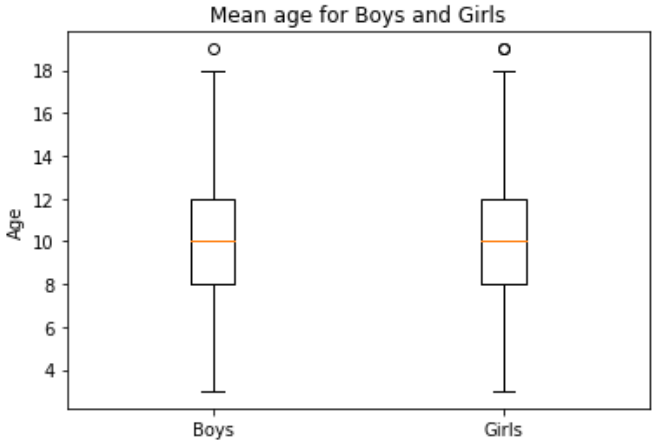
Project 1: Summarization of Data
Lincoln Nordquist

Case Study 1:

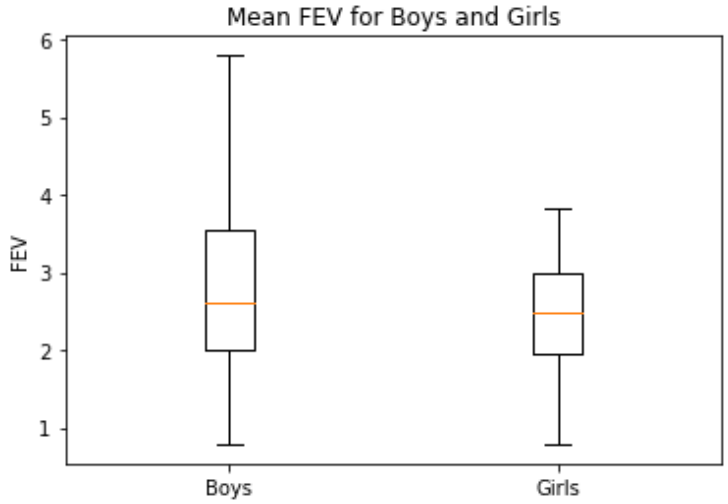
i) Descriptive statistics for each variable

Code for question 1:

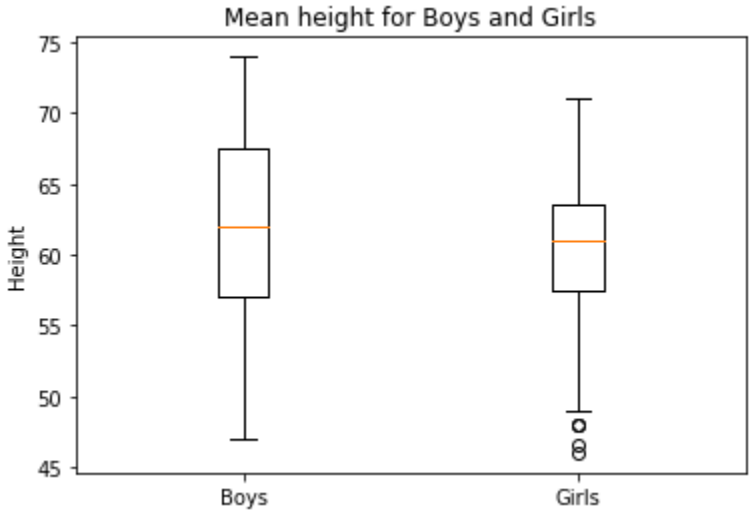
```
data = pd.read_csv('data.csv')
d = pd.DataFrame(data)

d.describe()

boys = d[d['Sex'] == 1]
girls = d[d['Sex'] == 0]



# question 1

print('Mean of Age (boys): ', round(np.mean(boys['Age']), 1))
print('Mean of Age (girls): ', round(np.mean(girls['Age']), 1))
plt.boxplot([boys.Age, girls.Age], labels=['Boys', 'Girls'])
plt.ylabel('Age')
plt.title('Mean age for Boys and Girls')

print('Mean of FEV (boys): ', round(np.mean(boys['FEV']),1))
print('Mean of FEV (girls): ', round(np.mean(girls['FEV']),1))
plt.boxplot([boys.FEV, girls.FEV], labels=['Boys', 'Girls'])
plt.ylabel('FEV')
plt.title('Mean FEV for Boys and Girls')

print('Mean of Height (boys): ',round(np.mean(boys['Hgt']),1))
print('Mean of Height (girls): ',round(np.mean(girls['Hgt']),1))
plt.boxplot([boys.Hgt, girls.Hgt], labels=['Boys', 'Girls'])
plt.ylabel('Height')
plt.title('Mean height for Boys and Girls')

print('Mean of Smoke (boys): ',round(np.mean(boys['Smoke']),0))
print('Mean of Smoke (girls): ',round(np.mean(girls['Smoke']),0))
plt.boxplot([boys.Smoke, girls.Smoke], labels=['Boys', 'Girls'])
plt.ylabel('Smoking Status')
plt.title('Mean smoking status for Boys and Girls')
```



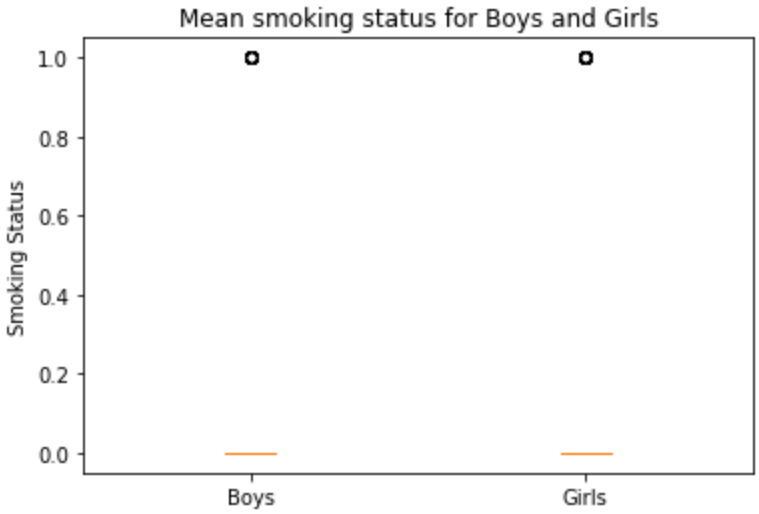
In this boxplot, we can see the mean age between boys and girls. We can see that the boys and girls have roughly the same mean and the same variation



In this boxplot, we can see the mean FEV between boys and girls, and once again the boys and girls have roughly the same mean, but this time, the boys have a wider range than the girls.



