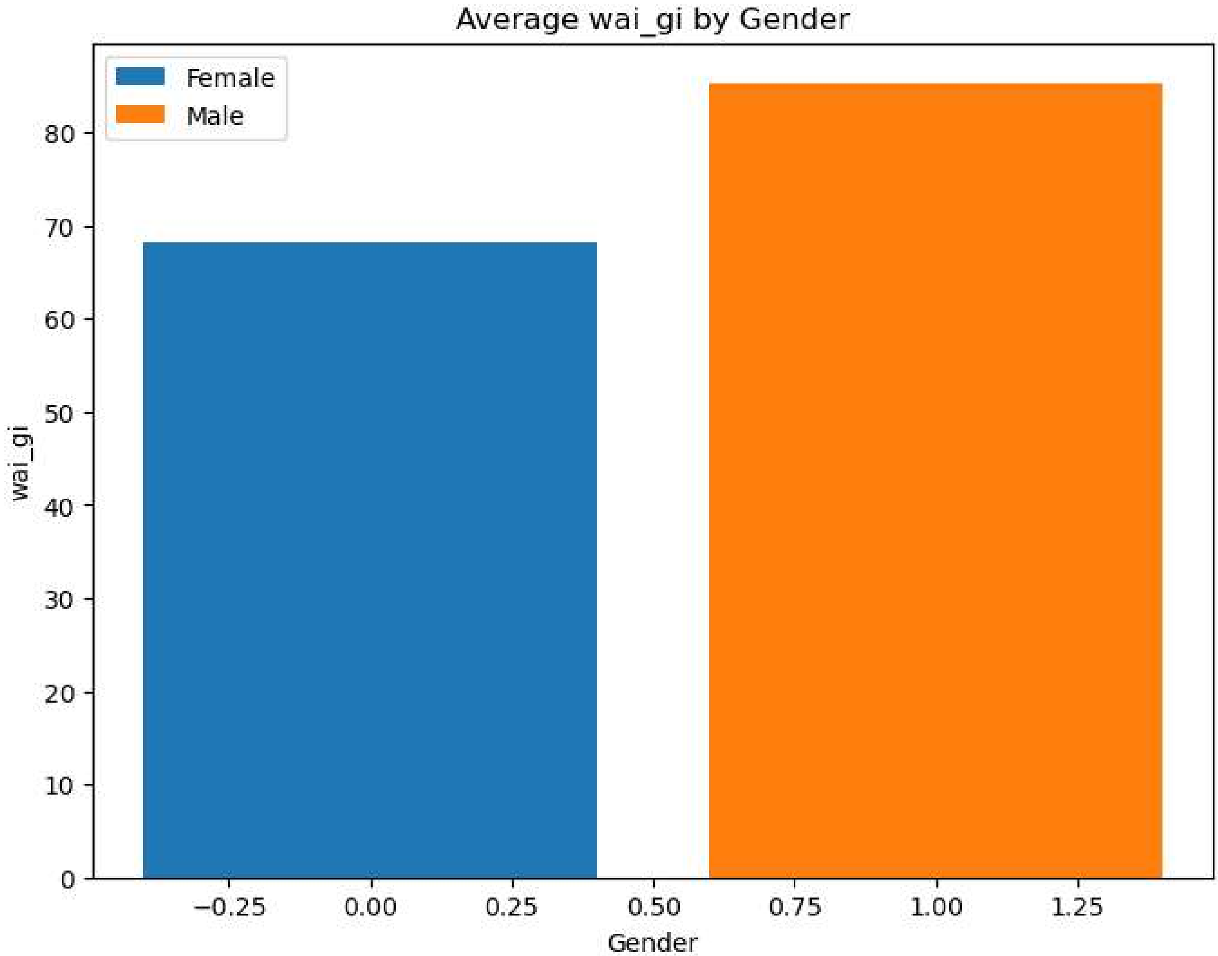

Average wri_di by Gender
Female
Male
wri_di
0
2
4
6
8
10
−0.25
0.00
0.25
0.50
0.75
1.00
1.25
Gender

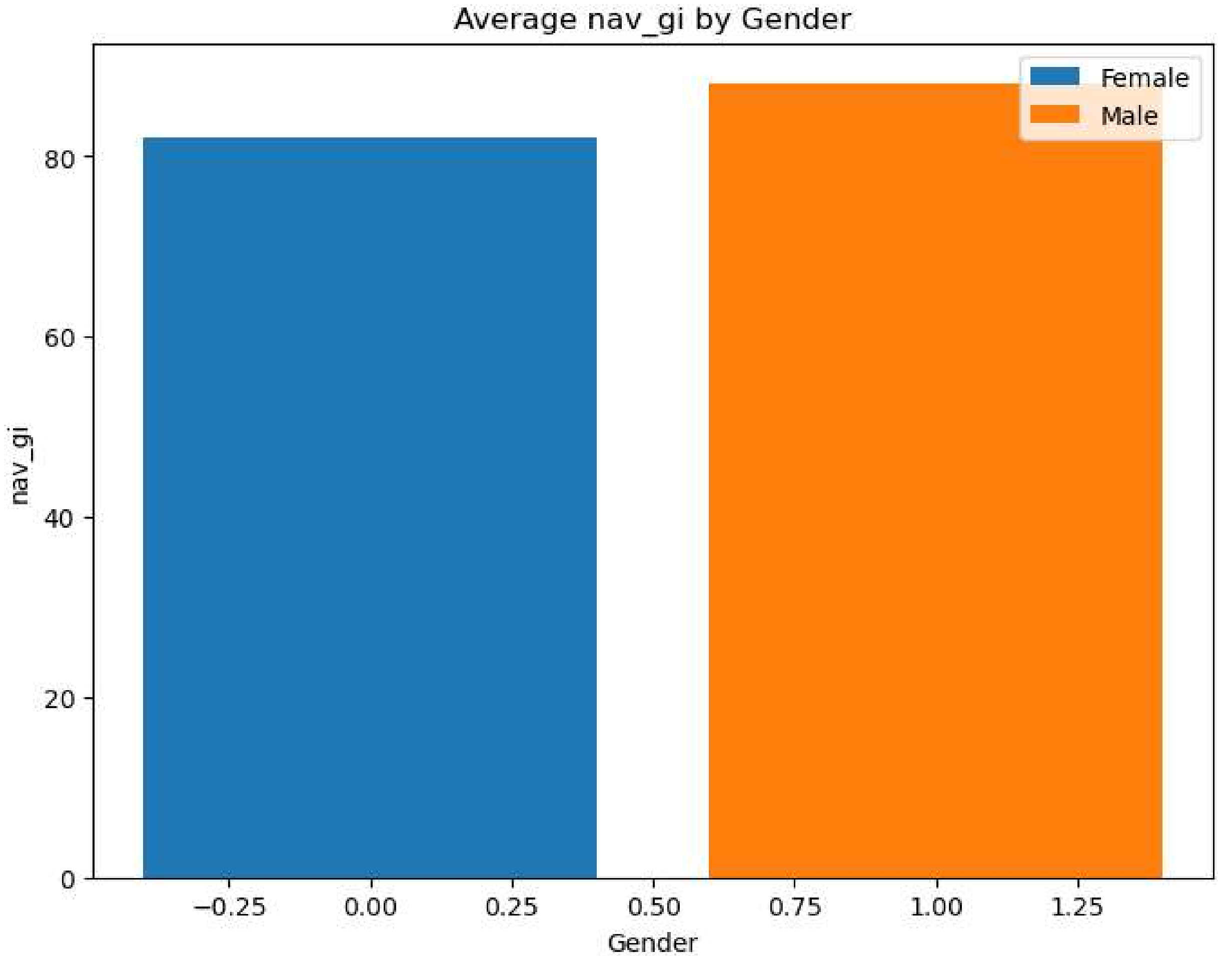

Average kne_di by Gender
Female
Male
kne_di
0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5
20.0
−0.25
0.00
0.25
0.50
0.75
1.00
1.25
Gender

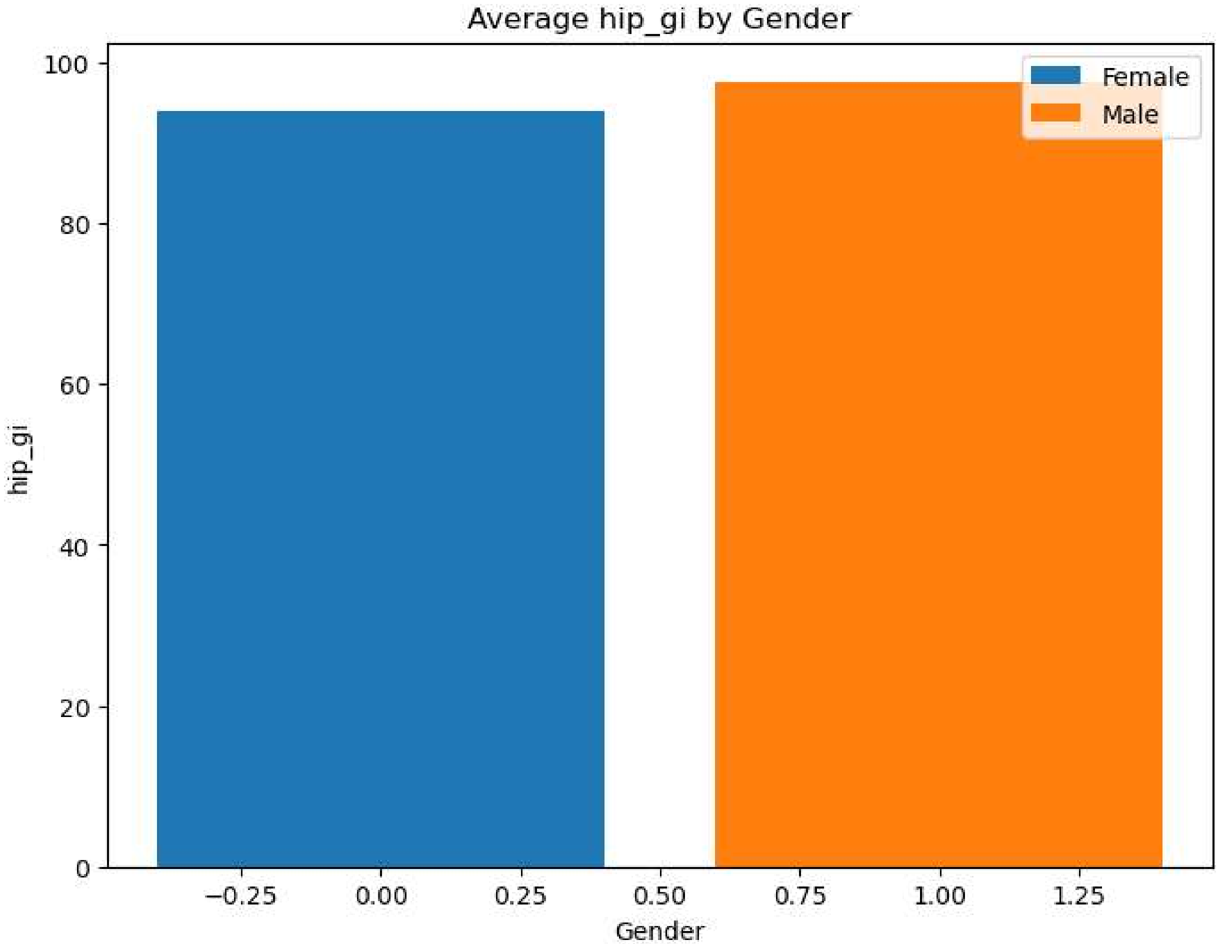

Average ank_di by Gender
Female
Male
ank_di
0
2
4
6
8
10
12
14
−0.25
0.00
0.25
0.50
0.75
1.00
1.25
Gender

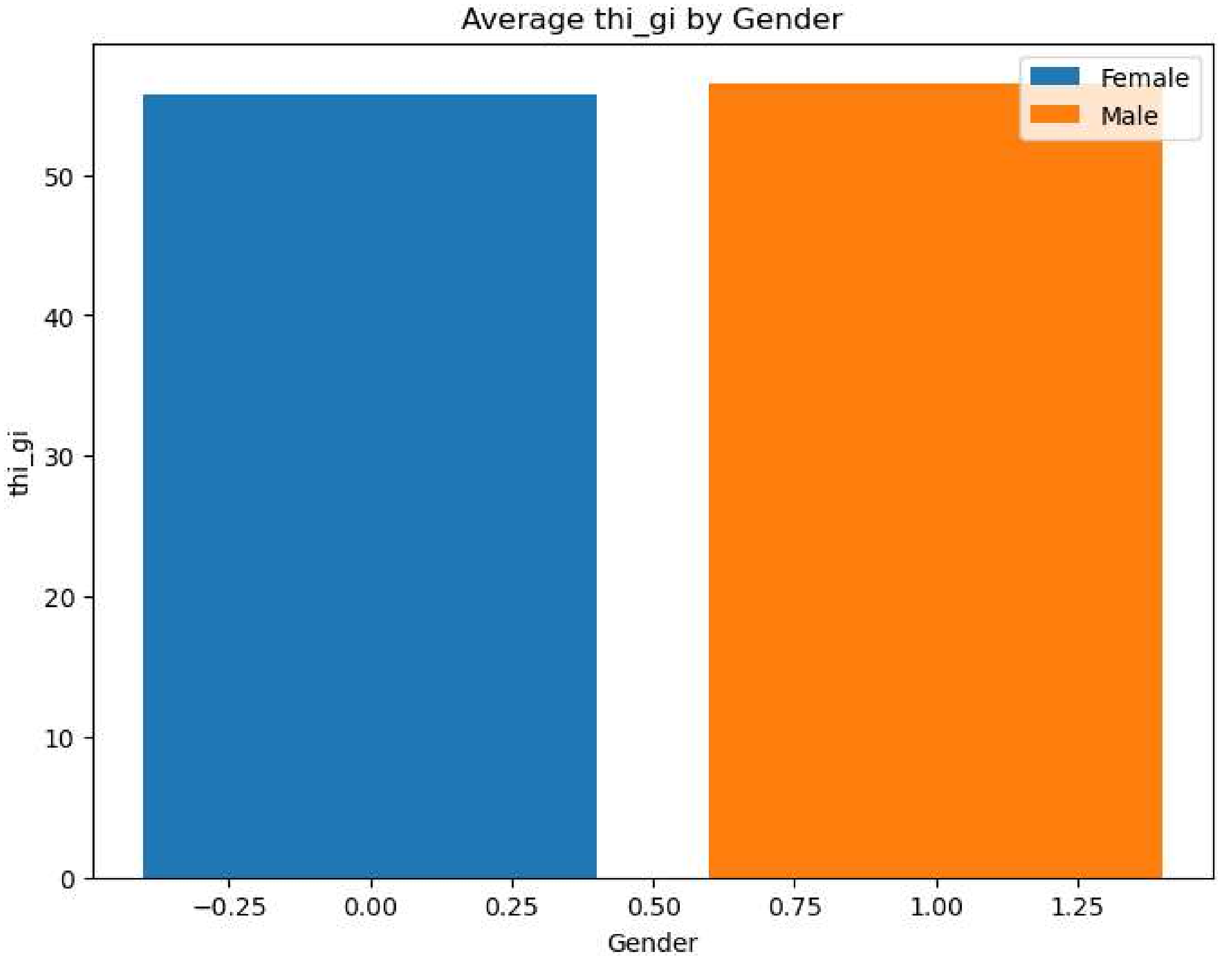

Average sho_gi by Gender
Female
Male
sho_gi
Gender
0
20
40
60
80
100
120
−0.25
0.00
0.25
0.50
0.75
1.00
1.25

