ELEC 390 Lab 4

Thursday, March 9th 2023

Section 3

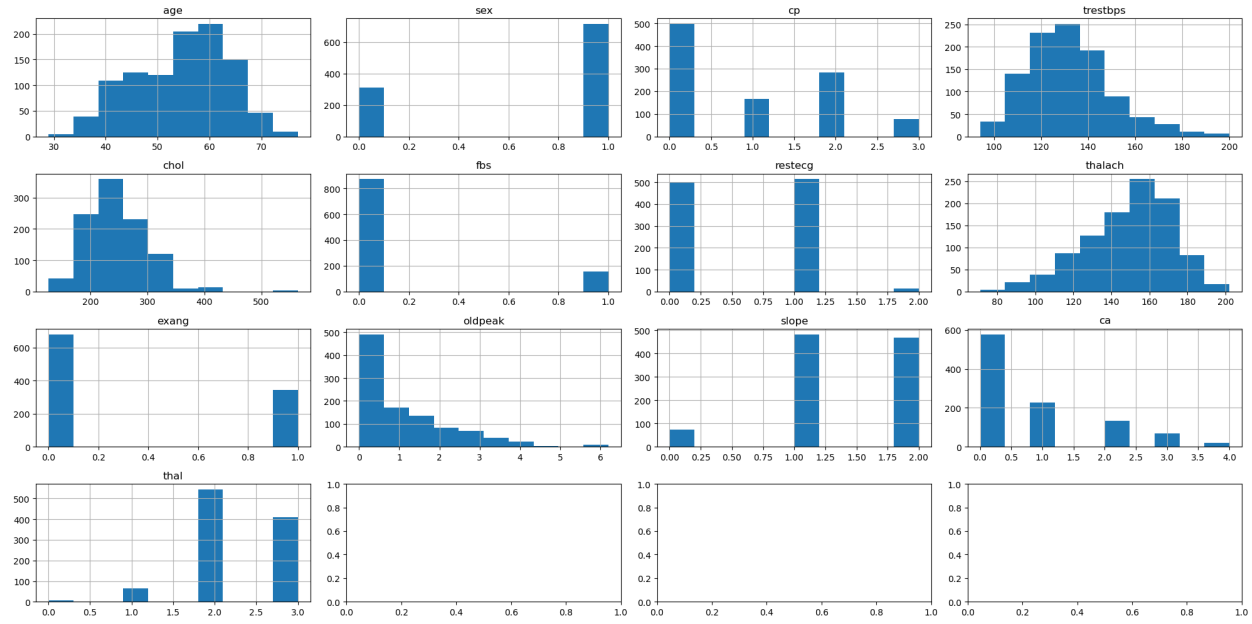
Charlotte Lombard (20232888)

Liam Salass (20229595)

Mile Stosic (20233349)

Question 1

```
# Question 1

dataset = pd.read_csv('heart.csv')
labels = dataset.iloc[:,13]
data = dataset.iloc[:,:13]

print (labels)
print (data)

fig, ax = plt.subplots(ncols=4, nrows=4, figsize= (20,10))

data.hist(ax=ax.flatten()[:13])
fig.tight_layout()

plt.show()
```