In this boxplot, we can see the mean height between boys and girls. The boys have a slightly higher mean, and the girls have more outliers.

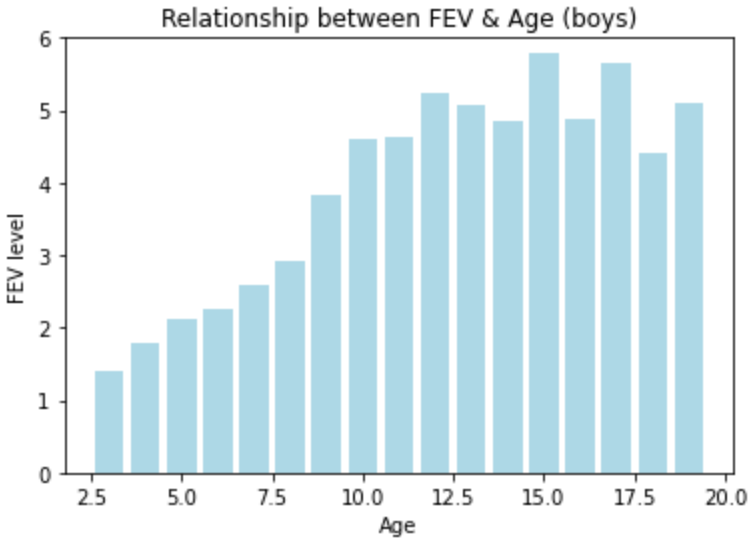


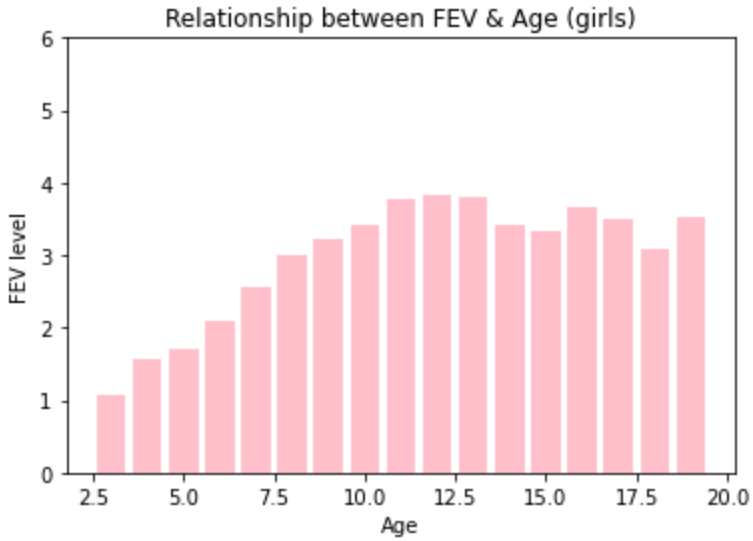
This boxplot shows the average smoking status between boys and girls. This plot is a bit harder to read, but we can see that the average boy and girl is not a smoker, but that there are still smokers between both genders

ii) Assess relationship of FEV to age, height, and smoking status (for boys and girls)

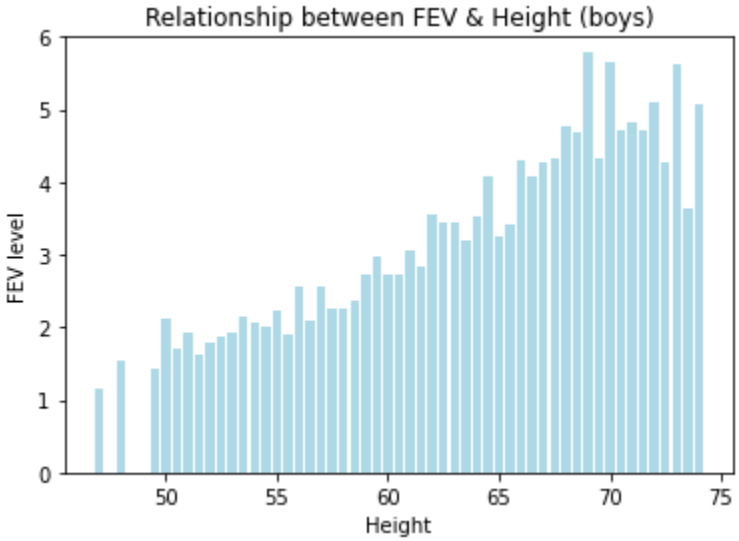
Code for question 2:

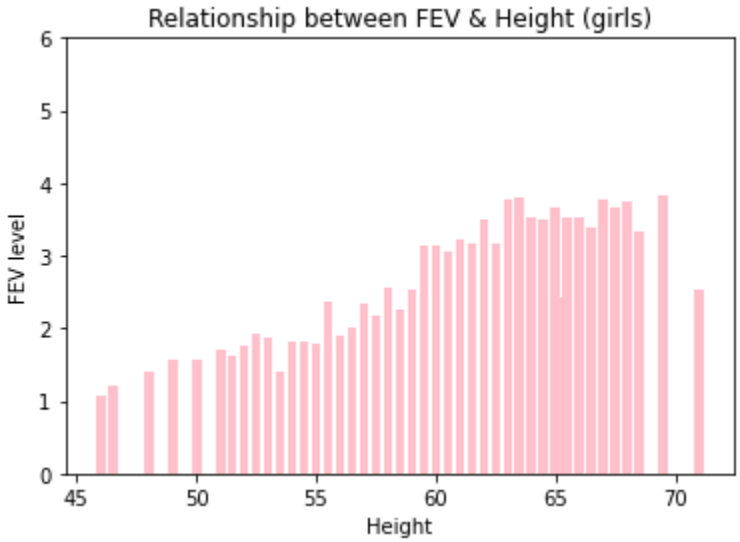
```
# comparison of FEV to age
plt.bar(boys.Age, boys.FEV, color="lightblue")
plt.ylim(0,6)
plt.xlabel('Age')
plt.ylabel('FEV level')
plt.title('Relationship between FEV & Age (boys)')
plt.show()

plt.bar(girls.Age, girls.FEV, color='pink')
plt.ylim(0,6)
plt.xlabel('Age')
plt.ylabel('FEV level')
plt.title('Relationship between FEV & Age (girls)')
plt.show()

# comparison of FEV to height
plt.bar(boys.Hgt, boys.FEV, color="lightblue", width=(0.4))
plt.ylim(0,6)
plt.xlabel('Height')
plt.ylabel('FEV level')
plt.title('Relationship between FEV & Height (boys)')
plt.show()

plt.bar(girls.Hgt, girls.FEV, color='pink', width=(0.4))
plt.ylim(0,6)
plt.xlabel('Height')
plt.ylabel('FEV level')
plt.title('Relationship between FEV & Height (girls)')
plt.show()

# comparison of FEV to smoking status

boys_smoke_yes = boys[d['Smoke'] == 1]
boys_smoke_no = boys[d['Smoke'] == 0]
girls_smoke_yes = girls[d['Smoke'] == 1]
girls_smoke_no = girls[d['Smoke'] == 0]

plt.scatter(boys.Smoke, boys.FEV, color='lightblue')
plt.xlabel('Smoking Status (no/yes)')
plt.ylabel('FEV level')
plt.title('Relationship between FEV & Smoking Status (boys)')

plt.scatter(girls.Smoke, girls.FEV, color='pink')
plt.xlabel('Smoking Status (yes/no)')
plt.ylabel('FEV level')
plt.title('Relationship between FEV & Smoking Status (girls)')
```