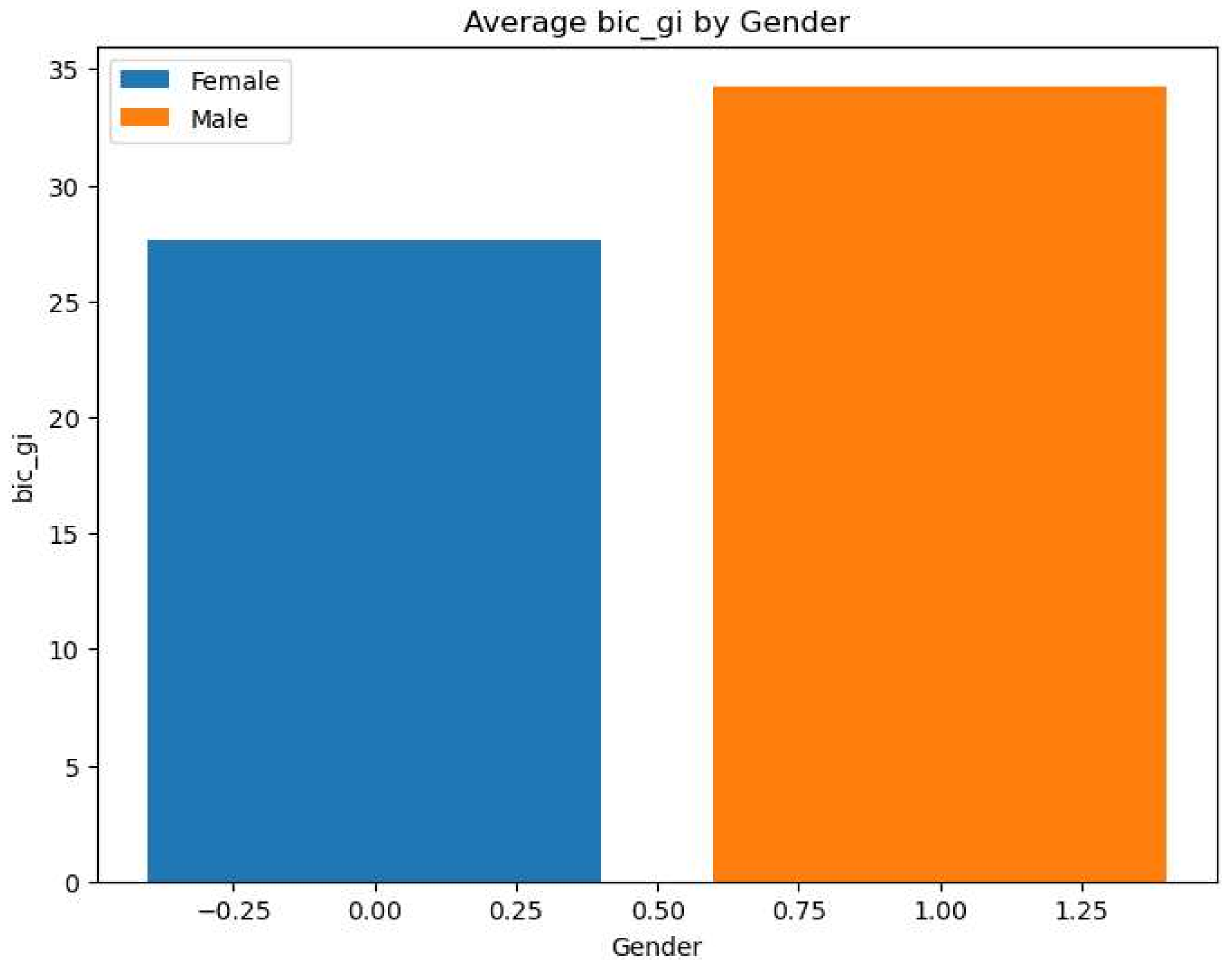

Average che_gi by Gender
Female
Male
che_gi
Gender
0
20
40
60
80
100
−0.25
0.00
0.25
0.50
0.75
1.00
1.25

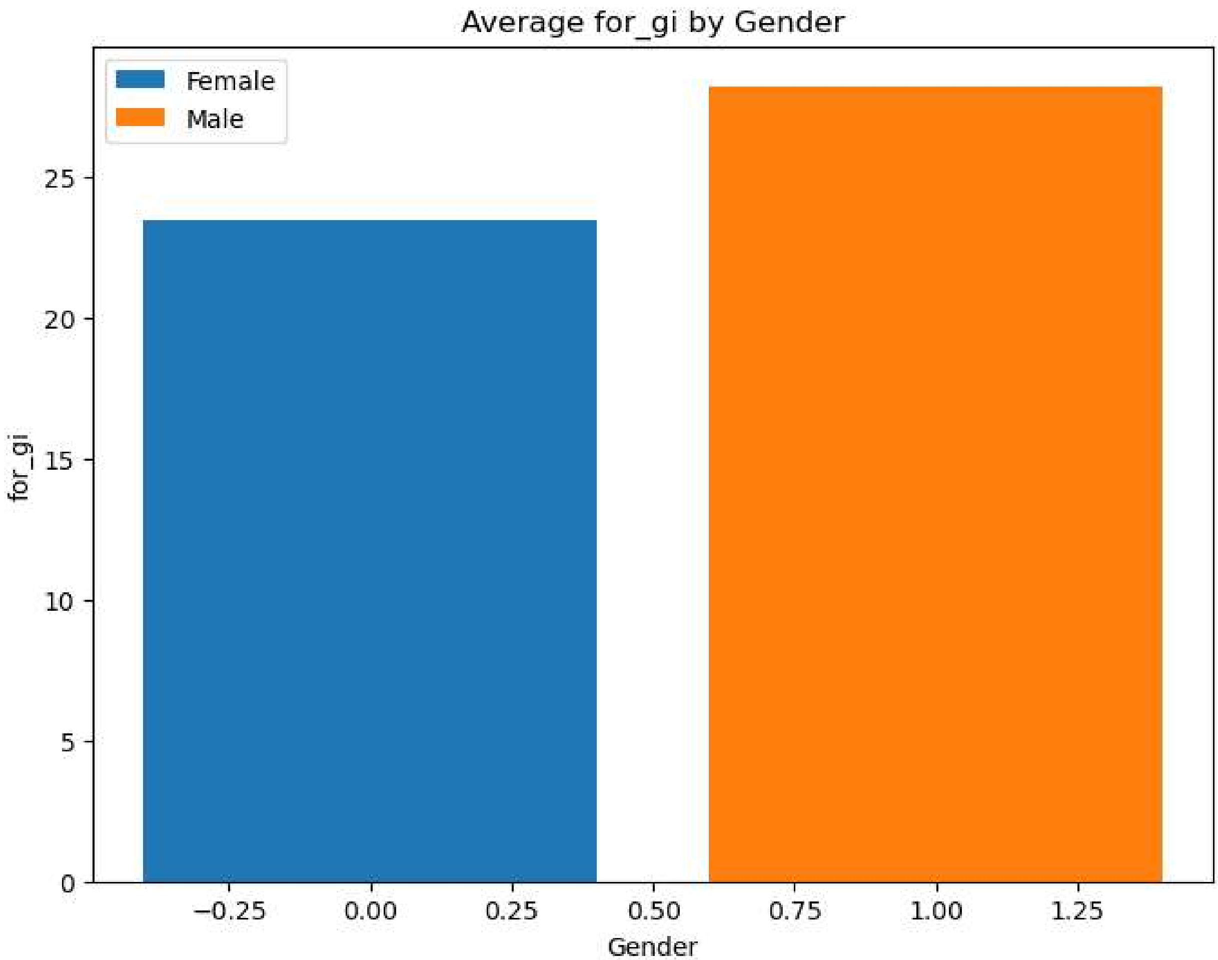

Average wai_gi by Gender
Female
Male
80
70
60
50
40
30
20
10
0
wai_gi
−0.25
0.00
0.25
0.50
0.75
1.00
1.25
Gender

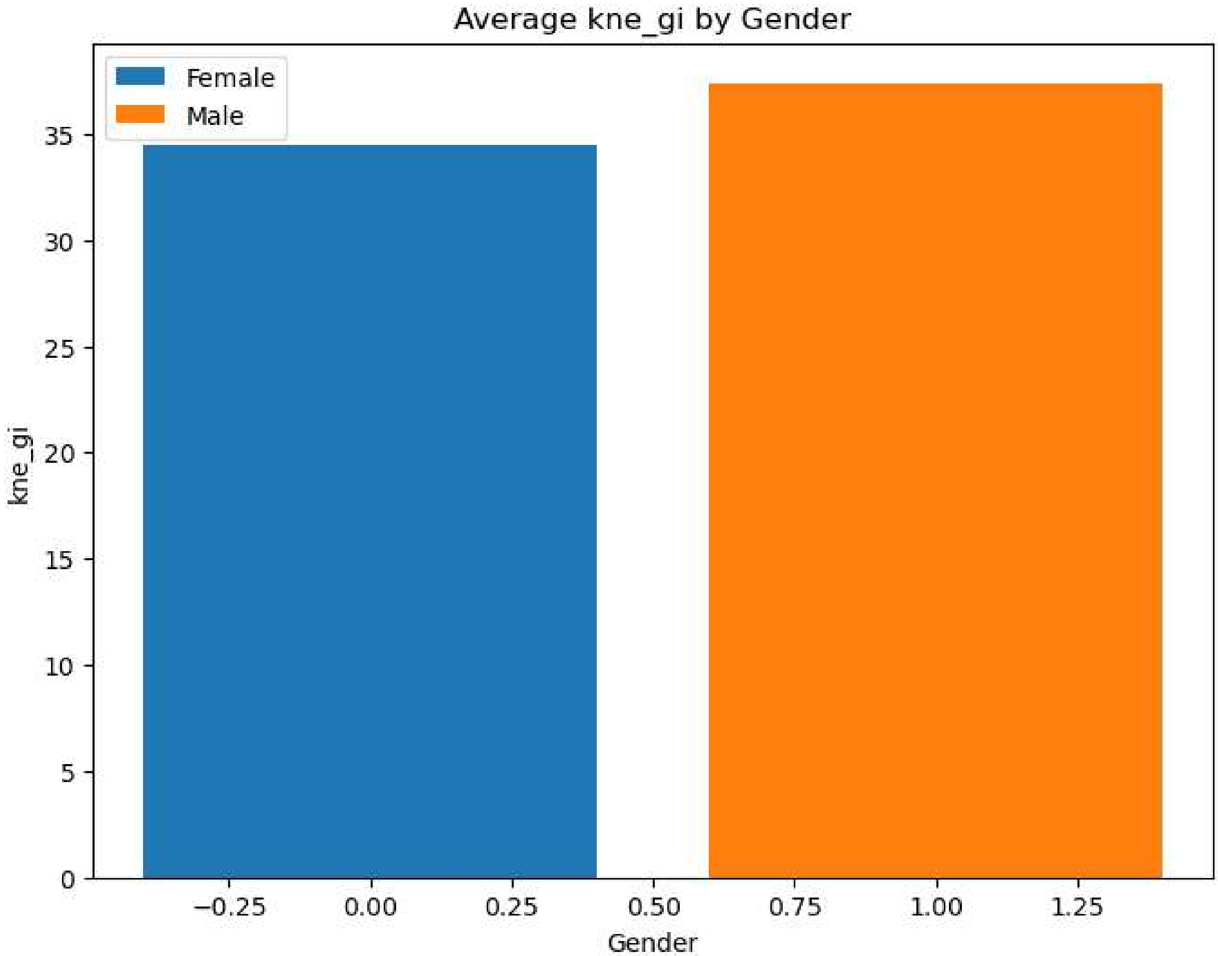

Average nav_gi by Gender
Female
Male
nav_gi
0
20
40
60
80
−0.25
0.00
0.25
0.50
0.75
1.00
1.25
Gender

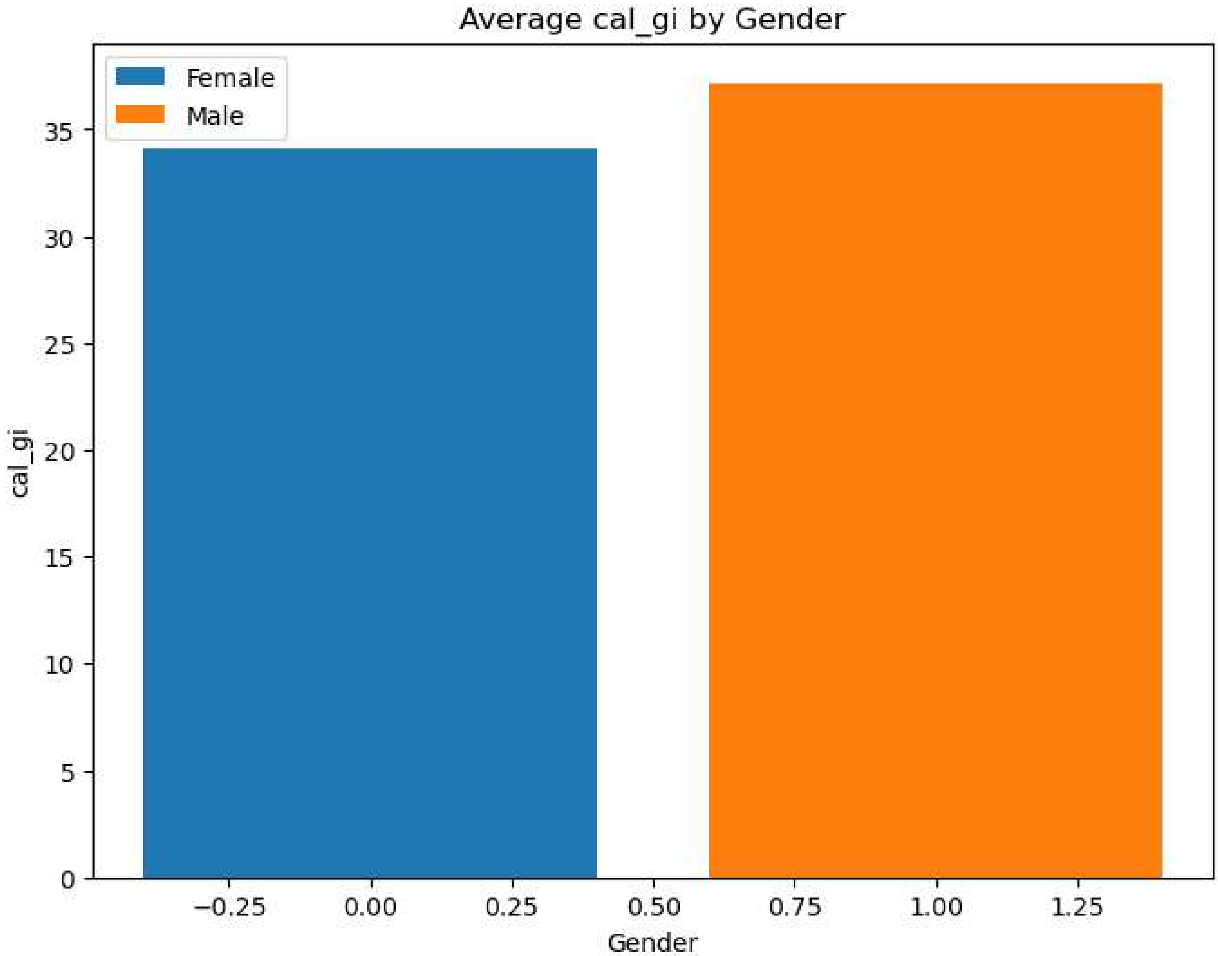

Average hip_gi by Gender
Female
Male
hip_gi
Gender
0
20
40
60
80
100
−0.25
0.00
0.25
0.50
0.75
1.00
1.25