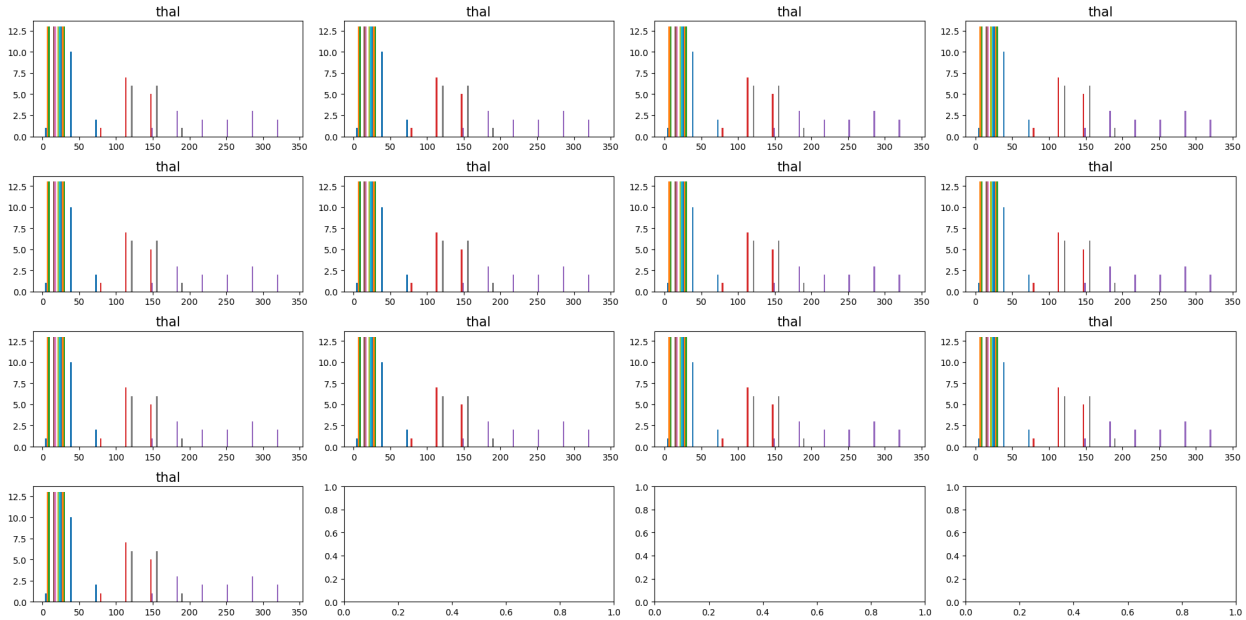


Question 2

a) The majority of the patients are older than 40.
b) The age with the highest likelihood of being selected is 60.
c) From the ‘chol’ histogram, we can infer that most patients have a cholesterol level between 150 and 350.
d) The binary features in the Heart Disease Dataset are sex, fbs, and exang