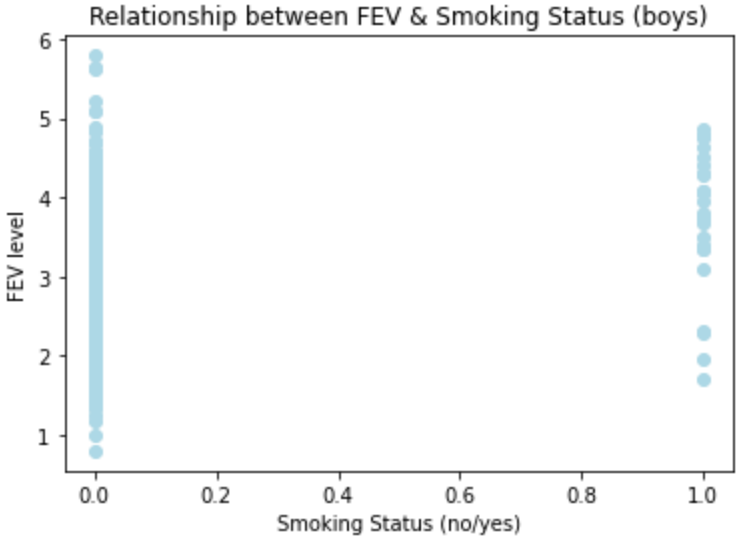


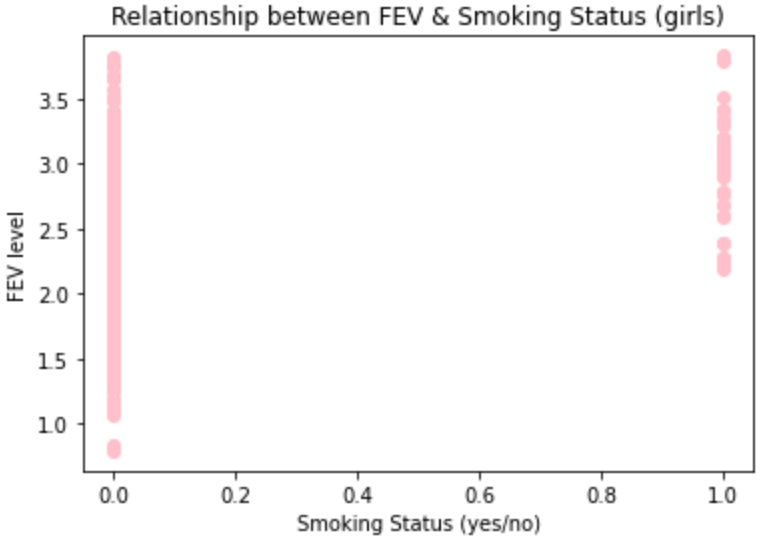
These graphs show the relationship between FEV and Age between boys and girls. As we can see, there seems to be a correlation between higher age and higher FEV levels. This correlation is present with both boys and girls, although age seems to have a greater effect on boys as shown in the greater FEV levels.





These graphs represent the relationship between height and FEV levels between boys and girls. Once again we are able to see a clear correlation. A greater height seems to correlate to higher FEV levels between both genders. Despite both genders having the same correlation, boys once again appear to have a higher spike than the girls.