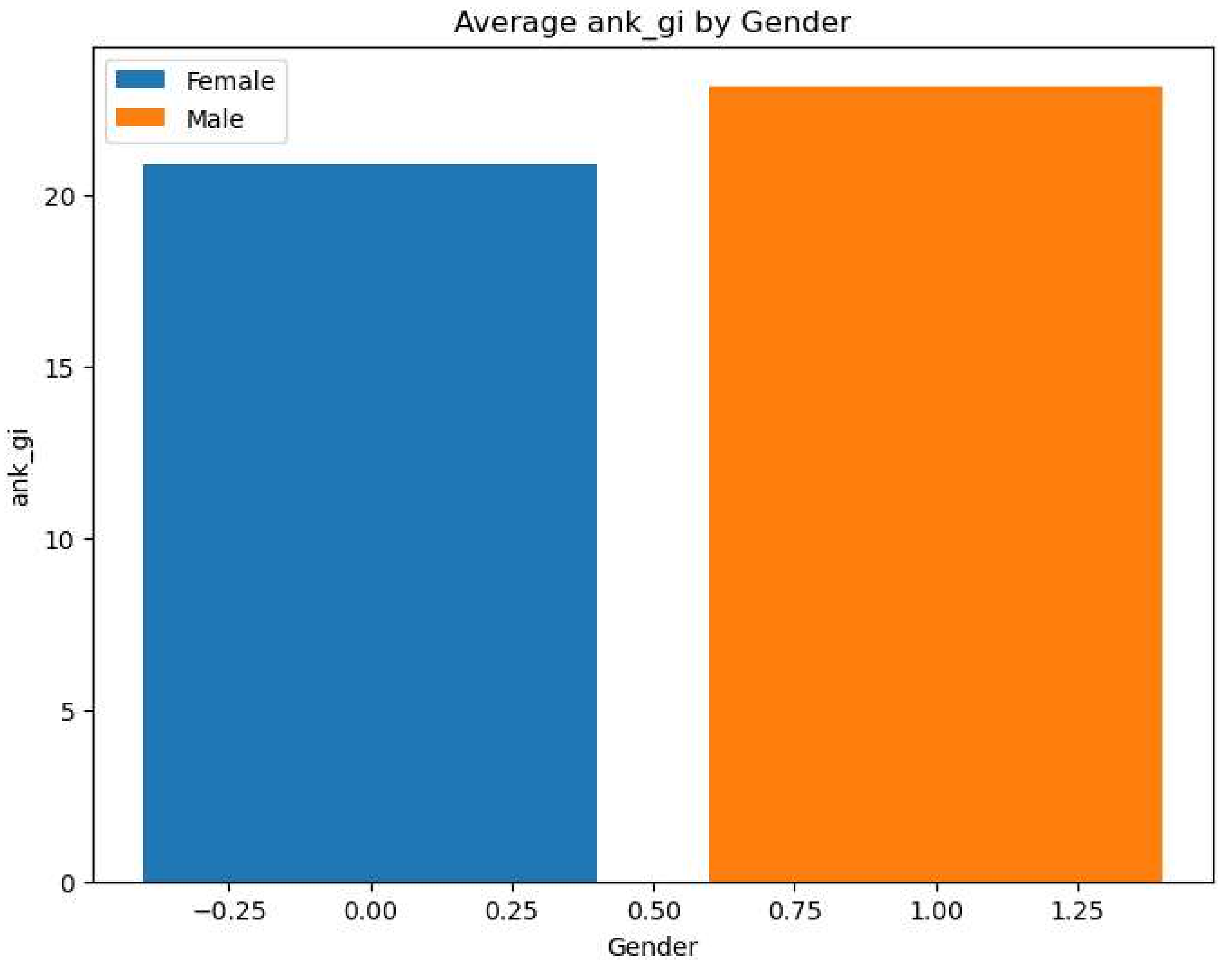

Average thi_gi by Gender
Female
Male
thi_gi
Gender
0
10
20
30
40
50
−0.25
0.00
0.25
0.50
0.75
1.00
1.25

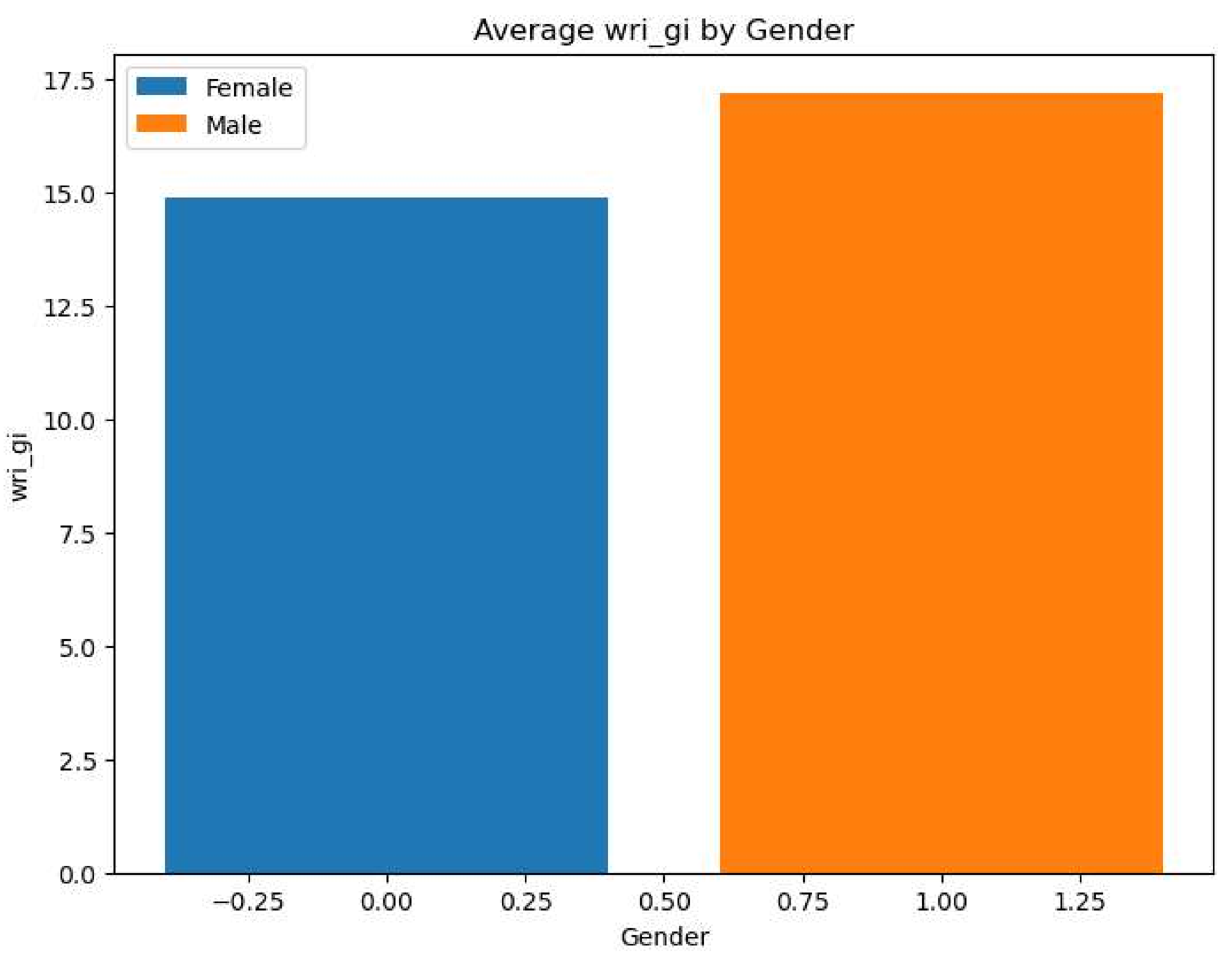

Average bic_gi by Gender
Female
Male
bic_gi
Gender
35
30
25
20
15
10
5
0
−0.25
0.00
0.25
0.50
0.75
1.00
1.25

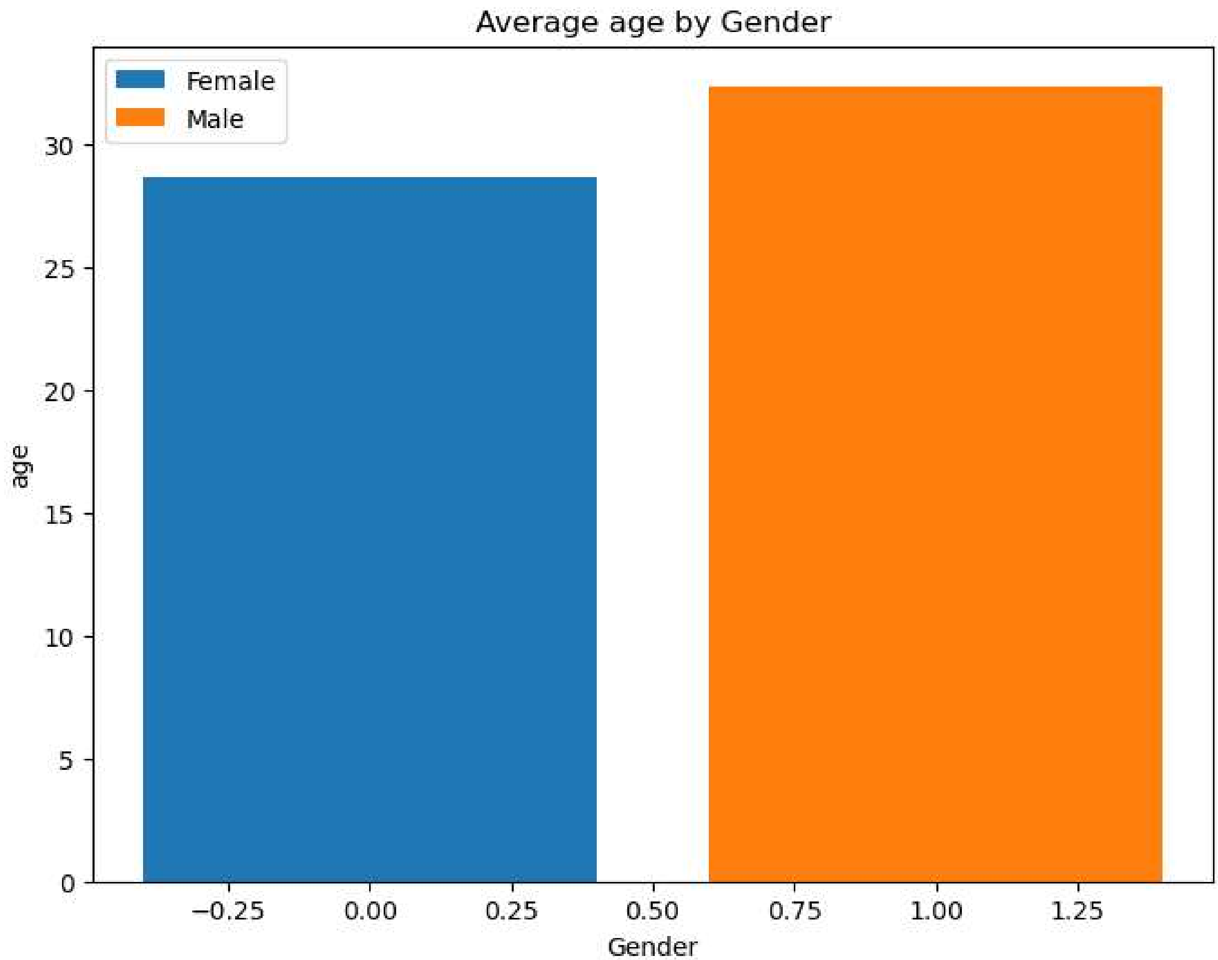

Average for_gi by Gender
Female
Male
for_gi
0
5
10
15
20
25
−0.25
0.00
0.25
0.50
0.75
1.00
1.25
Gender

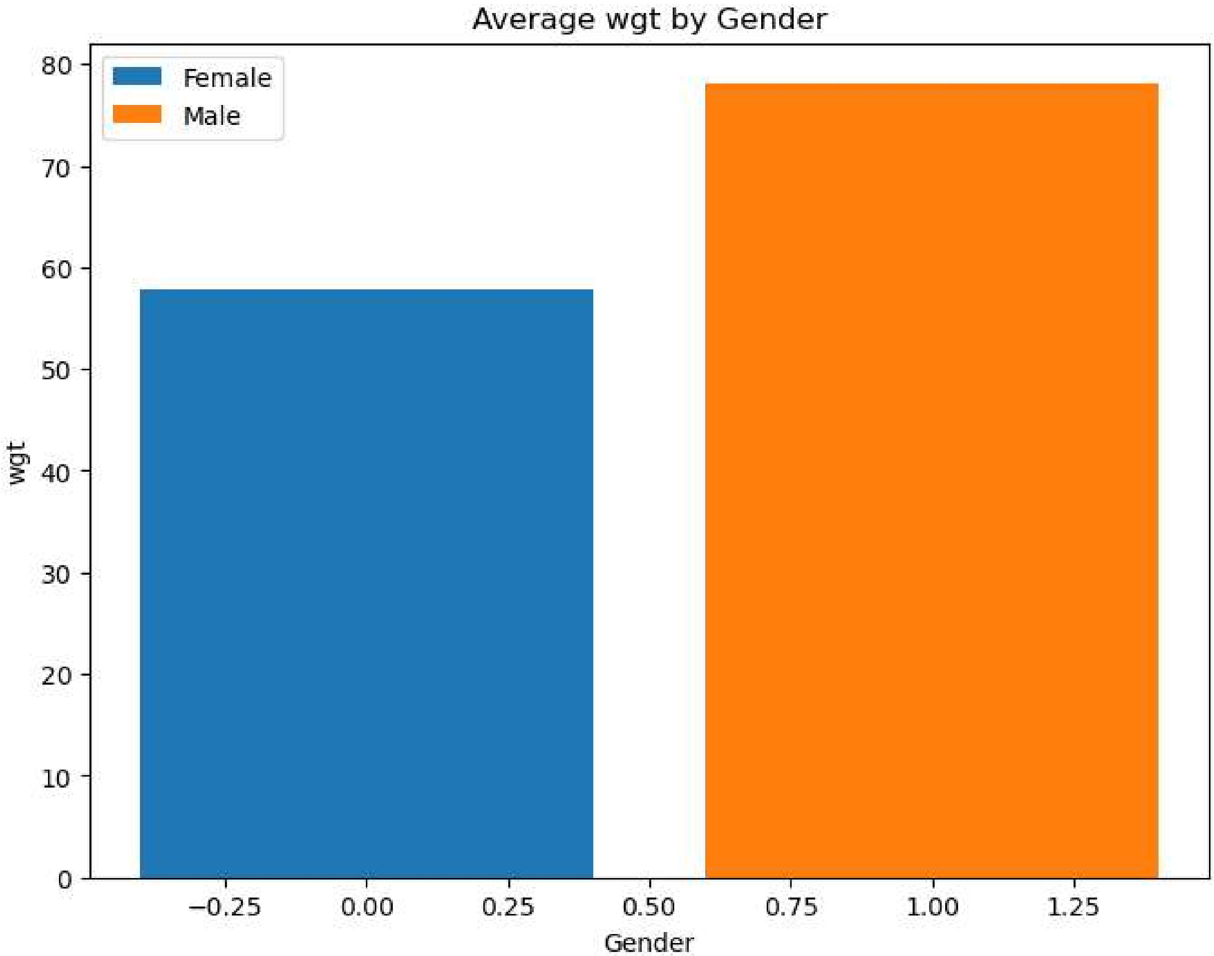

Average kne_gi by Gender
Female
Male
kne_gi
0
5
10
15
20
25
30
35
−0.25
0.00
0.25
0.50
0.75
1.00
1.25
Gender