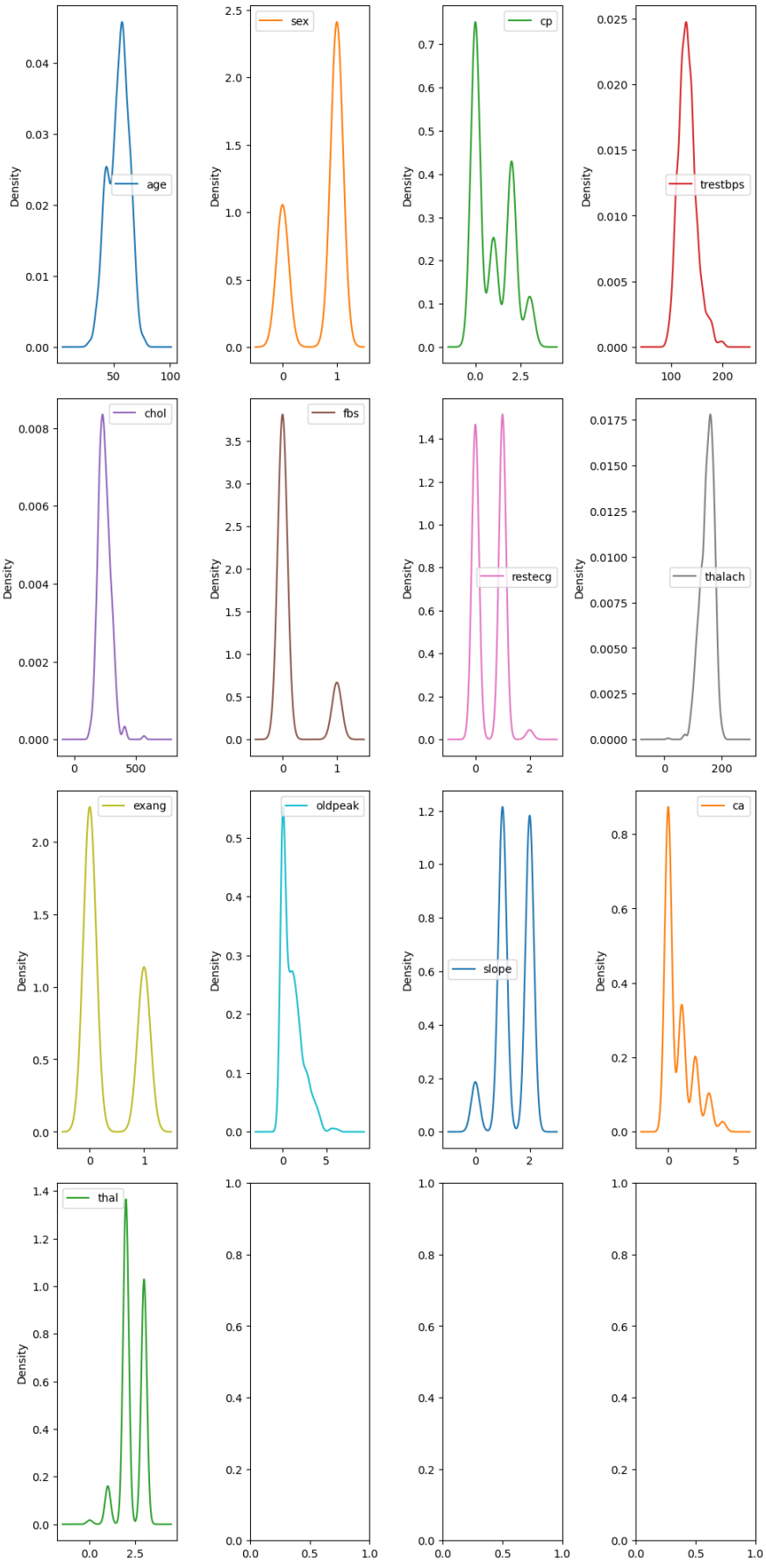
Question 3

```
# Question 3
fig, ax = plt.subplots(ncols=4, nrows=4, figsize= (20,10))
for i in range(13):
    ax.flatten()[i].hist(data.iloc[:13])