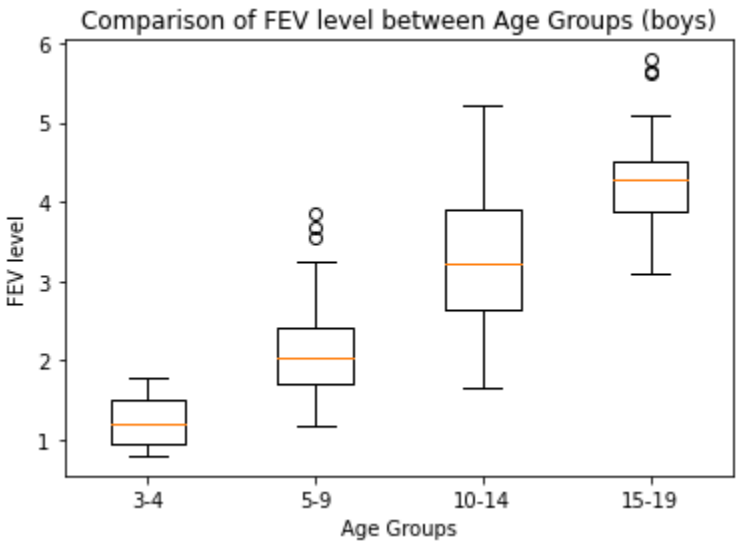


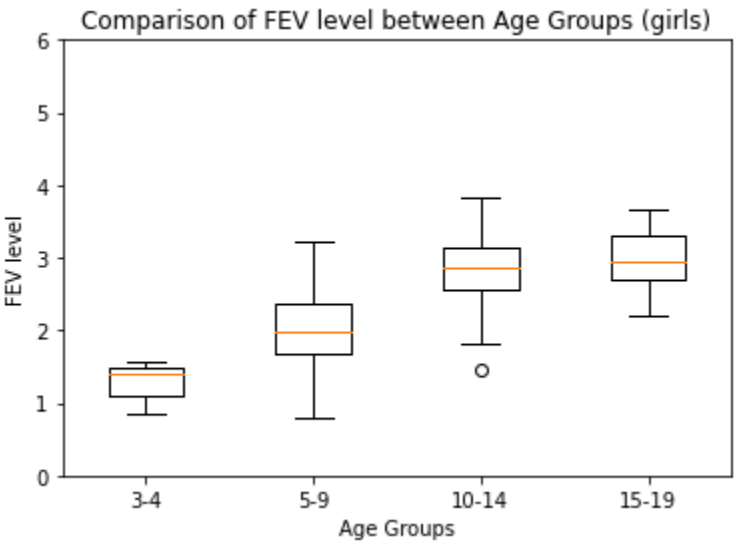
These scatter plots show us the relationship between smoking status and FEV levels between boys and girls. This graph is very interesting because there appears to be higher FEV levels between boys who don't smoke. The girls, however, seem to have consistently high FEV levels with smoking, and much more variance without smoking. This correlation appears to be much less consistent and makes me think that smoking status doesn't affect FEV levels as much as the previous criteria.

iii) Compare patterns of growth of FEV by age group for boys and girls

Question 3 code:

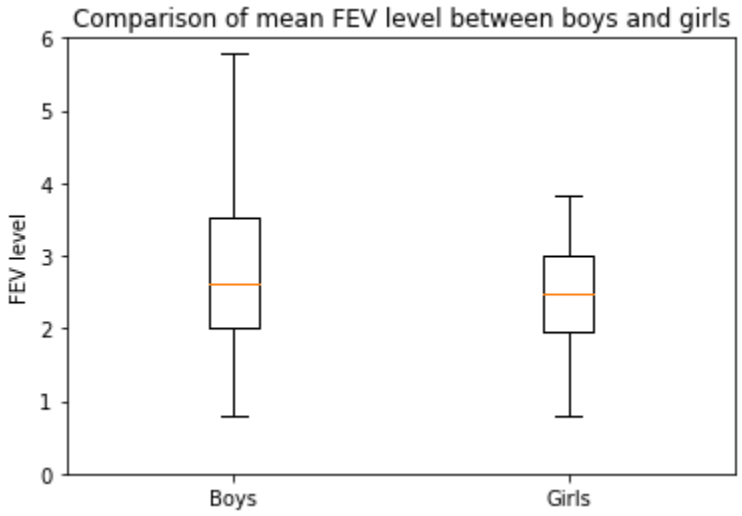
```
# question 3
bin1_b = (boys[(boys['Age'] >=3) & (boys['Age'] <=4)])
bin2_b = (boys[(boys['Age'] >=5) & (boys['Age'] <=9)])
bin3_b = (boys[(boys['Age'] >=10) & (boys['Age'] <=14)])
bin4_b = (boys[(boys['Age'] >=15) & (boys['Age'] <=19)])

plt.boxplot([bin1_b.FEV, bin2_b.FEV, bin3_b.FEV, bin4_b.FEV], labels=['3-4', '5-9', '10-14', '15-19'])
plt.xlabel('Age Groups')
plt.ylabel('FEV level')
plt.title('Comparison of FEV level between Age Groups (boys)')

bin1_g = (girls[(girls['Age'] >=3) & (girls['Age'] <=4)])
bin2_g = (girls[(girls['Age'] >=5) & (girls['Age'] <=9)])
bin3_g = (girls[(girls['Age'] >=10) & (girls['Age'] <=14)])
bin4_g = (girls[(girls['Age'] >=15) & (girls['Age'] <=19)])

plt.boxplot([bin1_g.FEV, bin2_g.FEV, bin3_g.FEV, bin4_g.FEV], labels=['3-4', '5-9', '10-14', '15-19'])
plt.xlabel('Age Groups')
plt.ylim(0,6)
plt.ylabel('FEV level')
plt.title('Comparison of FEV level between Age Groups (girls)')

plt.boxplot([boys.FEV, girls.FEV], labels=['Boys', 'Girls'])
plt.ylim(0,6)
plt.ylabel('FEV level')
plt.title('Comparison of mean FEV level between boys and girls')
```





These boxplots show the pattern of growth of FEV between age groups for boys and girls. As we can see, the mean of each age group becomes higher, each time we go up an age group. This is consistent with our previous findings of there being a positive correlation between age and FEV levels. The data for the boys seems to be more spread apart, while the data for the

girls seems to be more closely aligned. The boys data also has more outliers as shown in the 5-9 age group, while the girls group has much less outliers.



In this boxplot, we are able to see, side-by-side, the difference in mean between FEV levels of boys and girls. This boxplot is able to illustrate how similar the means are between the two groups. Despite the similarity in mean, we are able to see that the boys have a higher range, as well as a higher interquartile range than the girls.

Thoughts:

Overall, I think that this case study had very accurate data due to very similar sample sizes between the boys and girls. We were able to see the correlations between age, height, and smoking status between both genders. Despite these similarities in data, the boys did appear to have a higher range of data between each graph, as well as more outliers.

Case Study 2:

i) Descriptive statistics

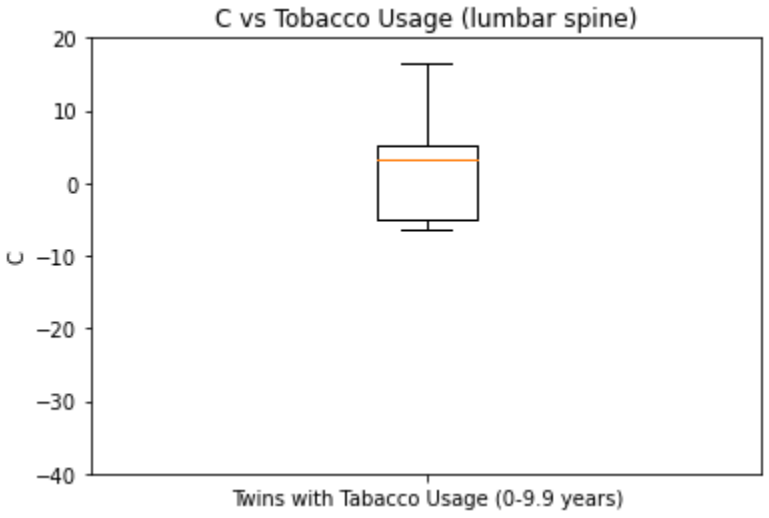
Code for question 1

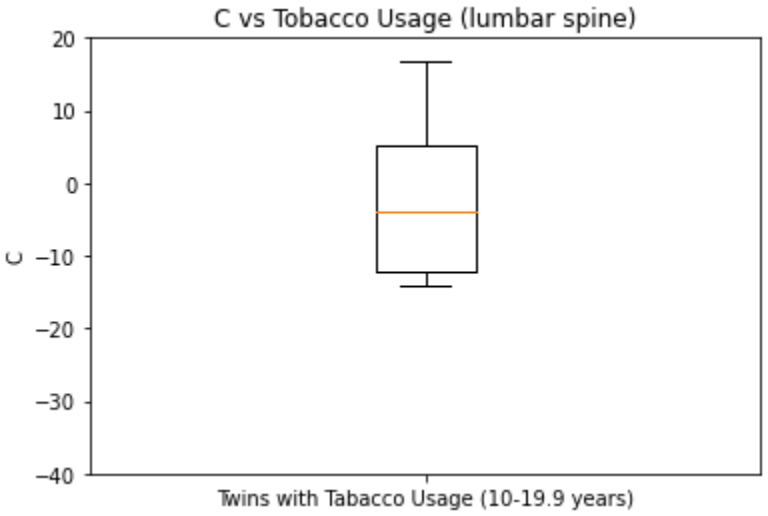
```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns

data = pd.read_csv('BONEDEN.csv')
d = pd.DataFrame(data)

# question 1

d['A1'] = (d['ls2'] - d['ls1'])
d['B1'] = (d['ls2'] + d['ls1']) / 2
d['C1'] = (d['A1']/d['B1']) * 100
d['smoke_diff'] = (d['pyr2'] - d['pyr1'])

d['A2'] = (d['fn2'] - d['fn1'])
d['B2'] = (d['fn2'] + d['fn1']) / 2
d['C2'] = (d['A2']/d['B2']) * 100

d['A3'] = (d['fs2'] - d['fs1'])
d['B3'] = (d['fs2'] + d['fs1']) / 2
d['C3'] = (d['A3']/d['B3']) * 100


d.describe()
```

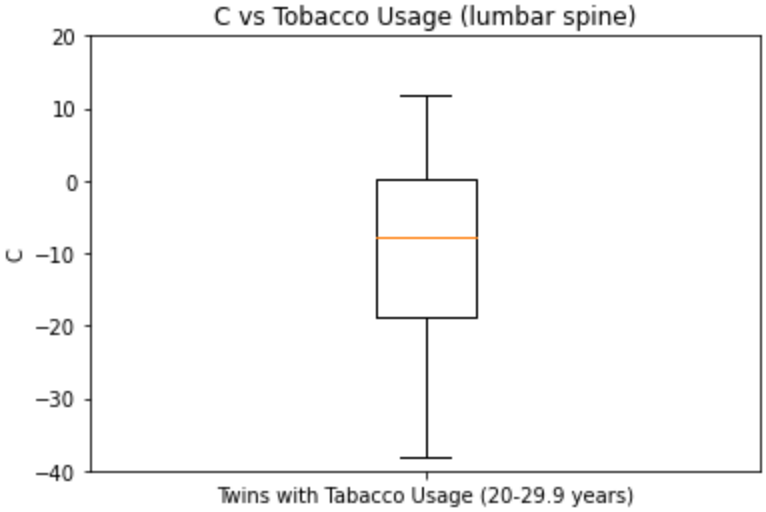
ii) Difference in tobacco expressed in 10 year groups (lumbar spine, femoral neck, femoral shaft)

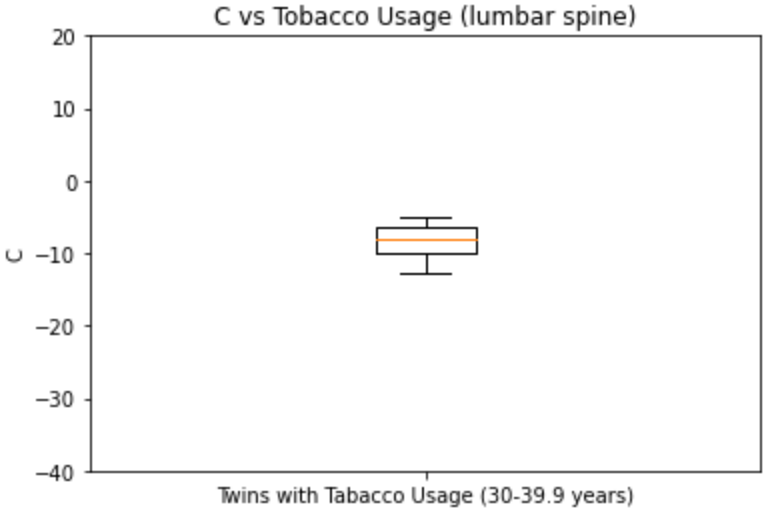
Lumbar spine data:

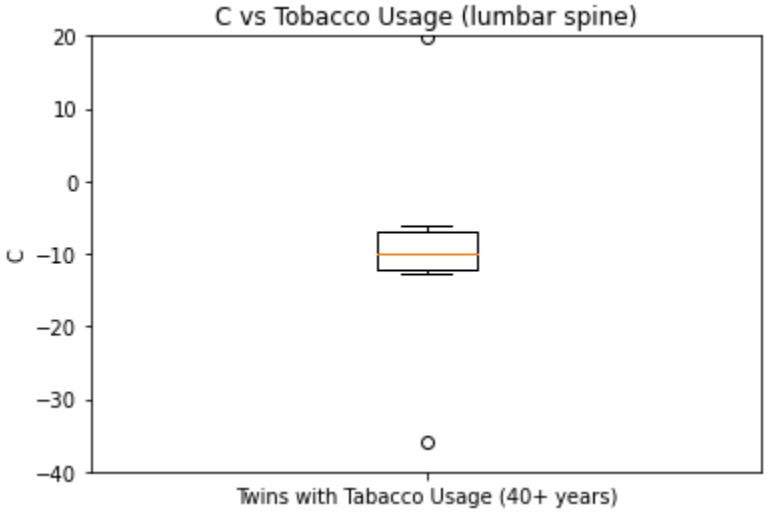
```
# question 2

bin1_1 = d[(d['pyr2'] - d['pyr1']>=0) & (d['pyr2'] - d['pyr1']<=9.9)]
bin2_1 = d[(d['pyr2'] - d['pyr1']>=10) & (d['pyr2'] - d['pyr1']<=19.9)]
bin3_1 = d[(d['pyr2'] - d['pyr1']>=20) & (d['pyr2'] - d['pyr1']<=29.9)]
bin4_1 = d[(d['pyr2'] - d['pyr1']>=30) & (d['pyr2'] - d['pyr1']<=39.9)]
bin5_1 = d[(d['pyr2'] - d['pyr1']>=40)]

print(bin1_1)

plt.boxplot(bin1_1.C1, labels=['Twins with Tabacco Usage (0-9.9 years)'])
plt.ylim(-40,20)
plt.ylabel('C')
plt.title('C vs Tobacco Usage (lumbar spine)')
plt.show()

plt.boxplot(bin2_1.C1, labels=['Twins with Tabacco Usage (10-19.9 years)'])
plt.ylim(-40,20)
plt.ylabel('C')
plt.title('C vs Tobacco Usage (lumbar spine)')
plt.show()

plt.boxplot(bin3_1.C1, labels=['Twins with Tabacco Usage (20-29.9 years)'])
plt.ylim(-40,20)
plt.ylabel('C')
plt.title('C vs Tobacco Usage (lumbar spine)')
plt.show()

plt.boxplot(bin4_1.C1, labels=['Twins with Tabacco Usage (30-39.9 years)'])
plt.ylim(-40,20)
plt.ylabel('C')
plt.title('C vs Tobacco Usage (lumbar spine)')
plt.show()

plt.boxplot(bin5_1.C1, labels=['Twins with Tabacco Usage (40+ years)'])
plt.ylim(-40,20)
plt.ylabel('C')
plt.title('C vs Tobacco Usage (lumbar spine)')
plt.show()
```







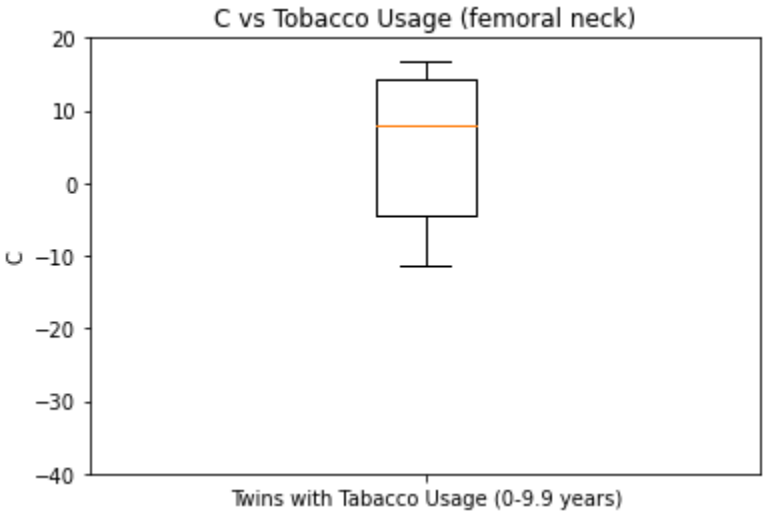


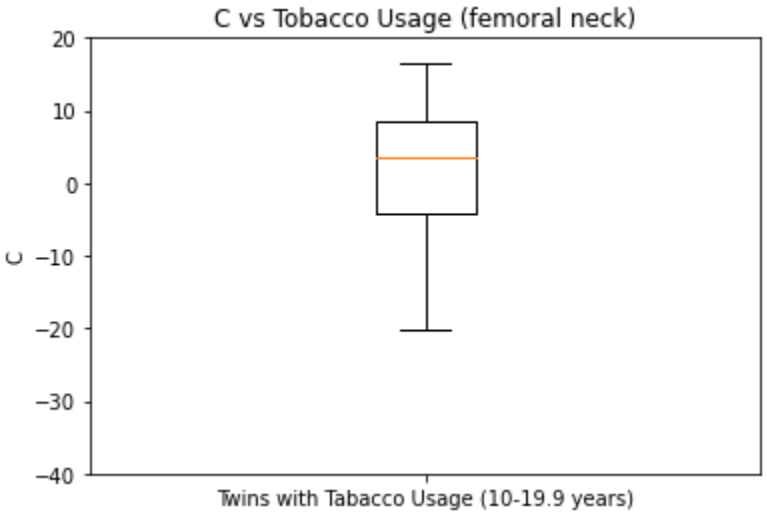


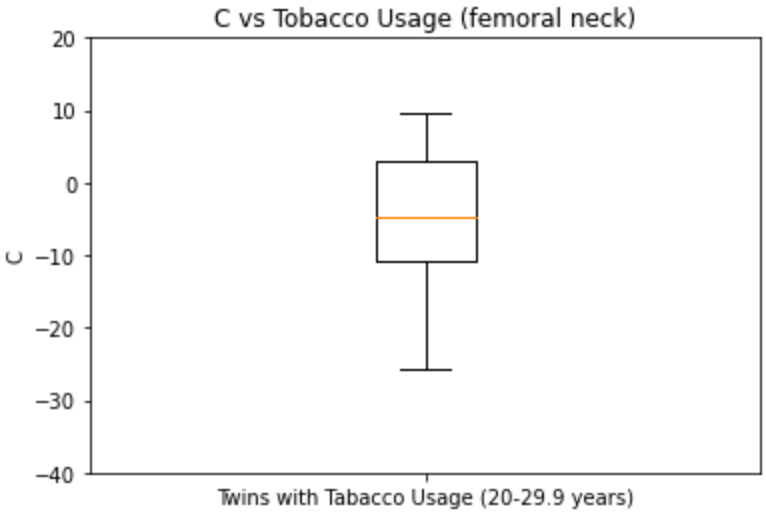
The 5 boxplots above represent the C values between twins with different tobacco usage for the lumbar spine. Based on the data, we can conclude that higher tobacco usage results in a lower C value. This correlation is consistent across all graphs for the lumbar spine data.

