DTSC 2301 Spring 2025 Homework #1

Turn in your assignment via Gradescope

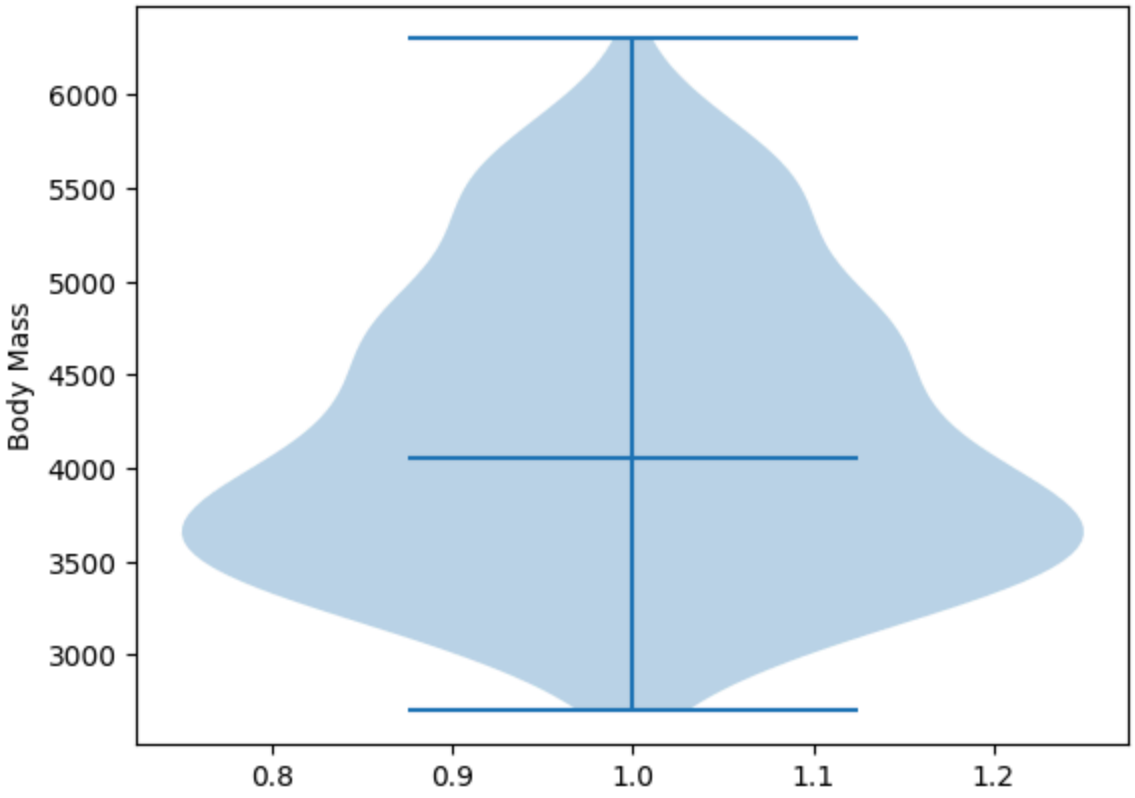
Due 1/17/25, 11: 59pm

For this assignment you may *not* use any generative AI and you may only use python commands and code we used in class.

Question 1

Read in the Penguins dataset (https://webpages.charlotte.edu/mschuck1/classes/DTSC2301/Data/penguins.csv). Create a violinplot for penguin body mass.

In [6]:
```
import pandas as pd
import matplotlib.pyplot as plt
import numpy as np
penguins = pd.read_csv("https://webpages.charlotte.edu/mschuck1/classes/DTSC2301/Da
penguins_clean = penguins.dropna(subset=['body_mass_g'])
plt.violinplot(penguins_clean['body_mass_g'],showmedians=True, showextrema=True)
plt.ylabel('Body Mass')
plt.show()
```



Question 2

Again using the data on body mass of penguins, create and interpret a 95% confidence interval for the mean body weight of a penguin based upon this sample.

In [11]:
```
import scipy.stats as st
body_mass_g = penguins_clean['body_mass_g']
st.t.interval(confidence=0.95,
              df=len(body_mass_g)-1,
              loc=np.mean(body_mass_g),
              scale=st.sem(body_mass_g))
```

Out[11]:
```
(np.float64(4116.458332024052), np.float64(4287.050439905773))
```

This is saying that the mean body mass of penguins has a 95% chance of being between 4,116.46 grams and 4,287.05 grams

Question 3

Using the Ames Housing Data (https://webpages.charlotte.edu/mschuck1/classes/DTSC2301/Data/Ames_house_prices.csv), create a 92% confidence interval for the mean above grade (ground) living area square feet (GrLivArea). Interpret this interval in the context of these data.

In [21]:
```
ames = pd.read_csv("https://webpages.charlotte.edu/mschuck1/classes/DTSC2301/Data/A
grlivarea_clean = ames.dropna(subset=['GrLivArea'])['GrLivArea']
st.t.interval(confidence=0.92,
              df=len(grlivarea_clean)-1,
              loc=np.mean(grlivarea_clean),
              scale=st.sem(grlivarea_clean))
```

Out[21]:
```
(np.float64(1491.3706943731436), np.float64(1539.5567028871305))
```

This is saying that there is a 92% chance the average square footage above ground in Ames is between 1491.37 and 1539.56 square feet.

Question 4

Using the Ames Housing Data, create a 90% confidence interval for the standard deviation of Sale Price using the bootstrap. Interpret this interval in the context of these data.

In [30]:
```
saleprice_clean = ames.dropna(subset=['SalePrice'])['SalePrice']
n_bootstraps = 1000
bootstrap_samples = np.random.choice(saleprice_clean, (n_bootstraps, len(saleprice_
bootstrap_std_devs = np.std(bootstrap_samples, axis=1)
lower_bound = np.percentile(bootstrap_std_devs, 5)
upper_bound = np.percentile(bootstrap_std_devs, 95)
print(f"90% confidence interval for the standard deviation of SalePrice: ({lower_bo
```

```
90% confidence interval for the standard deviation of SalePrice: (74490.99, 84099.6
4)
```

This means we are 90% confident the standard deviation of Sale Prices is between $74,490.99 and 84,099.64$

Question 5

Write a short 80-120 word paragraph explaining how data scientists should ensure that the analysis and its interpretation that you made about the Ames Housing Data do not inadvertently reinforce bias or inequalities in the housing market.

There are multiple ways, we as data scientists can prevent bias from being shown in our work. I think one way we can do this is by familiarizing ourself with the historical context of the data and identifying any potentially bias sources. I think another way we can prevent bias in this case is by ensuring the data sample we are using is reprensentative of the population we are aiming to look at (Ames). Finally, I think in all cases we must have transparency s stakeholder can understand our decision making process.