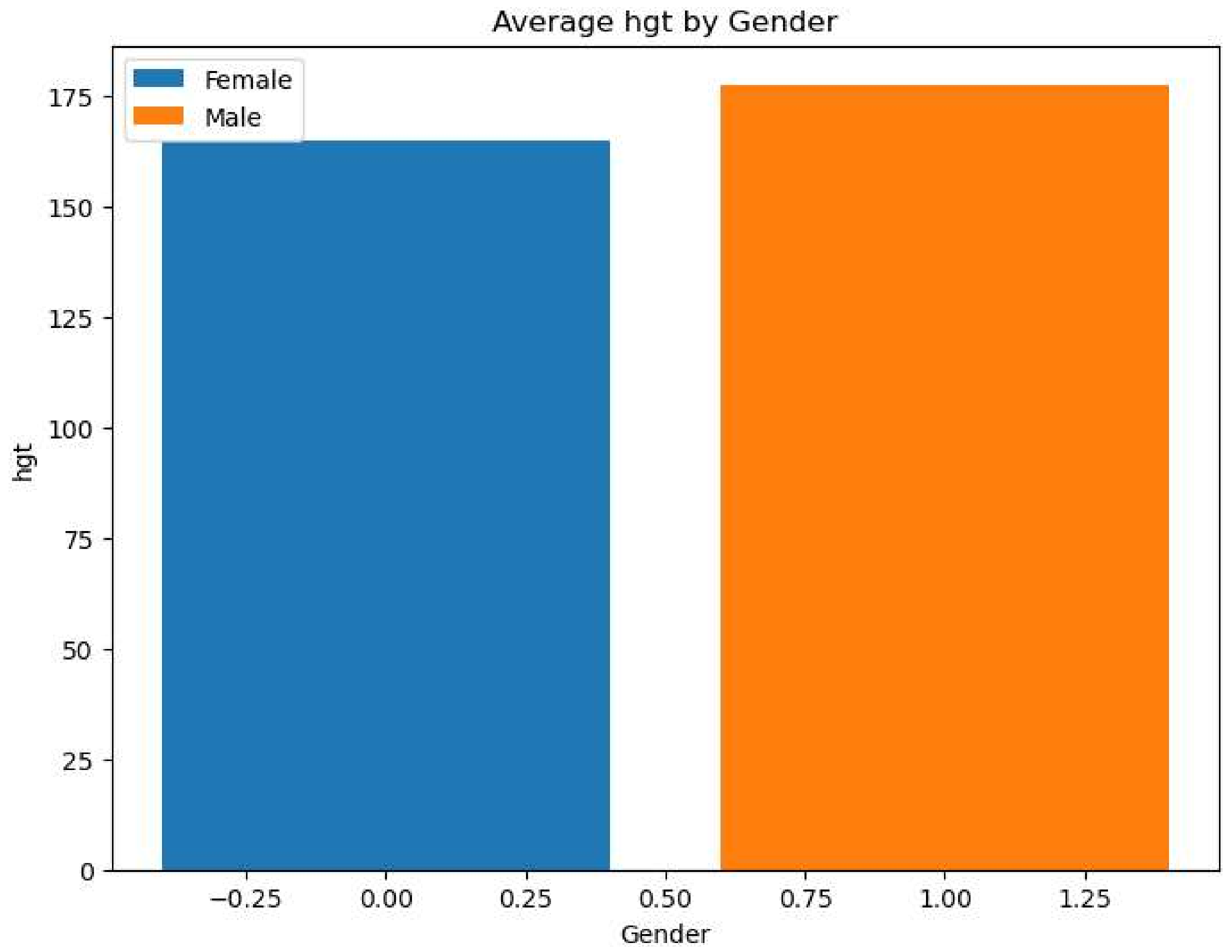

Average cal_gi by Gender
Female
Male
cal_gi
Gender
35
30
25
20
15
10
5
0
−0.25
0.00
0.25
0.50
0.75
1.00
1.25


Average ank_gi by Gender
Female
Male
ank_gi
0
5
10
15
20
−0.25
0.00
0.25
0.50
0.75
1.00
1.25
Gender


Average wri_gi by Gender
Female
Male
wri_gi
0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5
−0.25
0.00
0.25
0.50
0.75
1.00
1.25
Gender


Average age by Gender
Female
Male
age
Gender
30
25
20
15
10
5
0
−0.25
0.00
0.25
0.50
0.75
1.00
1.25


Average wgt by Gender
Female
Male
80
70
60
50
40
30
20
10
0
wgt
−0.25
0.00
0.25
0.50
0.75
1.00
1.25
Gender

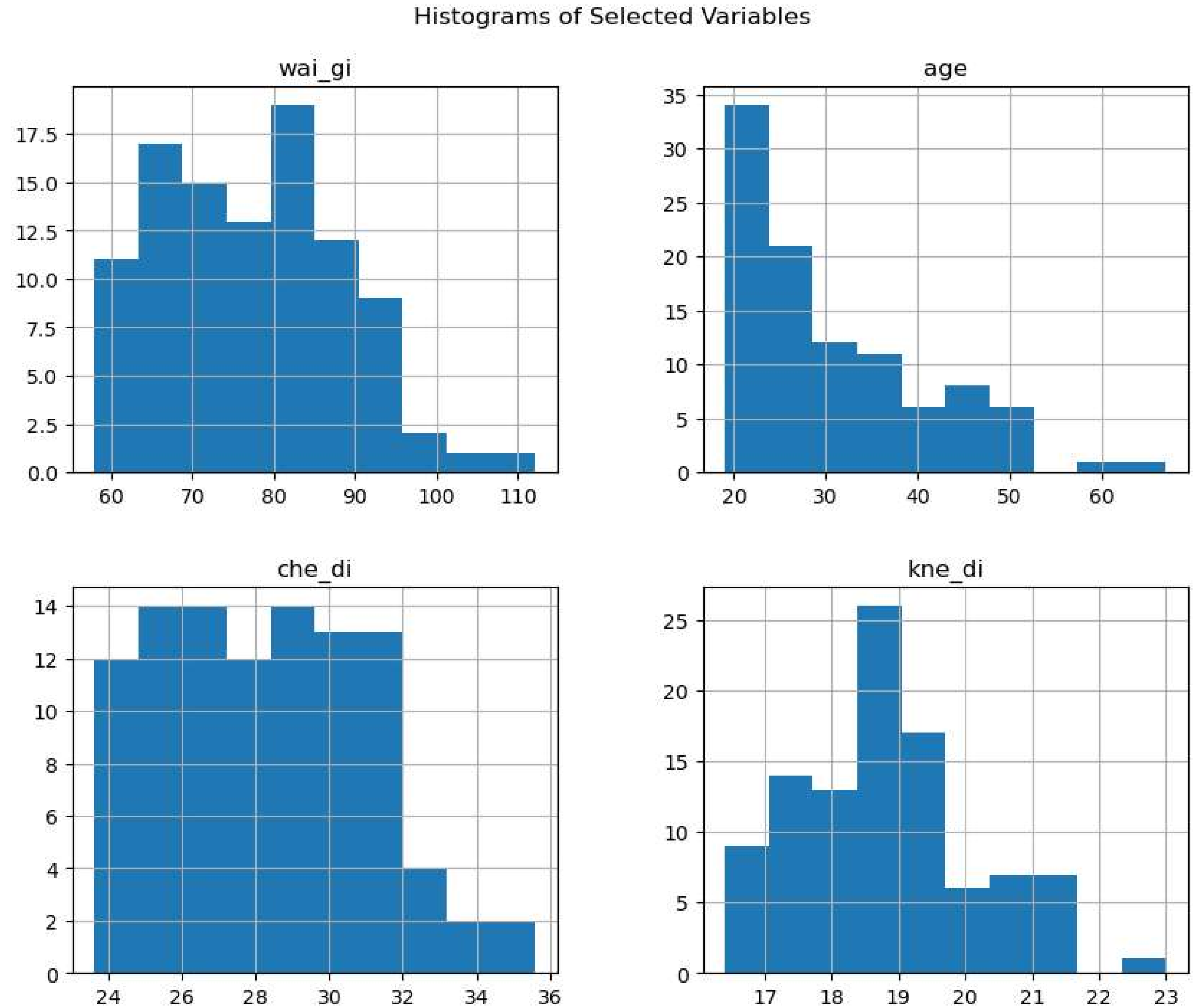


Task 3

In [ ]:
```
import random

column_titles = ["bia_di", "bii_di", "bit_di","che_de","che_di","elb_di","wri_di","kne
# Randomly select 4 titles from the first 22 columns
selected_columns = random.sample(column_titles, 4)

# Print the four variables that were randomly selected
print("Selected Variables:", selected_columns)
```

```
Selected Variables: ['wai_gi', 'age', 'che_di', 'kne_di']
```

In [ ]:
```
import matplotlib.pyplot as plt

# Select the desired columns
selected_df = df[selected_columns]

# Generate histograms
selected_df.hist(figsize=(10, 8))
plt.suptitle('Histograms of Selected Variables', y=0.95)
plt.show()

# Generate descriptive statistics
descriptive_stats = selected_df.describe()
print(descriptive_stats)
```




```
        wai_gi        age     che_di     kne_di
count  100.00000  100.000000  100.000000  100.000000
mean    77.41300   30.650000   28.262000   18.789000
std     11.32288   10.287165    2.758706    1.349814
min     57.90000   19.000000   23.600000   16.400000
25%     67.87500   22.000000   25.900000   17.875000
50%     77.00000   27.000000   28.300000   18.800000
75%     85.17500   36.000000   30.225000   19.525000
max    112.10000   67.000000   35.600000   23.000000
```

1. **Waist Girth (`wai_gi`)**:

   - Center: The mean waist girth is approximately 77.41 centimeters.
   - Spread: The standard deviation is approximately 11.32 centimeters, indicating variability in waist girth measurements.
   - Shape: The distribution appears to be relatively symmetric, as the mean and median are close in value.
   - Outliers: There might be potential outliers on the higher end, as the maximum waist girth is 112.10 centimeters.

2. **Age**:

   - Center: The mean age is approximately 30.65 years.

- Spread: The standard deviation is approximately 10.29 years, indicating variability in ages.
- Shape: The distribution may have a slightly right-skewed shape, as the mean is slightly higher than the median.
- Outliers: There might be potential outliers on the higher end, as the maximum age is 67 years.

3. **Chest Diameter (`che_di`)**:

   - Center: The mean chest diameter is approximately 28.26 centimeters.
   - Spread: The standard deviation is approximately 2.76 centimeters, indicating relatively low variability in chest diameter measurements.
   - Shape: The distribution appears to be relatively symmetric, as the mean and median are close in value.
   - Outliers: No outliers are evident based on the provided statistics.

4. **Knee Diameter (`kne_di`)**:

   - Center: The mean knee diameter is approximately 18.79 centimeters.
   - Spread: The standard deviation is approximately 1.35 centimeters, indicating relatively low variability in knee diameter measurements.
   - Shape: The distribution appears to be relatively symmetric, as the mean and median are close in value.
   - Outliers: No outliers are evident based on the provided statistics.

Task 4.

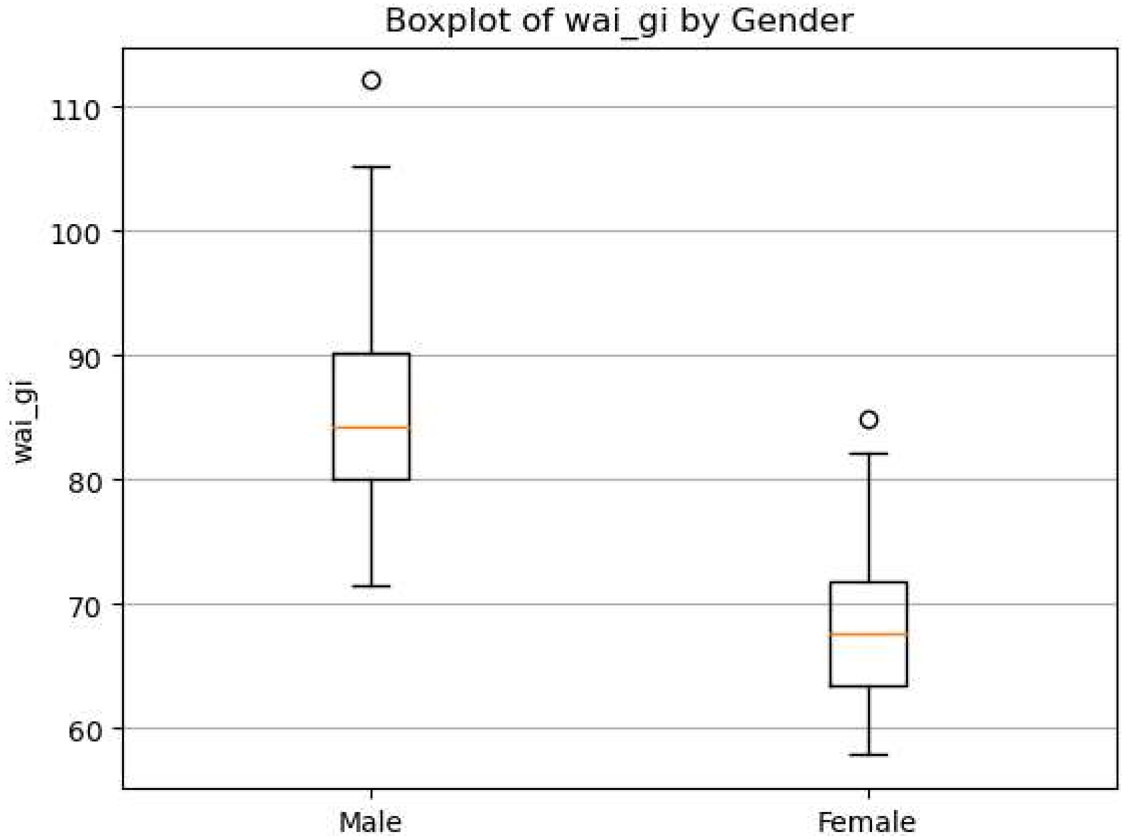
Now that you have described and plotted data, let's explore if the data differ for male and female participants.

- Generate grouped box plots for each of the 4 variables in Task 3.
- Your boxplot should compare the distributions for males and females in your dataset.
- Afterwards, you should describe what you observe in each case.
- Make sure you mention the five-number summaries for both genders.