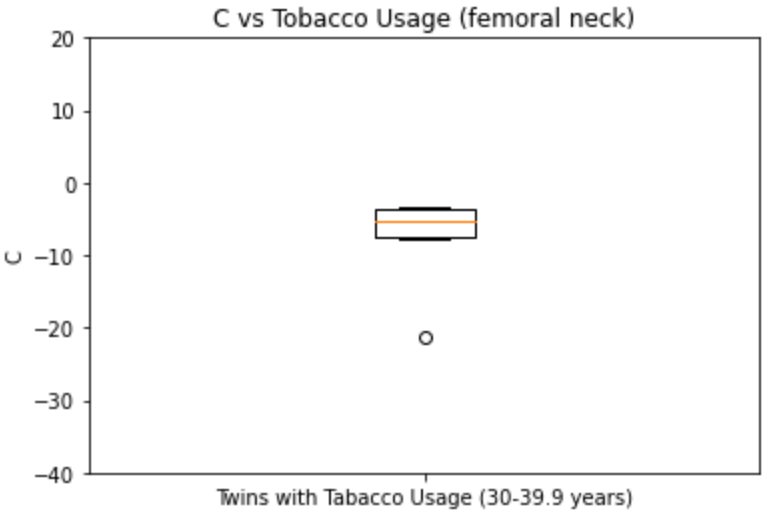
Femoral neck data:

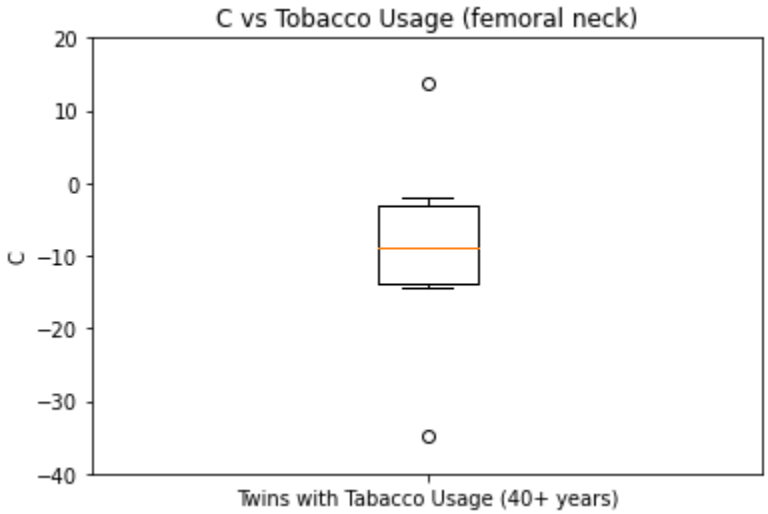
```
# FEMORAL NECK
plt.boxplot(bin1_1.C2, labels=['Twins with Tabacco Usage (0-9.9 years)'])
plt.ylim(-40,20)
plt.ylabel('C')
plt.title('C vs Tobacco Usage (femoral neck)')
plt.show()

plt.boxplot(bin2_1.C2, labels=['Twins with Tabacco Usage (10-19.9 years)'])
plt.ylim(-40,20)
plt.ylabel('C')
plt.title('C vs Tobacco Usage (femoral neck)')
plt.show()

plt.boxplot(bin3_1.C2, labels=['Twins with Tabacco Usage (20-29.9 years)'])
plt.ylim(-40,20)
plt.ylabel('C')
plt.title('C vs Tobacco Usage (femoral neck)')
plt.show()

plt.boxplot(bin4_1.C2, labels=['Twins with Tabacco Usage (30-39.9 years)'])
plt.ylim(-40,20)
plt.ylabel('C')
plt.title('C vs Tobacco Usage (femoral neck)')
plt.show()

plt.boxplot(bin5_1.C2, labels=['Twins with Tabacco Usage (40+ years)'])
plt.ylim(-40,20)
plt.ylabel('C')
plt.title('C vs Tobacco Usage (femoral neck)')
plt.show()
```







