Week 3 Assignment 1

Summer 2025 GRAD 50500-001 DIS
Caleb Welsh
Date Submitted: 05/22/2025
Date Due: 05/25/2025

Code

```
import pandas as pd
import matplotlib.pyplot as plt
from sklearn import datasets
import numpy as np
from itertools import combinations


# 1. Iris dataset questions

# Load iris dataset
iris = datasets.load_iris()
iris_df = pd.DataFrame(iris.data, columns=iris.feature_names)
iris_df['species'] = pd.Categorical.from_codes(iris.target,
iris.target_names)


# 1a. Make a histogram of the variable Sepal.Width.
print("#1a: Histogram of Sepal Width (cm)")
print()
fig1 = plt.figure(figsize=(7, 5))
fig1.canvas.manager.set_window_title('Figure 1')
plt.hist(iris_df['sepal width (cm)'], bins=22, color='skyblue',
edgecolor='black')
plt.title('Histogram of Sepal Width (cm)')
plt.xlabel('Sepal Width (cm)')
plt.ylabel('Frequency')
plt.grid(axis='y', alpha=0.75)
plt.show()

# 1b. Based on the histogram from #1a, which would you expect to be
higher, the mean or the median? Why?
print("#1b: Based on the histogram in #1a, I would say the mean seems to
be greater than the median because it is right skewed, its hard to tell
with the histogram.")
print()

# 1c. Confirm your answer to #1b by actually finding these values.
mean_sw = iris_df['sepal width (cm)'].mean()
median_sw = iris_df['sepal width (cm)'].median()
print(f"#1c: Mean Sepal.Width and Median Sepal.Width are {mean_sw:.2f} and
{median_sw:.2f} respectively.")
print()

# 1d. Only 27% of the flowers have a Sepal.Width higher than ________ cm.
```

```
threshold = iris_df['sepal width (cm)'].quantile(0.73)
print(f"#1d: Only 27% of the flowers have a Sepal.Width higher than
{threshold:.2f} cm.\n")
print()

# 1e. Make scatterplots of each pair of the numerical variables in iris
(There should be 6 pairs/plots).
fig2 = plt.figure(figsize=(15, 10))
fig2.canvas.manager.set_window_title('Figure 2')
num_cols = iris.feature_names
print(num_cols)
print()
for i, (x, y) in enumerate(combinations(num_cols, 2), 1):
    print(f"Plot {i}: {x} vs {y}")
    plt.subplot(2, 3, i)
    plt.scatter(iris_df[x], iris_df[y], alpha=0.7,
c=iris_df['species'].cat.codes, cmap='Accent')
    plt.xlabel(x)
    plt.ylabel(y)
    plt.title(f'{x} vs {y}')
plt.tight_layout()
plt.show()
print()

# 1f. Based on #1e, which two variables appear to have the strongest
relationship? And which two appear to have the weakest relationship?
print("#1f: Based on the scatterplots in #1e, petal length and petal width
are strongly related, and sepal length and sepal width are weakly
related.")
print()
```

```
# 2. PlantGrowth dataset questions

# Load PlantGrowth dataset
data = {
    "weight": [4.17, 5.58, 5.18, 6.11, 4.50, 4.61, 5.17, 4.53, 5.33, 5.14,
4.81, 4.17, 4.41, 3.59, 5.87, 3.83, 6.03, 4.89, 4.32, 4.69, 6.31, 5.12,
5.54, 5.50, 5.37, 5.29, 4.92, 6.15, 5.80, 5.26],
    "group": ["ctrl"] * 10 + ["trt1"] * 10 + ["trt2"] * 10
}
PlantGrowth = pd.DataFrame(data)

# 2a. Make a histogram of the variable weight with breakpoints (bin edges)
at every 0.3 units, starting at 3.3.
print("#2a: Histogram of Plant Weight")
print()
bins = np.arange(3.3, PlantGrowth['weight'].max() + 0.3, 0.3)
fig3 = plt.figure(figsize=(7, 5))
fig3.canvas.manager.set_window_title('Figure 3')
plt.hist(PlantGrowth['weight'], bins=bins, color='olive',
edgecolor='black')
plt.title('Histogram of Plant Weight')
plt.xlabel('Weight')
plt.ylabel('Frequency')
plt.grid(axis='y', alpha=0.75)
plt.show()

# 2b. Make boxplots of weight separated by group in a single graph.
print("#2b: Boxplot of Plant Weight by Group")
print()
fig4 = plt.figure(figsize=(7, 5))
fig4.canvas.manager.set_window_title('Figure 4')
ax = fig4.add_subplot(111)
PlantGrowth.boxplot(column='weight', by='group', grid=False,
patch_artist=True, boxprops=dict(facecolor='magenta'), ax=ax)
ax.set_title('Boxplot of Plant Weight by Group')
fig4.suptitle('')
ax.set_xlabel('Group')
ax.set_ylabel('Weight')
plt.show()

# 2c. Based on the boxplots in #2b, approximately what percentage of the
"trt1" weights are below the minimum "trt2" weight?
print(f"#2c: Based on the boxplots about 8 of the 10 elements in the
'trt1' weights are below the minimum 'trt2' weight so about 80%.")
print()
```