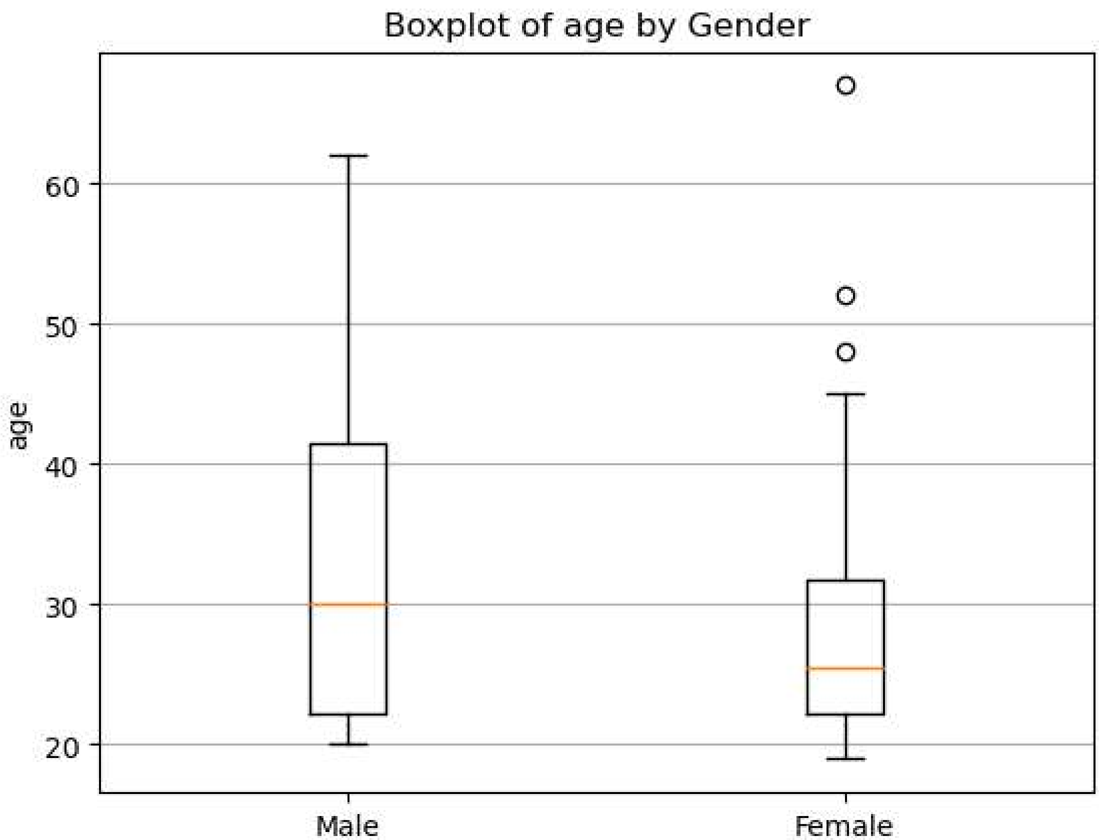
In [ ]:
```
import matplotlib.pyplot as plt

# Filter data for males and females
male = df[df["sex"] == 1]
female = df[df["sex"] == 0]


# Loop through each selected column
for column_name in selected_columns:
    # Create boxplots side by side
    data1 = male[column_name]
    data2 = female[column_name]
    data = [data1, data2]

    fig, ax = plt.subplots()
```

```
    ax.set_xticklabels(['Male', 'Female'])
    plt.grid(axis="y")
    plt.boxplot(data)
    plt.title(f"Boxplot of {column_name} by Gender")
    plt.ylabel(column_name)

    # Descriptive statistics
    print(f"Descriptive Statistics for {column_name}")
    print("Male:")
    print(male[column_name].describe())
    print("\nFemale:")
    print(female[column_name].describe())

    plt.show()
```

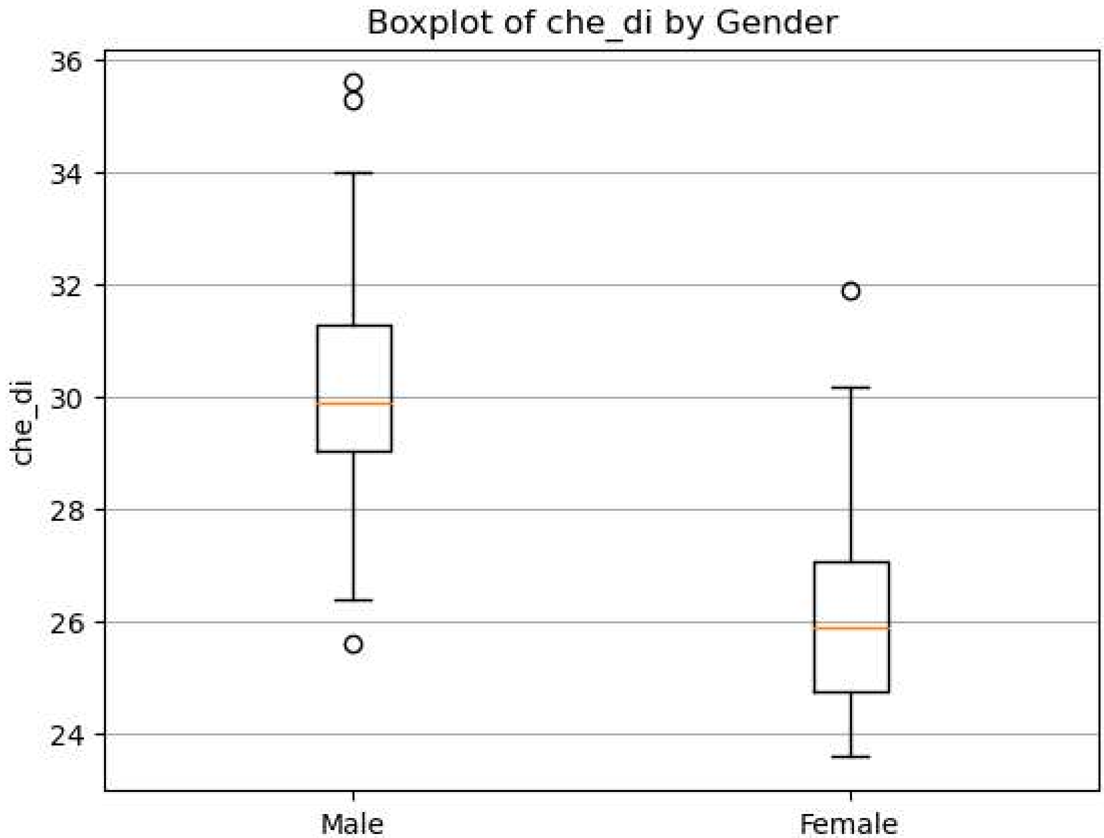
```
C:\Users\ROBERT\AppData\Local\Temp\ipykernel_15832\2024948347.py:16: UserWarning: Fix
edFormatter should only be used together with FixedLocator
  ax.set_xticklabels(['Male', 'Female'])
Descriptive Statistics for wai_gi
Male:
count     54.000000
mean      85.300000
std        8.245754
min       71.500000
25%       80.050000
50%       84.200000
75%       90.225000
max      112.100000
Name: wai_gi, dtype: float64

Female:
count    46.000000
mean     68.154348
std       6.301065
min      57.900000
25%      63.500000
50%      67.600000
75%      71.925000
max      84.900000
Name: wai_gi, dtype: float64
```



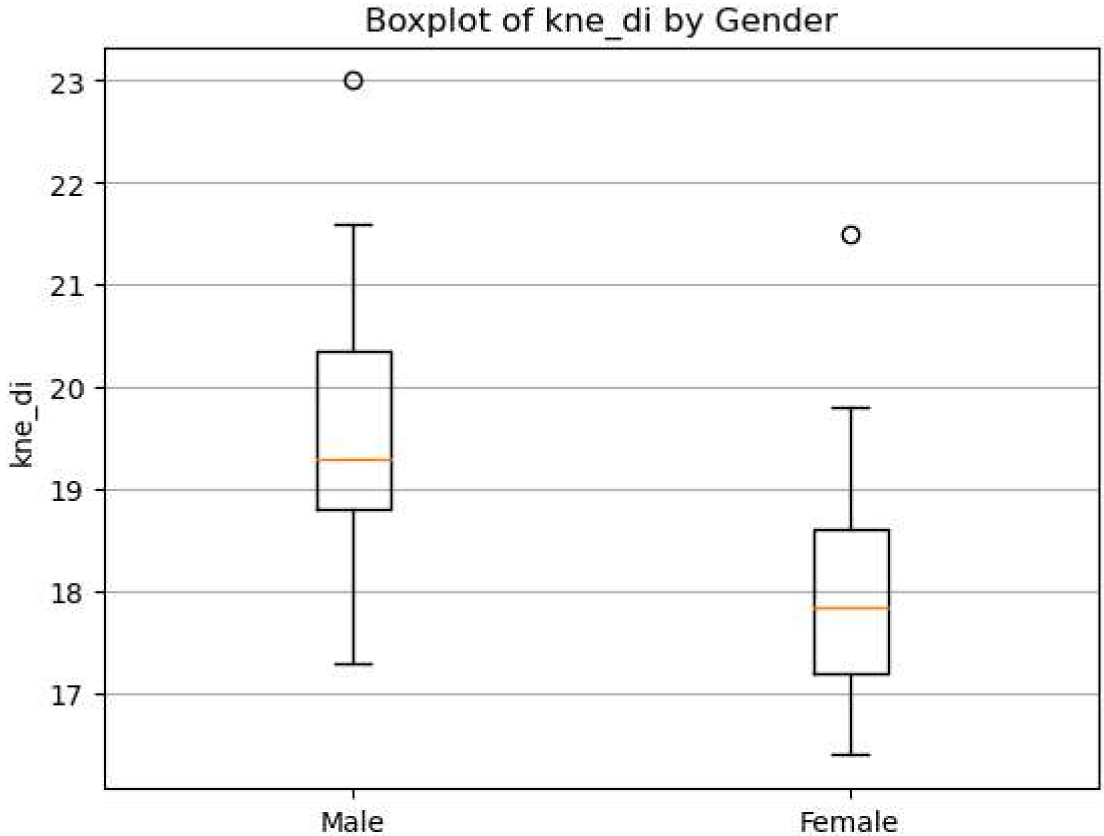

```
C:\Users\ROBERT\AppData\Local\Temp\ipykernel_15832\2024948347.py:16: UserWarning: Fix
edFormatter should only be used together with FixedLocator
  ax.set_xticklabels(['Male', 'Female'])
Descriptive Statistics for age
Male:
count    54.000000
mean     32.351852
std      10.410730
min      20.000000
25%      22.250000
50%      30.000000
75%      41.500000
max      62.000000
Name: age, dtype: float64

Female:
count    46.000000
mean     28.652174
std       9.879760
min      19.000000
25%      22.250000
50%      25.500000
75%      31.750000
max      67.000000
Name: age, dtype: float64
```




```
C:\Users\ROBERT\AppData\Local\Temp\ipykernel_15832\2024948347.py:16: UserWarning: Fix
edFormatter should only be used together with FixedLocator
  ax.set_xticklabels(['Male', 'Female'])
Descriptive Statistics for che_di
Male:
count    54.000000
mean     30.114815
std       1.992666
min      25.600000
25%      29.050000
50%      29.900000
75%      31.300000
max      35.600000
Name: che_di, dtype: float64

Female:
count    46.000000
mean     26.086957
std       1.763785
min      23.600000
25%      24.775000
50%      25.900000
75%      27.075000
max      31.900000
Name: che_di, dtype: float64
```




```
C:\Users\ROBERT\AppData\Local\Temp\ipykernel_15832\2024948347.py:16: UserWarning: Fix
edFormatter should only be used together with FixedLocator
  ax.set_xticklabels(['Male', 'Female'])
Descriptive Statistics for kne_di
Male:
count    54.000000
mean     19.529630
std       1.135969
min      17.300000
25%      18.800000
50%      19.300000
75%      20.350000
max      23.000000
Name: kne_di, dtype: float64

Female:
count    46.000000
mean     17.919565
std       1.028401
min      16.400000
25%      17.200000
50%      17.850000
75%      18.600000
max      21.500000
Name: kne_di, dtype: float64
```



Waist Girth (`wai_gi`):

The descriptive statistics for waist girth unveil intriguing gender disparities. Among males, the median waist girth stands at 84.20 centimeters, with measurements ranging from 71.50 to 112.10 centimeters. Contrastingly, females exhibit a notably lower median of 67.60 centimeters, spanning from 57.90 to 84.90 centimeters. This discrepancy underscores the distinct distribution of waist girths between genders, with males generally presenting broader waist measurements compared to females.

Age:

Analyzing age distributions reveals noteworthy differences between males and females. Among males, the median age is 30.00 years, with ages ranging from 20.00 to 62.00 years. In contrast, females exhibit a slightly lower median age of 25.50 years, with ages spanning from 19.00 to 67.00 years. These statistics shed light on the varying age compositions within the dataset, showcasing a broader age range among males compared to females.

Chest Diameter (`che_di`):

Examining chest diameter distributions illuminates distinctive patterns between genders. Among males, the median chest diameter is 29.90 centimeters, ranging from 25.60 to 35.60 centimeters. Conversely, females present a lower median of 25.90 centimeters, with chest diameters spanning from 23.60 to 31.90 centimeters. This divergence in chest measurements underscores the inherent anatomical differences between males and females, with males typically exhibiting broader chest dimensions than females.

Knee Diameter (`kne_di`):

Analysis of knee diameter distributions unveils noteworthy variations between genders. Among males, the median knee diameter is 19.30 centimeters, ranging from 17.30 to 23.00 centimeters. In contrast, females display a lower median of 17.85 centimeters, with knee diameters spanning from 16.40 to 21.50 centimeters. This disparity highlights the distinct anatomical characteristics between males and females, with males generally presenting wider knee measurements compared to females.

Task 5

Part A

- Select **TWO** variables from Task 3. Treat these as an independent variable.
- Now create a scatterplot for each variable.
- In each case, the plot should visualize the relationship between the variable and **weight** (dependent variable).
- Describe each scatterplot in terms of the **form**, **strength**, and **direction** of the relationship between the variables.