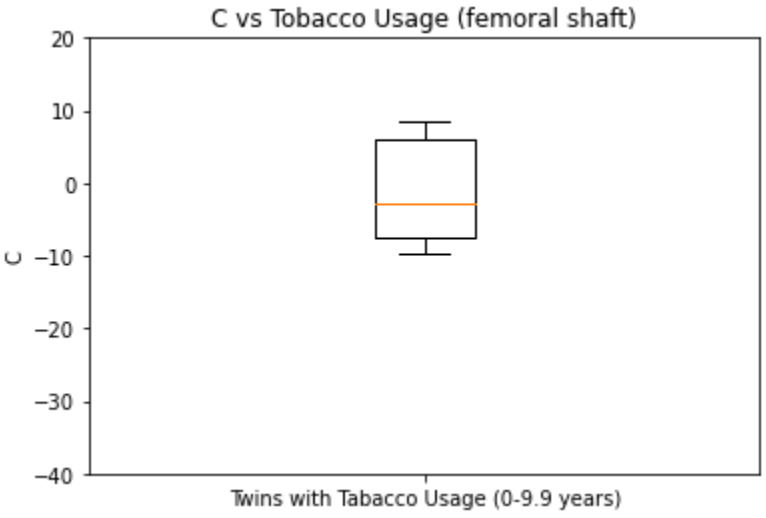


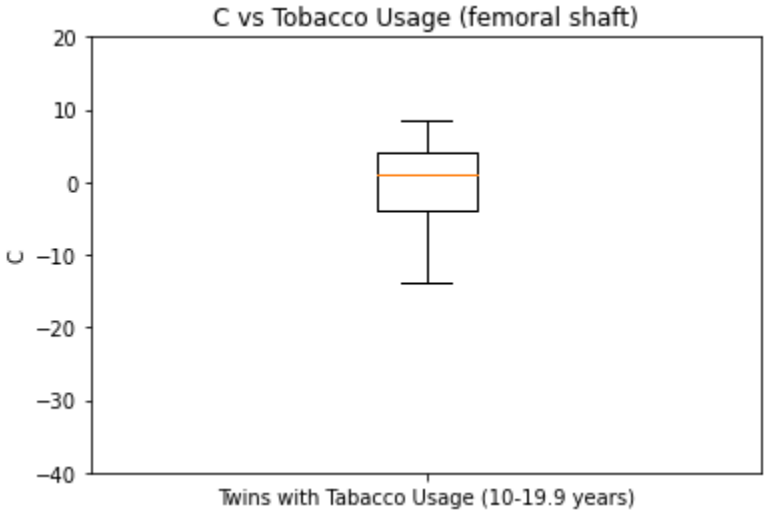


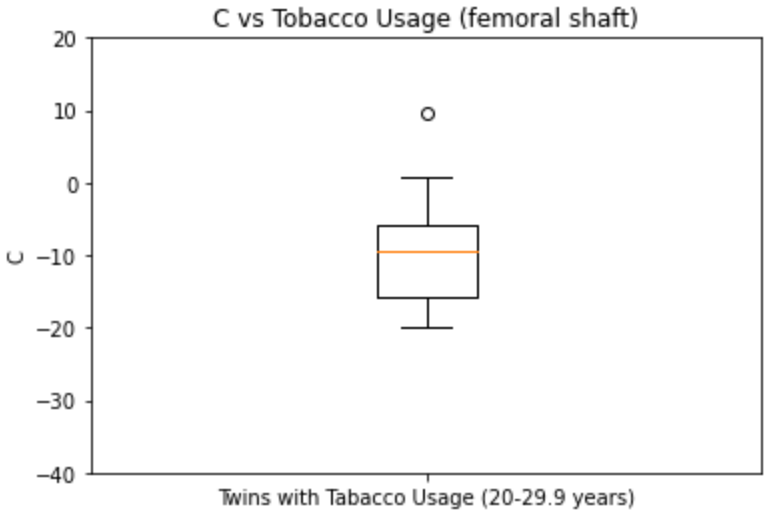
The 5 boxplots above represent the C values between twins with different tobacco usage for the femoral neck. Like with the lumbar spine data, we are able to see a correlation between higher tobacco usage and a lower C value.

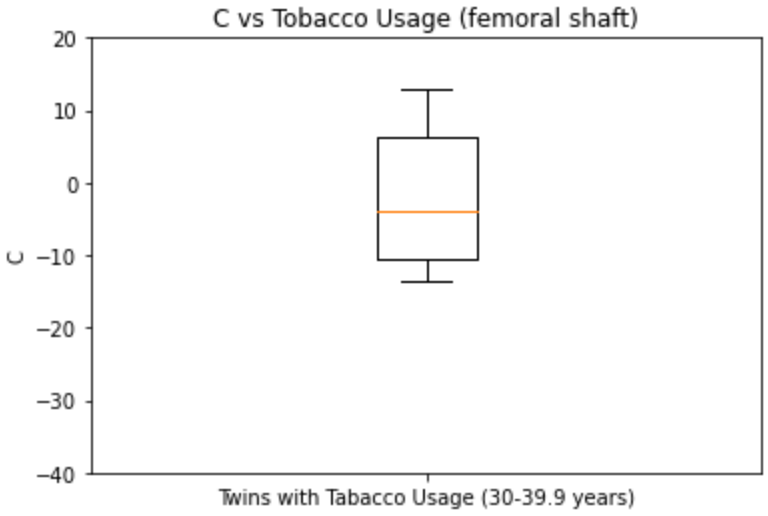
Femoral shaft data:

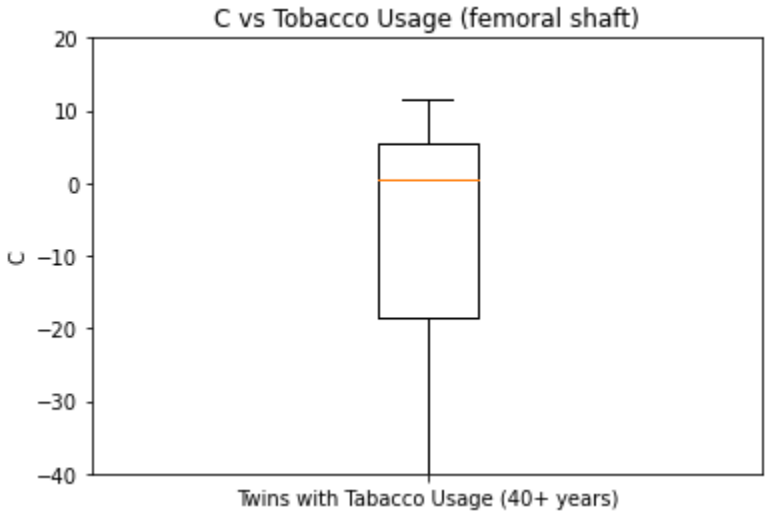
```
# FEMORAL SHAFT
plt.boxplot(bin1_1.C3, labels=['Twins with Tabacco Usage (0-9.9 years)'])
plt.ylim(-40,20)
plt.ylabel('C')
plt.title('C vs Tobacco Usage (femoral shaft)')
plt.show()

plt.boxplot(bin2_1.C3, labels=['Twins with Tabacco Usage (10-19.9 years)'])
plt.ylim(-40,20)
plt.ylabel('C')
plt.title('C vs Tobacco Usage (femoral shaft)')
plt.show()

plt.boxplot(bin3_1.C3, labels=['Twins with Tabacco Usage (20-29.9 years)'])
plt.ylim(-40,20)
plt.ylabel('C')
plt.title('C vs Tobacco Usage (femoral shaft)')
plt.show()

plt.boxplot(bin4_1.C3, labels=['Twins with Tabacco Usage (30-39.9 years)'])
plt.ylim(-40,20)
plt.ylabel('C')
plt.title('C vs Tobacco Usage (femoral shaft)')
plt.show()

plt.boxplot(bin5_1.C3, labels=['Twins with Tabacco Usage (40+ years)'])
plt.ylim(-40,20)
plt.ylabel('C')
plt.title('C vs Tobacco Usage (femoral shaft)')
plt.show()
```











These boxplots are very unlike the previous sets of data. As we can see from the five boxplots, the data doesn't seem to be following the same correlation as the previous two sets of data. The 20-29.9 tobacco usage group also contains an outlier, which is unlike previous data as well. This makes me think that tobacco usage has less of an effect on the C value for this specific bone than it does for the other bones. This could be due to the fact that a femur is the strongest bone in the body, so it might not be easily affected by outside influences.

Thoughts:

This case study was much harder for me to interpret. I think this could've been due to the fact that we were comparing data of a set of twins who had different variables such as tobacco usage. Based on the boxplots, there appeared to be a negative correlation between C and tobacco usage, but it was much less apparent than case study 1. The femoral shaft in particular appeared to have less of an effect from tobacco usage compared to the other two bones.