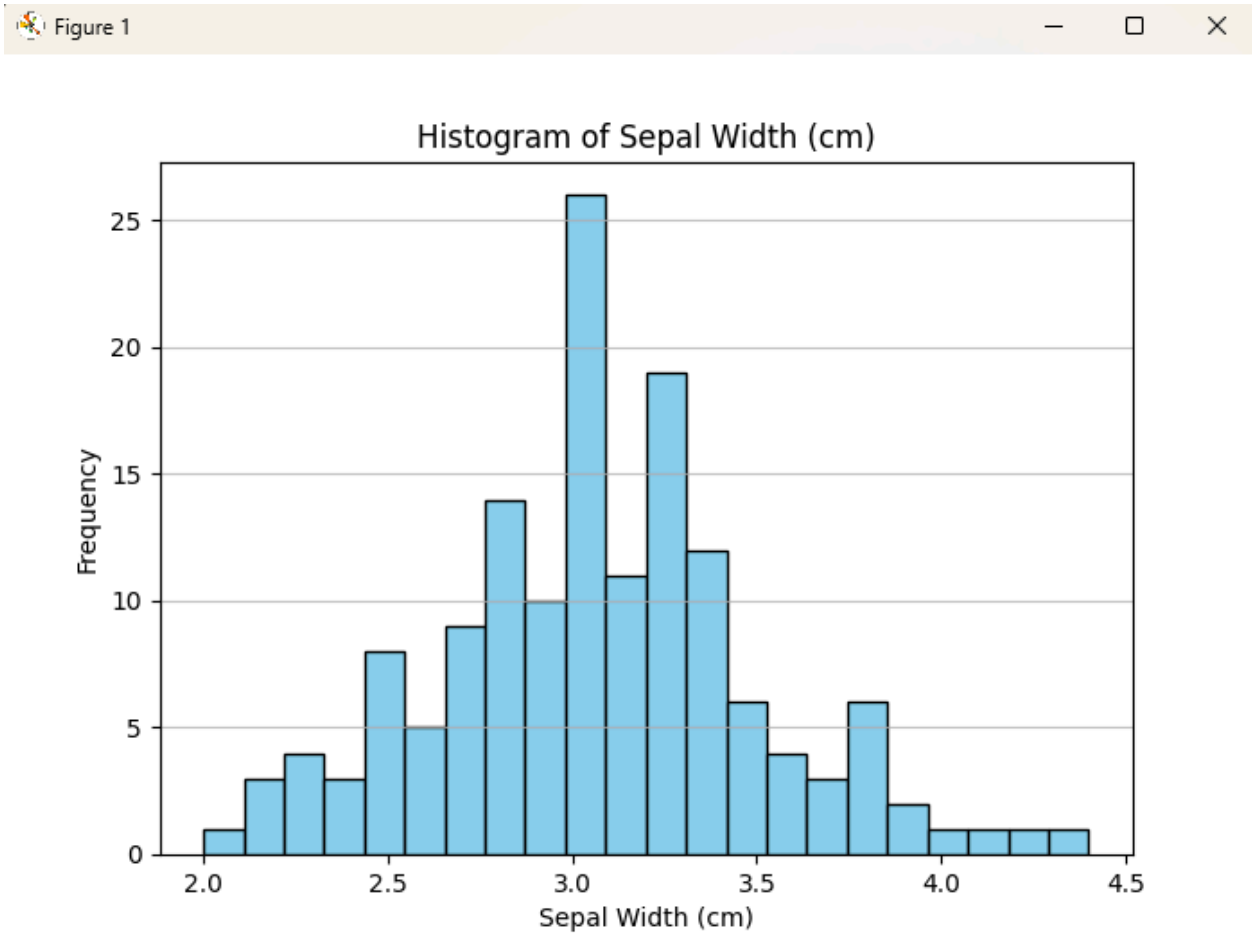
```
# 2d. Find the exact percentage of the "trt1" weights that are below the
minimum "trt2" weight.
min_trt2 = PlantGrowth[PlantGrowth['group'] == 'trt2']['weight'].min()
trt1_weights = PlantGrowth[PlantGrowth['group'] == 'trt1']['weight']
# print(f"(trt1_weights < min_trt2): {(trt1_weights < min_trt2)}")
exact_pct = (trt1_weights < min_trt2).mean() * 100
print(f"#2d: Exactly {exact_pct:.2f}% of the 'trt1' weights are below the
minimum 'trt2' weight.")
print()

# 2e. Only including plants with a weight above 5.5, make a barplot of the
variable group. Make the barplot colorful using a color palette.
print("#2e: Group Barplot Weights of 5.5 and above.")
threshold = PlantGrowth[PlantGrowth['weight'] > 5.5]
group_counts = threshold['group'].value_counts()
colors = ['tomato', 'gold', 'mediumseagreen']
fig5 = plt.figure(figsize=(6, 5))
fig5.canvas.manager.set_window_title('Figure 5')
ax = group_counts.plot(kind='bar', color=colors)
for bar in ax.patches:
    bar.set_edgecolor('black')
plt.title('Count of Groups (Weight > 5.5)')
plt.xlabel('Group')
plt.ylabel('Count')
plt.show()


# Used cursor to supliment comments and understand the code more.
```