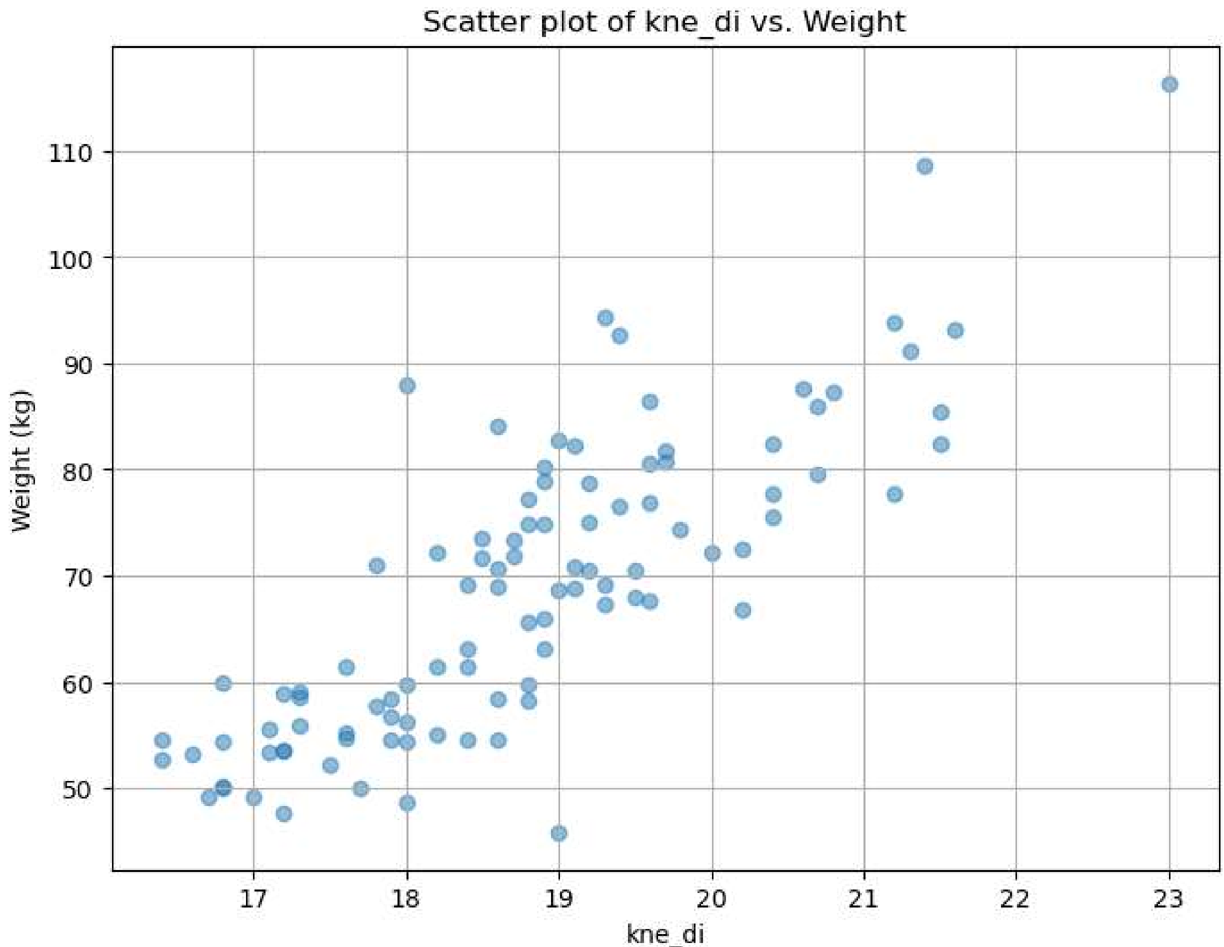
Part B

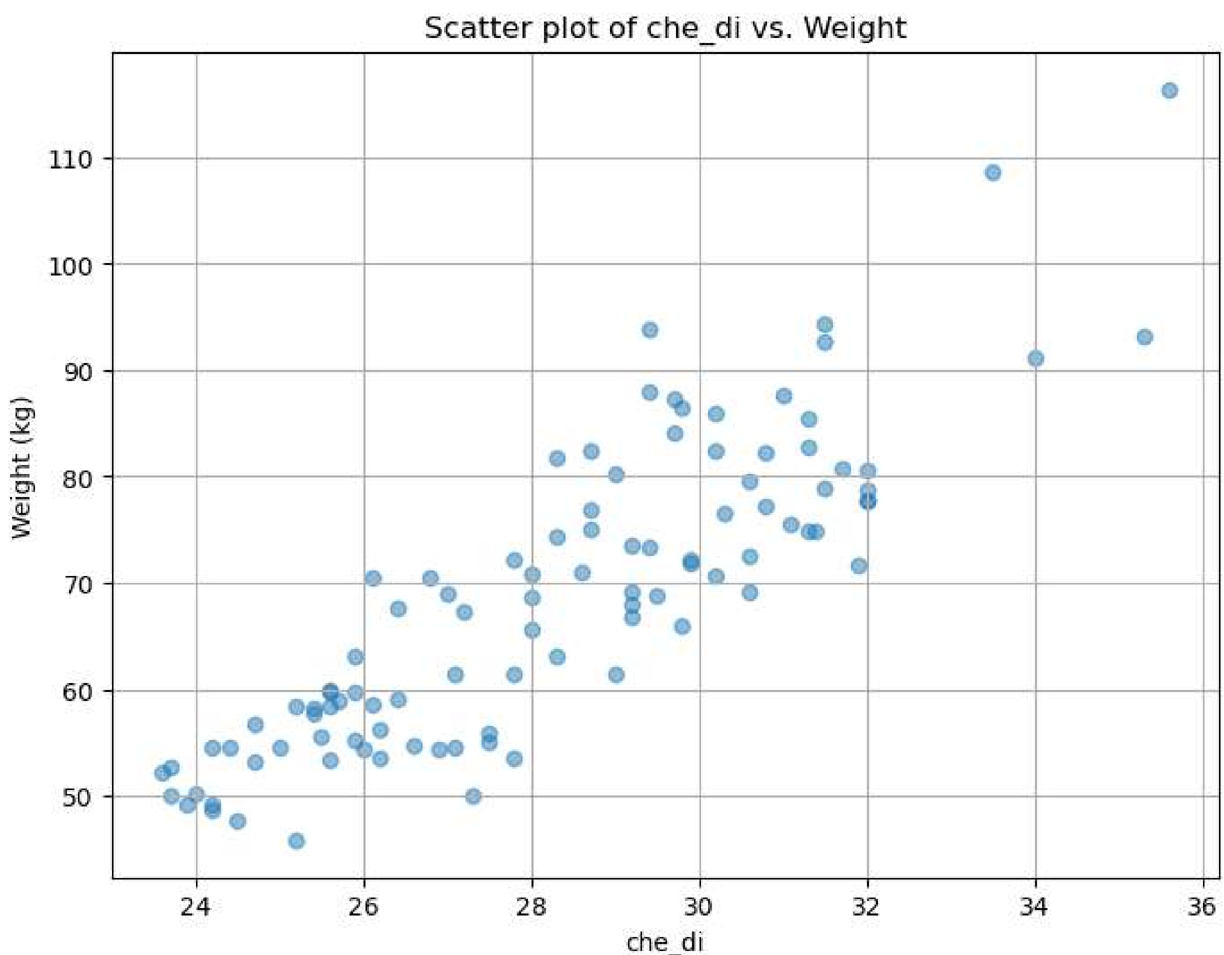
- Examine if the relationship explored in each scatterplot varies **by gender**.
- Hint: You will need to create scatterplots separately for each gender to answer this question.

In [ ]:

```
import random

# Randomly select two columns
selected_columns_random = random.sample(selected_columns, 2)
print("Randomly selected columns:", selected_columns_random)

import matplotlib.pyplot as plt

# Create scatter plots
for column_name in selected_columns_random:
    plt.figure(figsize=(8, 6))
    plt.scatter(df[column_name], df['wgt'], alpha=0.5)
    plt.title(f'Scatter plot of {column_name} vs. Weight')
    plt.xlabel(column_name)
    plt.ylabel('Weight (kg)')
    plt.grid(True)
    plt.show()
```

```
Randomly selected columns: ['kne_di', 'che_di']
```


Scatter plot of kne_di vs. Weight
Weight (kg)
50
60
70
80
90
100
110
17
18
19
20
21
22
23
kne_di


Scatter plot of che_di vs. Weight
Weight (kg)
50
60
70
80
90
100
110
24
26
28
30
32
34
36
che_di

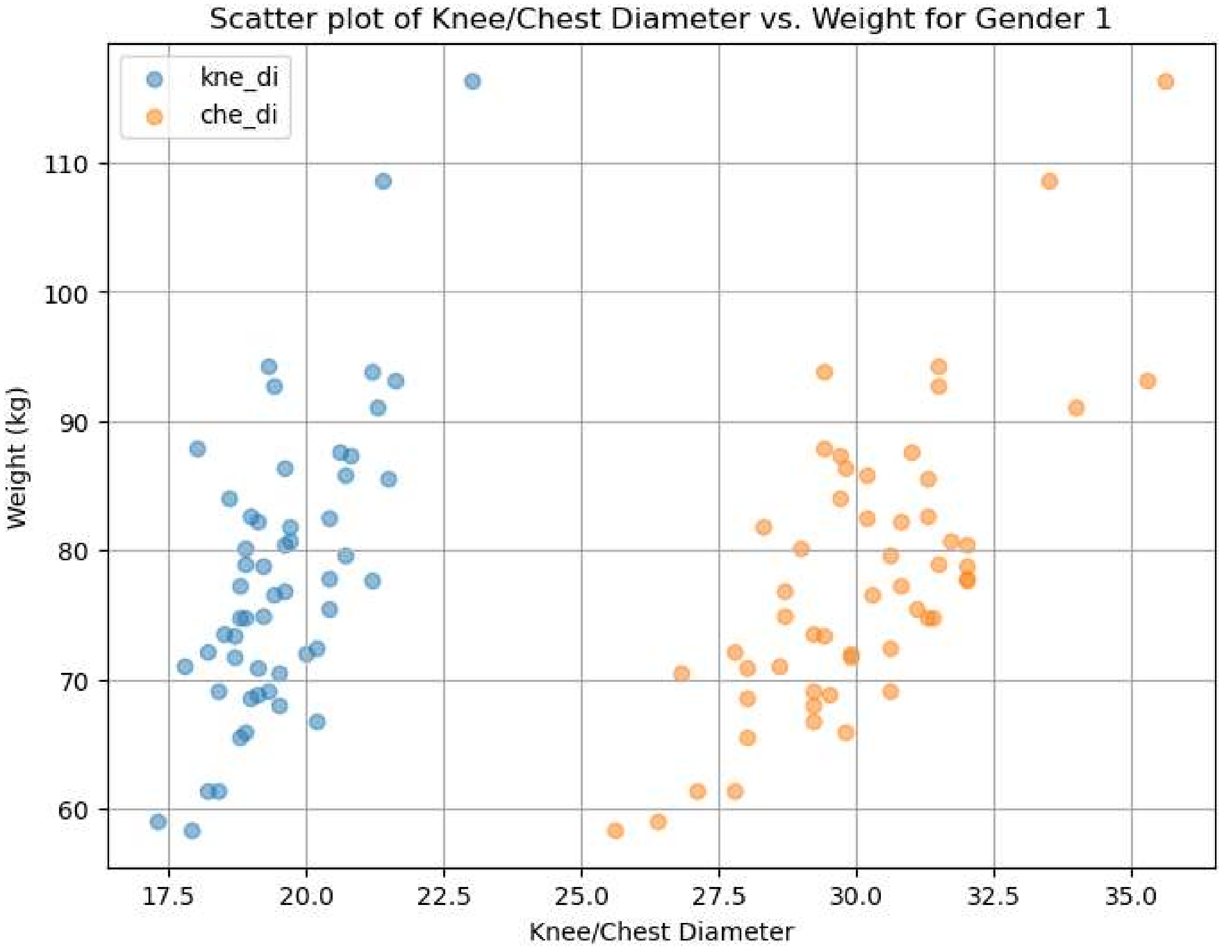
Scatterplot of Knee Diameter vs. Weight:

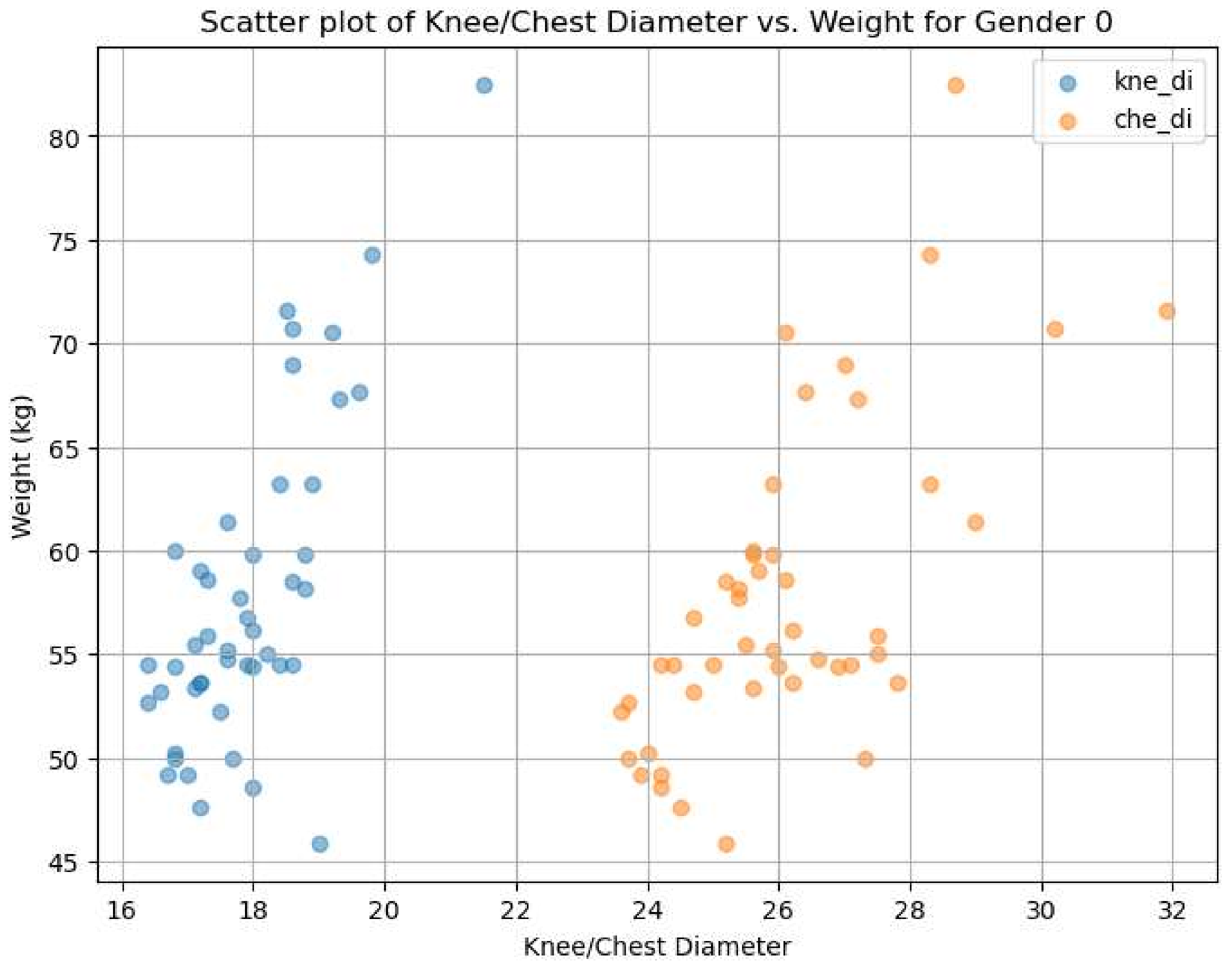
- **Form**: The relationship between knee diameter and weight appears to be approximately linear, with some scattered points.
- **Strength**: The strength of the relationship seems to be moderate, as there is noticeable variation in weight for a given knee diameter.
- **Direction**: The direction of the relationship is positive, indicating that as knee diameter increases, weight tends to increase as well. However, the relationship is not perfectly linear, suggesting some variability in weight even for similar knee diameters.

Scatterplot of Chest Diameter vs. Weight:

- **Form**: The scatterplot displays a roughly linear pattern, with points distributed along a line.
- **Strength**: The relationship appears to be moderately strong, as there is relatively little variability in weight for a given chest diameter.
- **Direction**: The direction of the relationship is positive, indicating that as chest diameter increases, weight tends to increase as well. The data points are more tightly clustered around the regression line compared to the knee diameter scatterplot, suggesting a stronger association between chest diameter and weight.