    ax.flatten()[i].set_title(data.columns.all(),fontsize = 15)

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



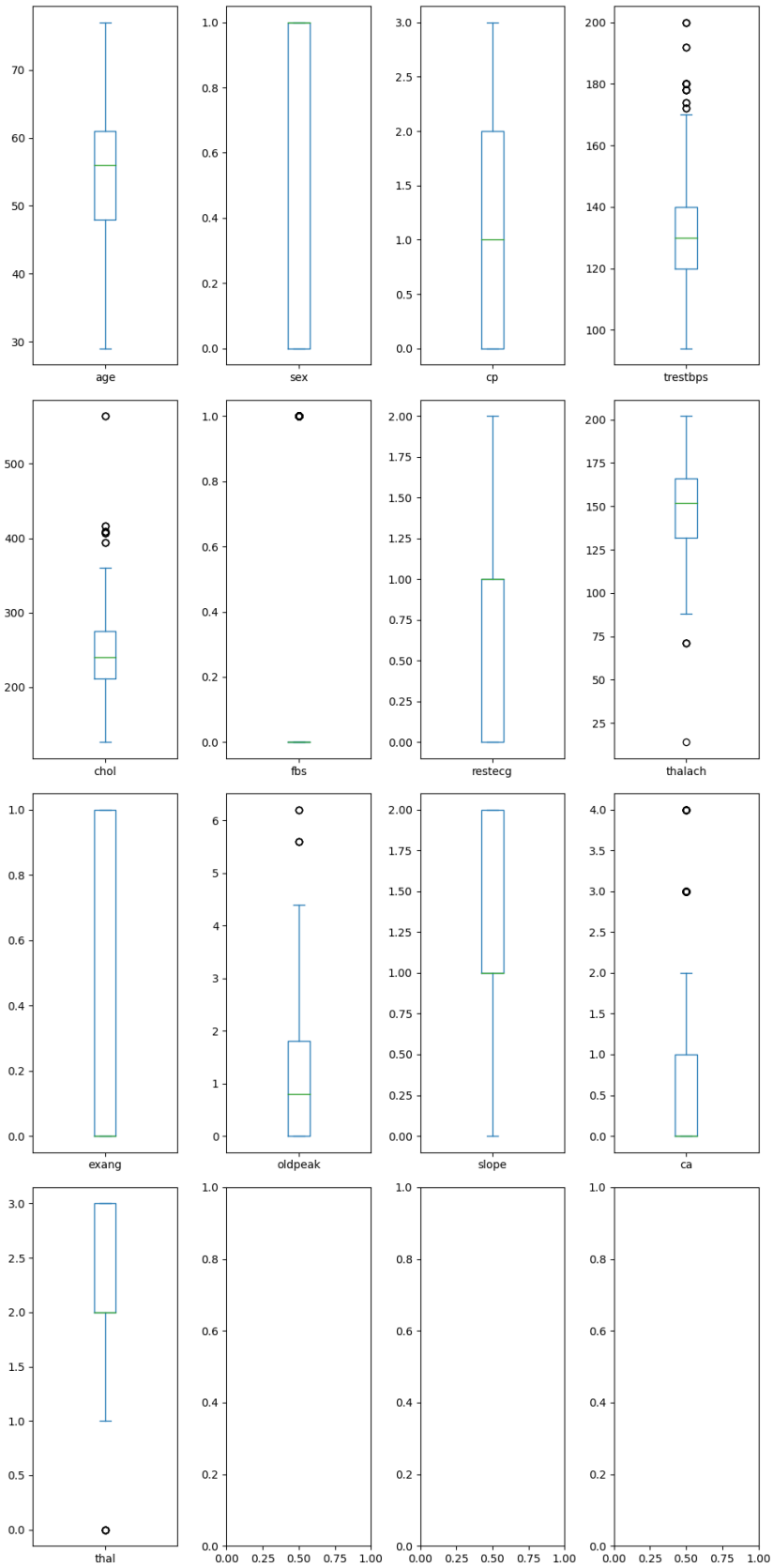
Question 4