Output

Iris dataset questions

#1a: Histogram of Sepal Width (cm)

Figure 1 - Histogram of Sepal Width (cm)



#1b: Based on the histogram in #1a, I would say the mean seems to be greater than the median because it is right skewed, its hard to tell with the histogram.

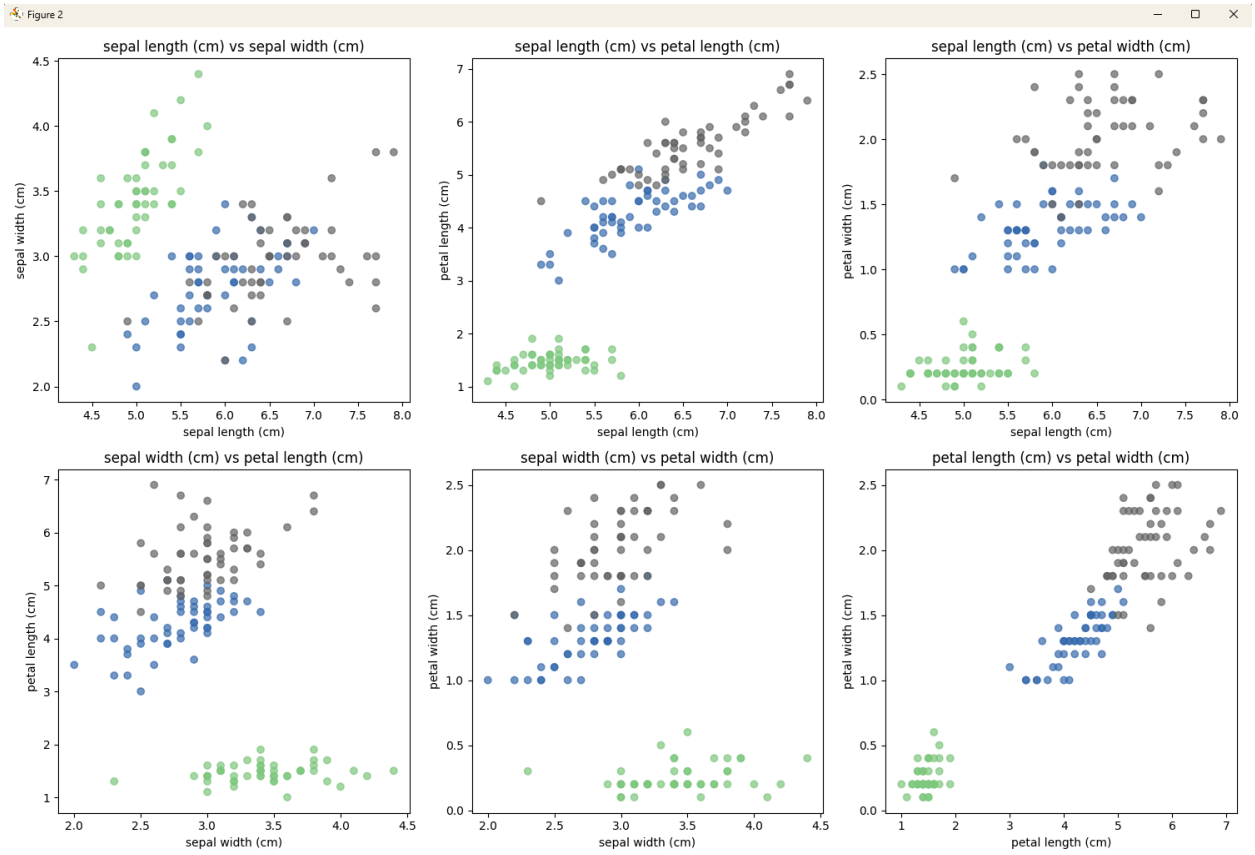
#1c: Mean Sepal.Width and Median Sepal.Width are 3.06 and 3.00, respectively.

#1d: Only 27% of the flowers have a Sepal.Width higher than 3.30 cm.

['sepal length (cm)', 'sepal width (cm)', 'petal length (cm)', 'petal width (cm)']

Plot 1: sepal length (cm) vs sepal width (cm)
Plot 2: sepal length (cm) vs petal length (cm)
Plot 3: sepal length (cm) vs petal width (cm)
Plot 4: sepal width (cm) vs petal length (cm)
Plot 5: sepal width (cm) vs petal width (cm)
Plot 6: petal length (cm) vs petal width (cm)