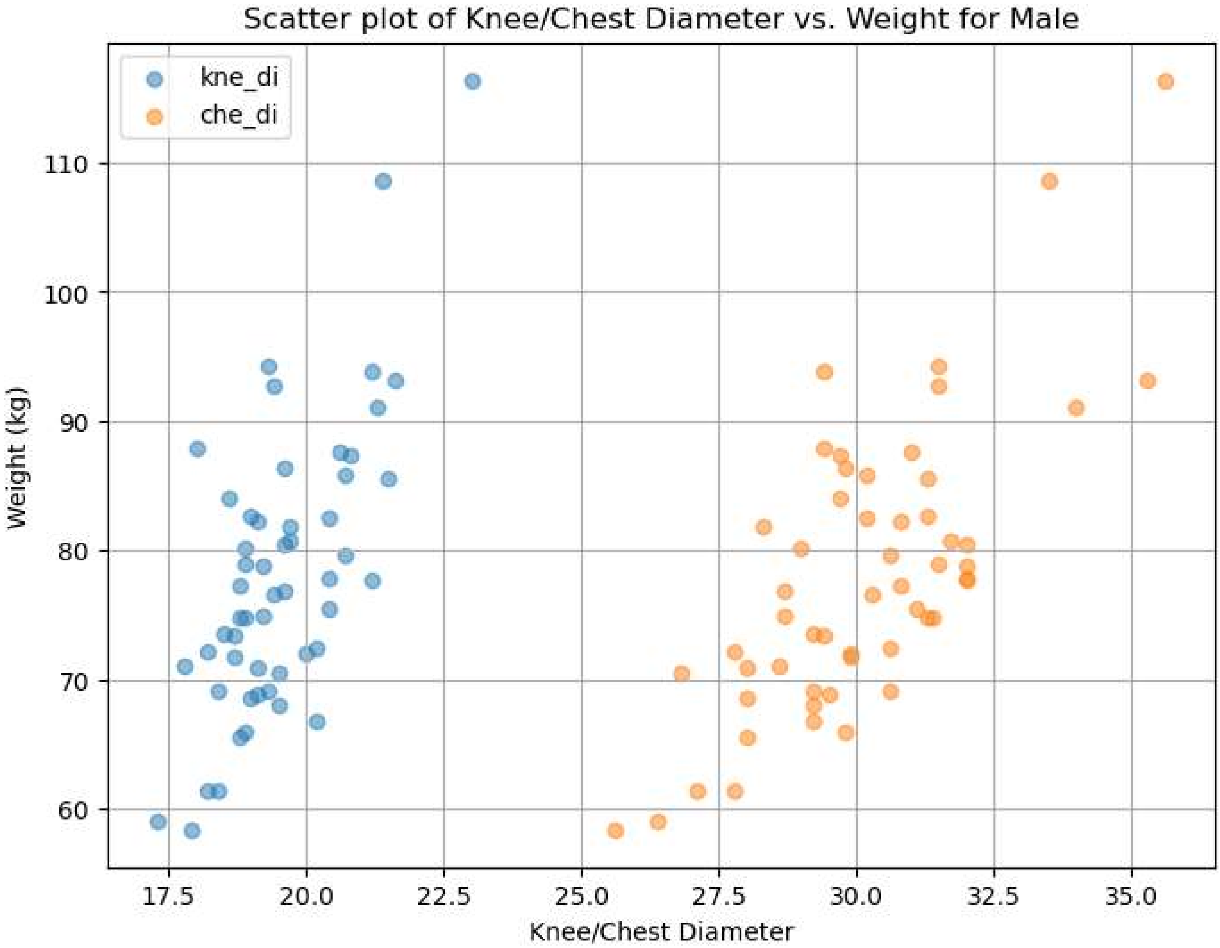
In [ ]:
```
import matplotlib.pyplot as plt


# Create scatter plots for each gender
for gender in df['sex'].unique():
    plt.figure(figsize=(8, 6))
    gender_df = df[df['sex'] == gender]
    for column_name in selected_columns_random:
        plt.scatter(gender_df[column_name], gender_df['wgt'], alpha=0.5, label=column_
    plt.title(f'Scatter plot of Knee/Chest Diameter vs. Weight for Gender {gender}')
    plt.xlabel('Knee/Chest Diameter')
    plt.ylabel('Weight (kg)')
    plt.legend()
    plt.grid(True)
    plt.show()
```

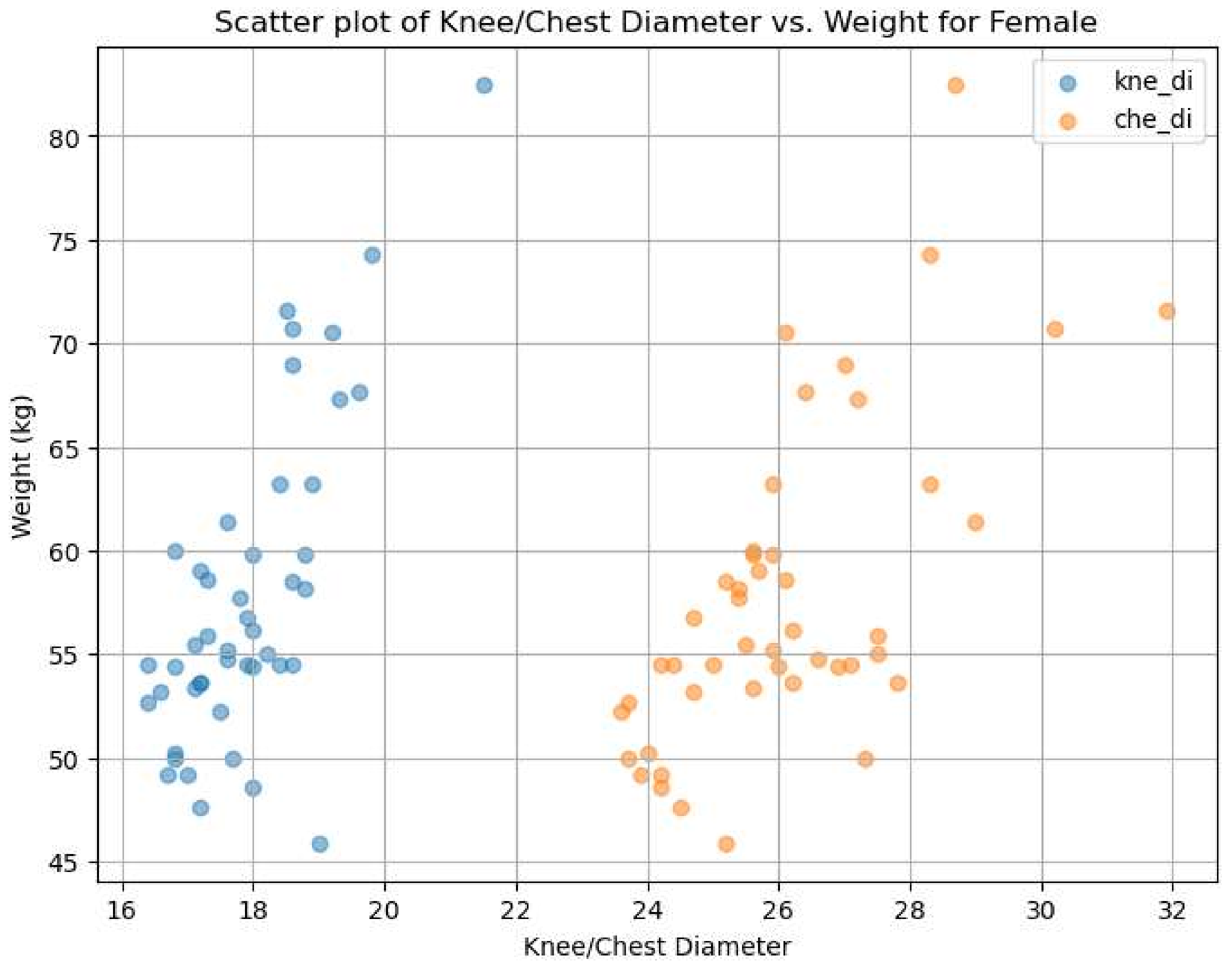

Scatter plot of Knee/Chest Diameter vs. Weight for Gender 1
kne_di
che_di
Weight (kg)
Knee/Chest Diameter
110
100
90
80
70
60
17.5
20.0
22.5
25.0
27.5
30.0
32.5
35.0



In [ ]:

```
import matplotlib.pyplot as plt

# Filter data for males and females
male = df[df["sex"] == 1]
female = df[df["sex"] == 0]

# Create scatter plots for each gender
for gender_df, gender_label in zip([male, female], ['Male', 'Female']):
    plt.figure(figsize=(8, 6))
    for column_name in selected_columns_random:
        plt.scatter(gender_df[column_name], gender_df['wgt'], alpha=0.5, label=column_
    plt.title(f'Scatter plot of Knee/Chest Diameter vs. Weight for {gender_label}')
    plt.xlabel('Knee/Chest Diameter')
    plt.ylabel('Weight (kg)')
    plt.legend()
    plt.grid(True)
    plt.show()
```


Scatter plot of Knee/Chest Diameter vs. Weight for Male
kne_di
che_di
Weight (kg)
60
70
80
90
100
110
17.5
20.0
22.5
25.0
27.5
30.0
32.5
35.0
Knee/Chest Diameter



the relationships od not vary by gender

Task 6.

PART A

Finally, for each of the variables you focused on in Task 5:

- Fit a simple linear regression model that predicts a participant's **Weight** based on the variable you selected.
- Make sure you generate, interpret, and use the residual plot, the standard error, and the R^2 to assess the fit of each linear model.
- If the model is a good fit, interpret the slope and the y-intercept.

PART B

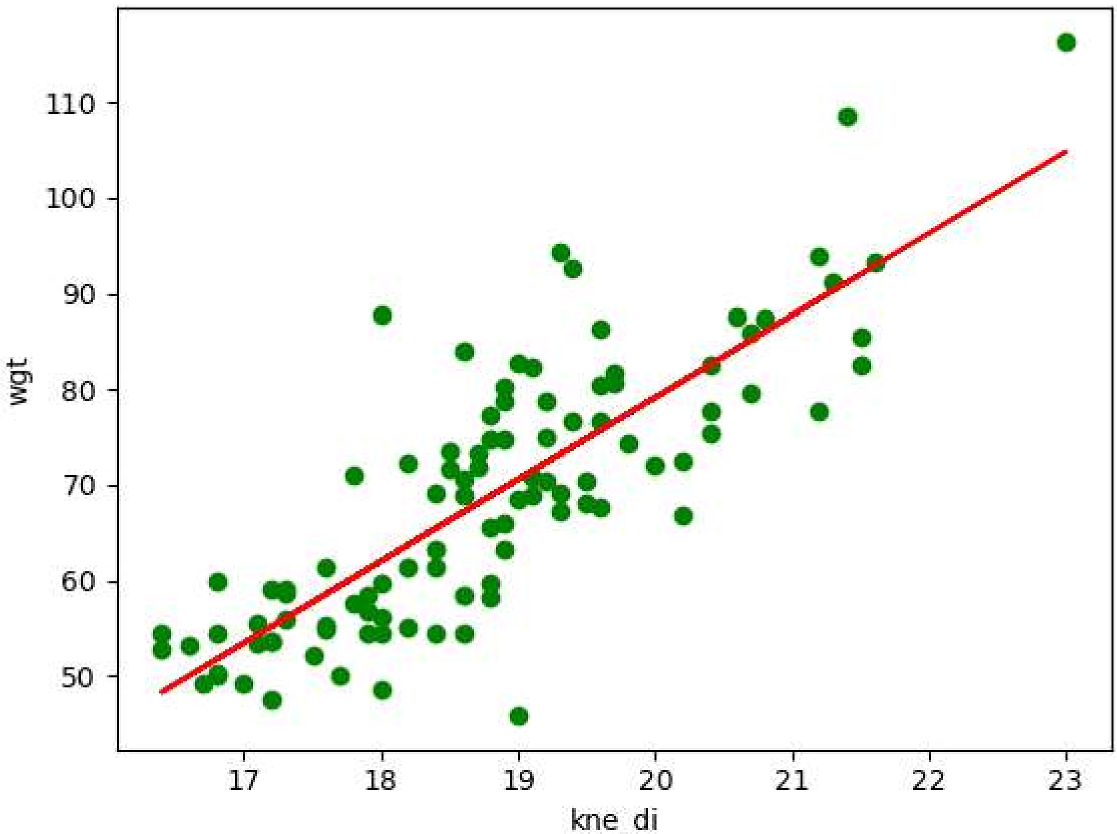
If you found that the relationship between weight and the variable you selected differed for males and females in Task 5 (Part B) then:

- Run the regression model for each gender separately and interpret your findings accordingly.

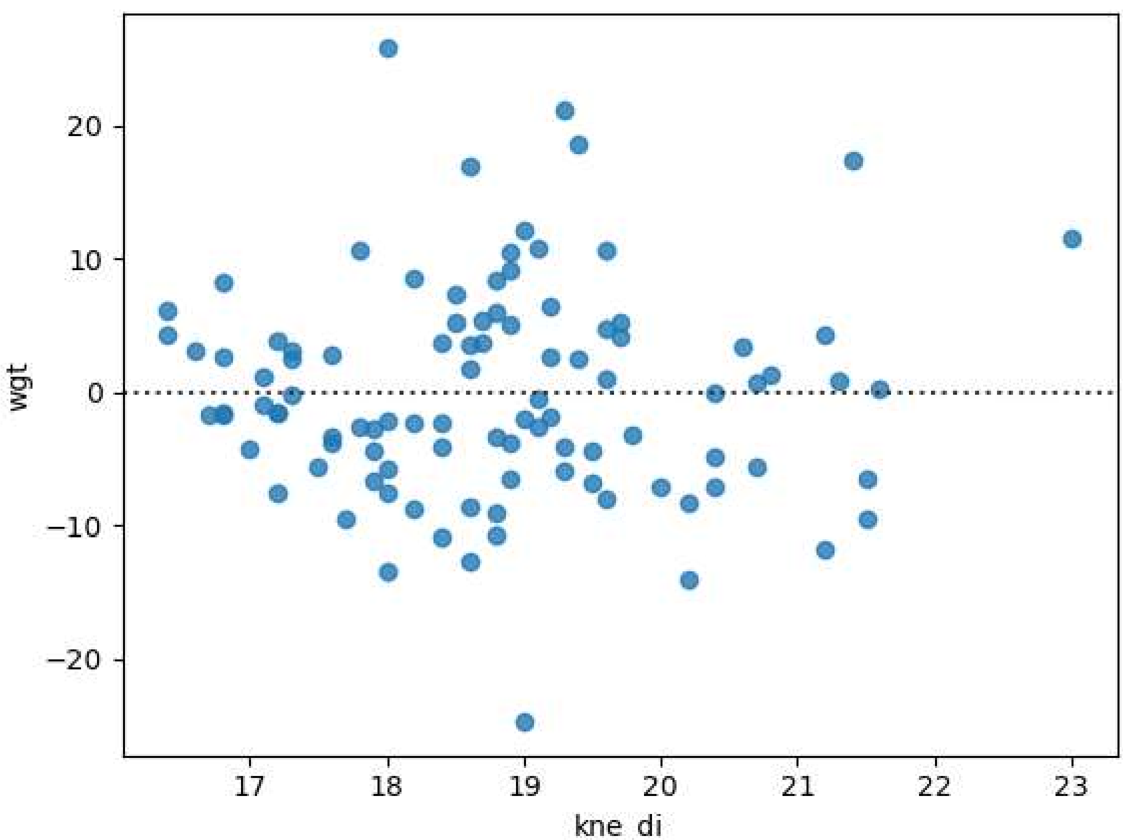
In [ ]:
```
import numpy as np
import matplotlib.pyplot as plt
```

```
import statsmodels.api as sm
import seaborn as sns

def regression_equation(column_x, column_y):
    # Fit the regression line using "statsmodels" library:
    X = df[column_x]
    X = sm.add_constant(X)
    Y = df[column_y]
    regression_model = sm.OLS(Y, X).fit()  # OLS stands for "ordinary least squares"

    # Calculate R^2
    R_squared = round(regression_model.rsquared, 3)
    print('R2: ', R_squared)

    # Calculate standard error
    SE = np.sqrt(regression_model.mse_resid)
    print('SE: ', round(SE, 3))

    # Calculate correlation coefficient
    correlation_coefficient = np.corrcoef(df[column_x], df[column_y])[0, 1]
    print('Correlation coefficient: ', round(correlation_coefficient, 3))

    # Extract regression parameters and print the regression equation
    slope = round(regression_model.params[1], 3)
    intercept = round(regression_model.params[0], 3)
    print("Regression equation: " + column_y + " = ", slope, "* " + column_x + " + ",

    # Display the scatter plot with the line of best fit
    plt.scatter(df[column_x], Y, color='green')
    plt.xlabel(column_x)
    plt.ylabel(column_y)
    plt.plot(df[column_x], regression_model.params[1]*df[column_x]+regression_model.pa
    plt.show()

    # Display the residual plot
    sns.residplot(x=column_x, y=column_y, data=df)
    plt.show()

    # Display the residual plot with SE
    sns.residplot(x=column_x, y=column_y, data=df)
    plt.axhline(y=SE, color='r', linestyle='--')
    plt.axhline(y=-SE, color='r', linestyle='--')
    plt.show()

# Call the function with the selected variables
regression_equation(selected_columns_random[0], "wgt")
regression_equation(selected_columns_random[1], "wgt")
```

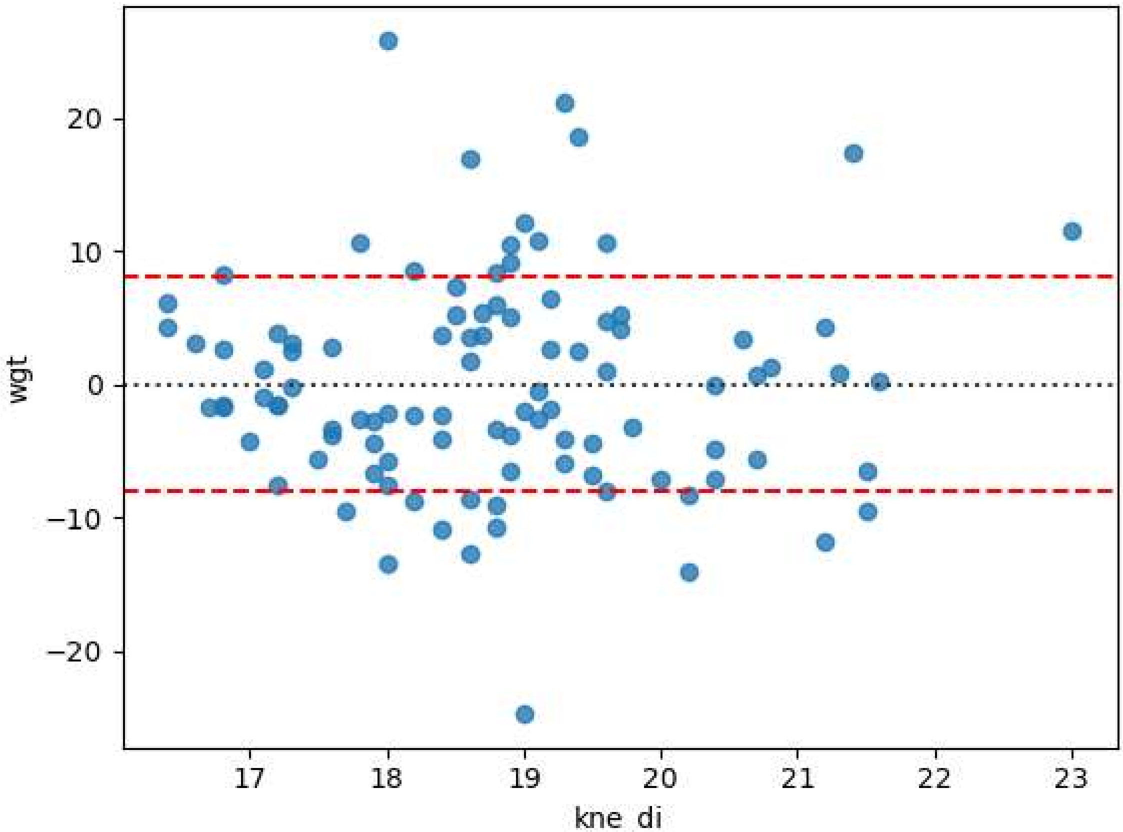
```
R2:  0.674
SE:  8.077
Correlation coefficient:  0.821
Regression equation: wgt =  8.57 * kne_di +  -92.237
```



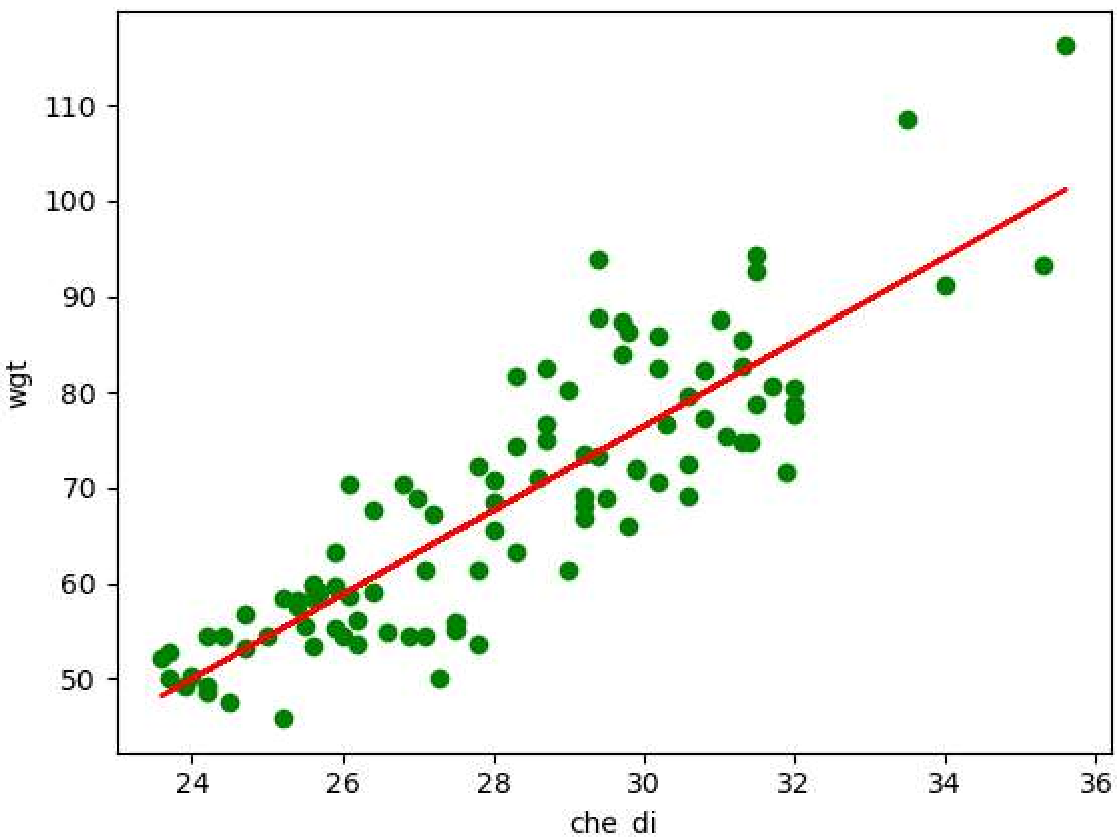
wgt
kne_di
110
100
90
80
70
60
50
17
18
19
20
21
22
23

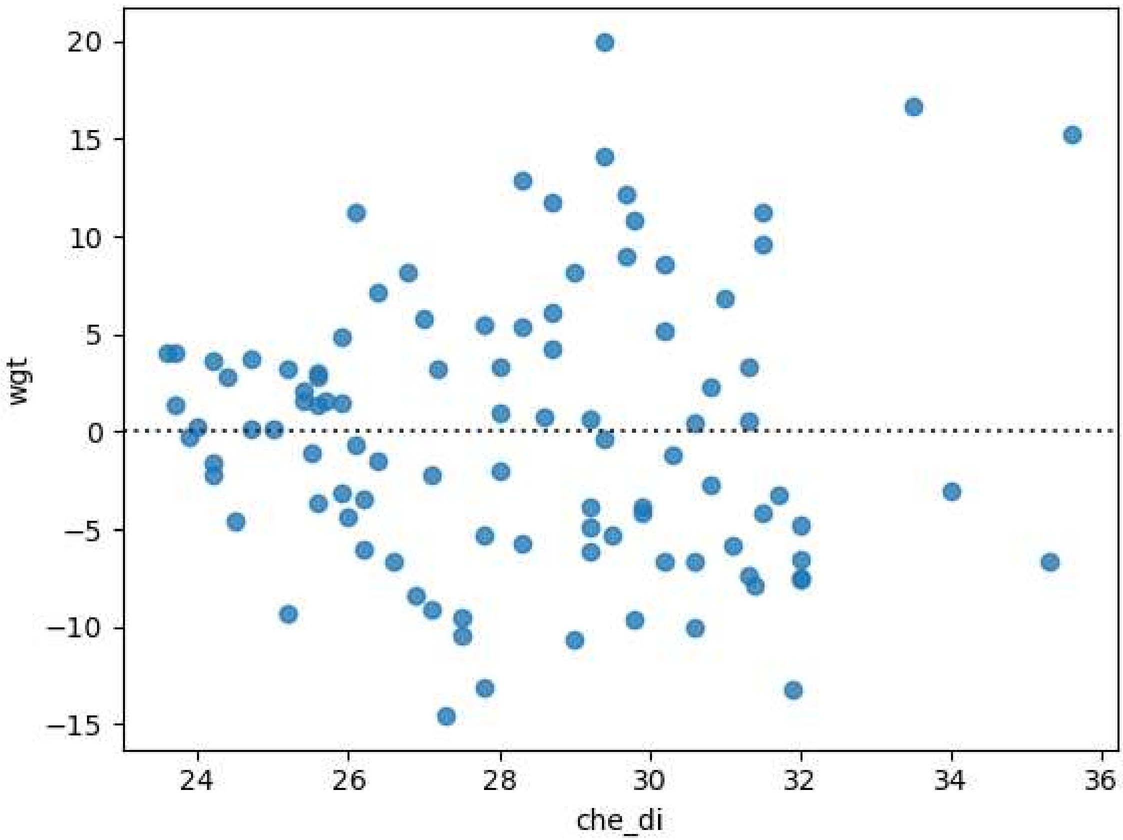


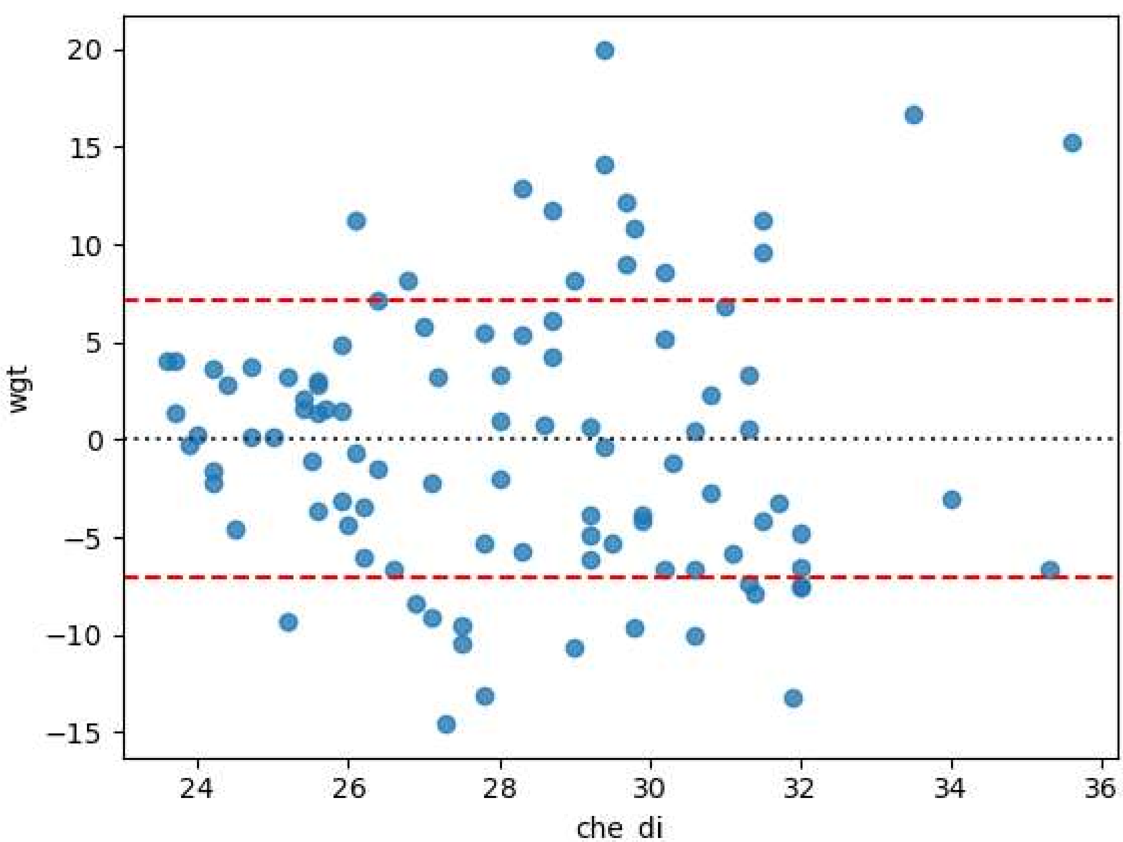
wgt
kne_di
20
10
0
−10
−20
17
18
19
20
21
22
23




```
R2:  0.748
SE:  7.102
Correlation coefficient:  0.865
Regression equation: wgt =  4.417 * che_di +  -56.048
```







1. **R^2 (Coefficient of Determination):** The R^2 value measures the proportion of the variance in the dependent variable (weight) that is predictable from the independent variable (knee diameter). In this case, the R^2 value is 0.674, indicating that approximately

67.4% of the variability in weight can be explained by knee diameter. A higher R^2 value closer to 1 indicates a better fit of the model.

2. **Standard Error (SE):** The standard error represents the average amount that the observed values differ from the predicted values by the regression line. In this case, the standard error is 8.077, which means that, on average, the actual weight values deviate from the predicted values by approximately 8.077 kilograms.

3. **Correlation Coefficient:** The correlation coefficient measures the strength and direction of the linear relationship between knee diameter and weight. In this case, the correlation coefficient is 0.821, indicating a strong positive linear relationship between knee diameter and weight. A value closer to 1 or -1 indicates a stronger linear relationship.

4. **Regression Equation:** The regression equation represents the relationship between the independent variable (knee diameter) and the dependent variable (weight). In this case, the regression equation is:
`Weight = 8.57 * Knee Diameter - 92.237`
Interpretation:

   - Slope (Coefficient): The slope of 8.57 means that, on average, for every one-unit increase in knee diameter, the weight increases by approximately 8.57 kilograms.
   - Y-intercept: The y-intercept of -92.237 represents the estimated weight when knee diameter is zero. However, in this context, interpreting the y-intercept may not be meaningful since knee diameter cannot be zero for a person.

5. **Residual Plot:** The residual plot is used to assess the goodness-of-fit of the regression model. It displays the residuals (the differences between observed and predicted values) against the independent variable (knee diameter). Ideally, the residuals should be randomly scattered around the horizontal line at zero, indicating that the errors are randomly distributed and the model is a good fit.

Interpreting these metrics and plots helps in understanding how well the linear regression model explains the relationship between knee diameter and weight. In this case, with a high R^2 value, strong correlation coefficient, and a well-scattered residual plot, the model appears to be a good fit for predicting weight based on knee diameter.

1. **R^2 (Coefficient of Determination):** The R^2 value is 0.748, indicating that approximately 74.8% of the variability in weight can be explained by chest diameter. This suggests that chest diameter is a good predictor of weight in the dataset.

2. **Standard Error (SE):** The standard error is 7.102, which represents the average amount that the observed weight values deviate from the predicted values by the regression line. A lower standard error indicates that the model's predictions are closer to the actual values.

3. **Correlation Coefficient:** The correlation coefficient is 0.865, indicating a strong positive linear relationship between chest diameter and weight. This suggests that as chest diameter increases, weight tends to increase as well, and vice versa.

4. **Regression Equation:** The regression equation is:
   `Weight = 4.417 * Chest Diameter - 56.048`

   - Slope (Coefficient): The slope of 4.417 means that, on average, for every one-unit increase in chest diameter, the weight increases by approximately 4.417 kilograms.
   - Y-intercept: The y-intercept of -56.048 represents the estimated weight when chest diameter is zero. However, in reality, chest diameter cannot be zero for a person, so the interpretation of the y-intercept may not be meaningful.
5. **Residual Plot:** The residual plot should be examined to assess the goodness-of-fit of the regression model. Ideally, the residuals should be randomly scattered around the horizontal line at zero, indicating that the errors are randomly distributed and the model is a good fit.

Overall, based on the high R^2 value, strong correlation coefficient, and well-scattered residual plot, the model appears to be a good fit for predicting weight based on chest diameter.

In [ ]:

```
```