Figure 2 - Iris Feature Scatterplots

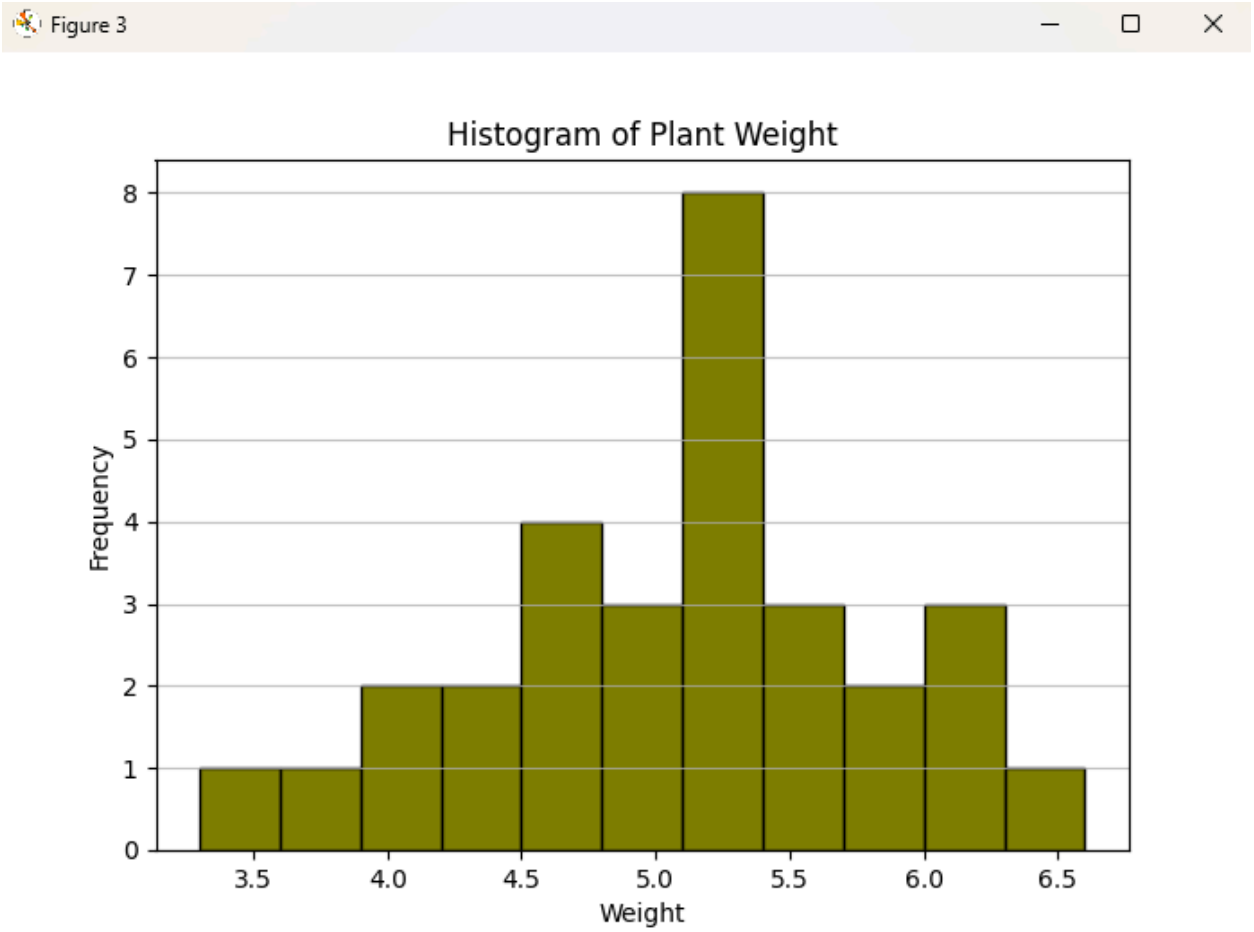


#1f: Based on the scatterplots in #1e, petal length and petal width are strongly related, and sepal length and sepal width are weakly related.

PlantGrowth dataset questions

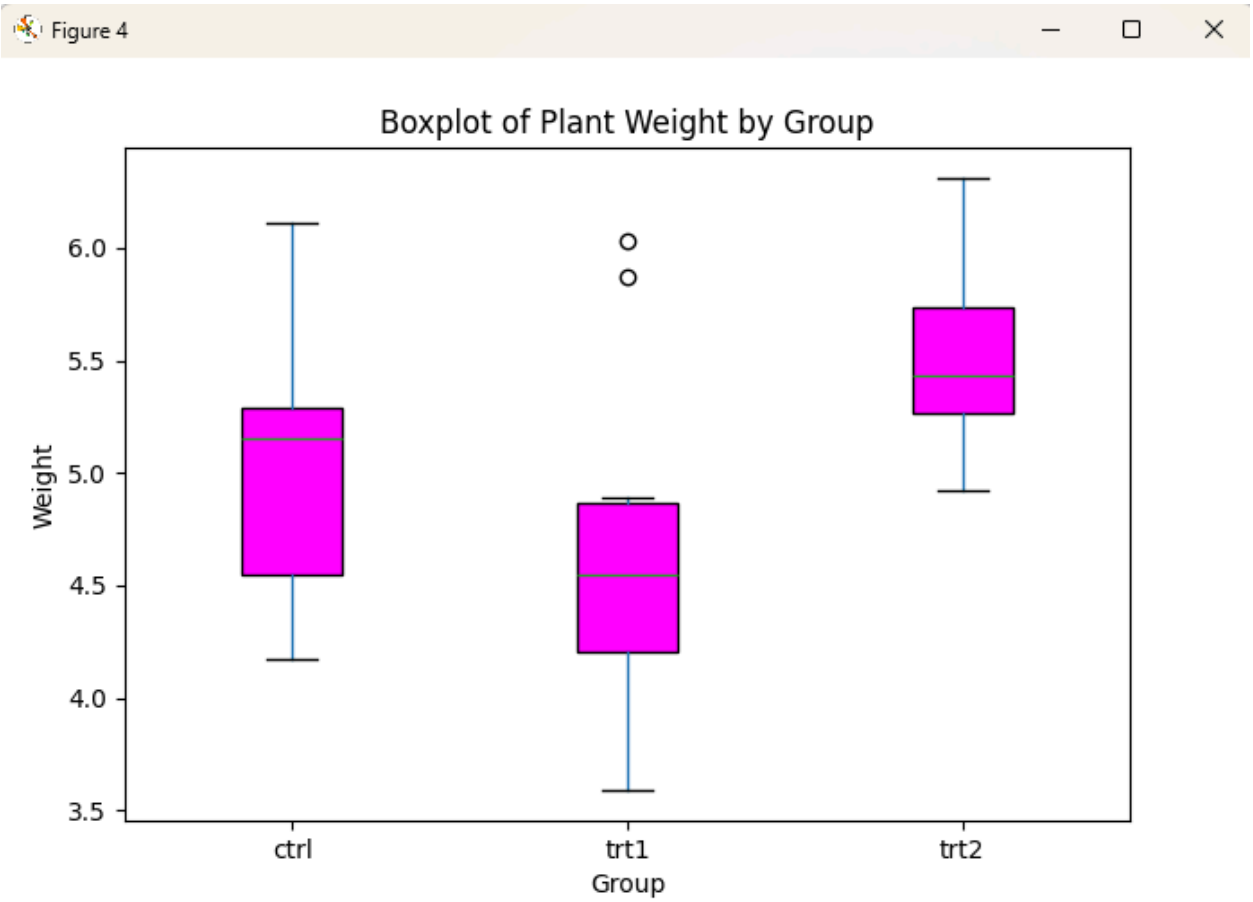
#2a: Histogram of Plant Weight

Figure 3 - Histogram of Plant Weight



#2b: Boxplot of Plant Weight by Group

Figure 4 - Boxplot of Plant Weight by Group

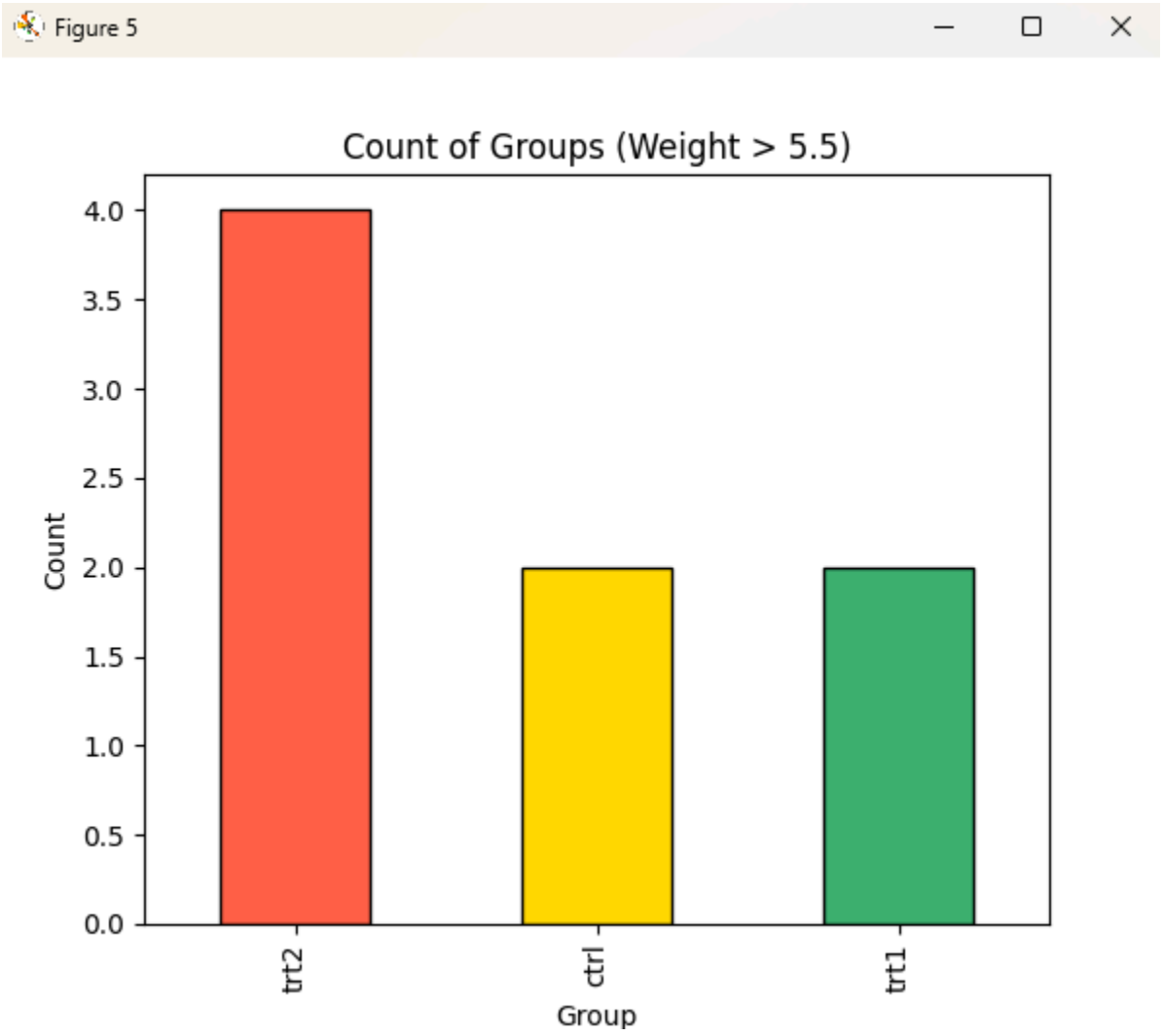


#2c: Based on the boxplots about 8 of the 10 elements in the 'trt1' weights are below the minimum 'trt2' weight so about 80%.

#2d: Exactly 80.00% of the 'trt1' weights are below the minimum 'trt2' weight.

#2e: Group Barplot Weights of 5.5 and above.

Figure 5 - Group Barplot Weights of 5.5 and above.



References

Cursor—The AI Code Editor. (n.d.). Retrieved May 22, 2025, from https://www.cursor.com/

pandas documentation—Pandas 2.2.3 documentation. (n.d.). Retrieved May 22, 2025, from https://pandas.pydata.org/docs/