```
# Question 4
dataset = pd.read_csv('heart.csv')
labels = dataset.iloc[:,13]
data = dataset.iloc[:,:13]

fig, ax = plt.subplots(ncols=4, nrows=4, figsize=(10,20))

data.plot(ax=ax.flatten()[:13], kind='density', subplots=True, sharex=False)
fig.tight_layout()
plt.show()
```


age
sex
cp
trestbps
chol
fbs
restecg
thalach
exang
oldpeak
slope
ca
thal
Density

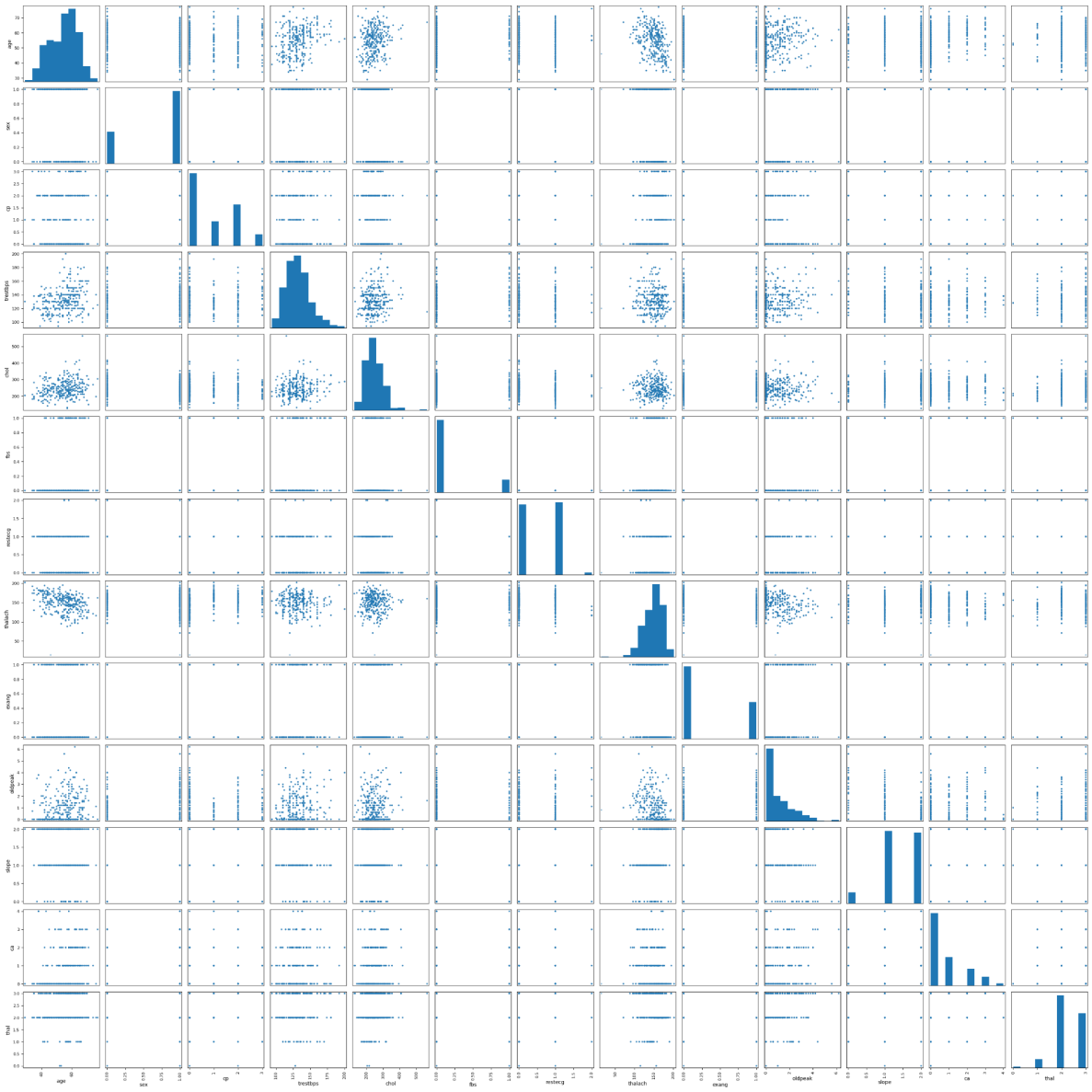
Question 5

```
# Question 5
dataset = pd.read_csv('heart.csv')
labels = dataset.iloc[:,13]
data = dataset.iloc[:,:13]

fig, ax = plt.subplots(ncols=4, nrows=4, figsize=(10,20))

data.plot(ax=ax.flatten()[:13], kind='box', subplots=True, sharex=False, sharey=False)
fig.tight_layout()
plt.show()
```


age
sex
cp
trestbps
chol
fbs
restecg
thalach
exang
oldpeak
slope
ca
thal

Question 6

```
# Question 6
dataset = pd.read_csv('heart.csv')
labels = dataset.iloc[:,13]
data = dataset.iloc[:,:13]

fig, ax = plt.subplots(ncols=13, nrows=13, figsize=(30,30))

pd.plotting.scatter_matrix(data, ax=ax)

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



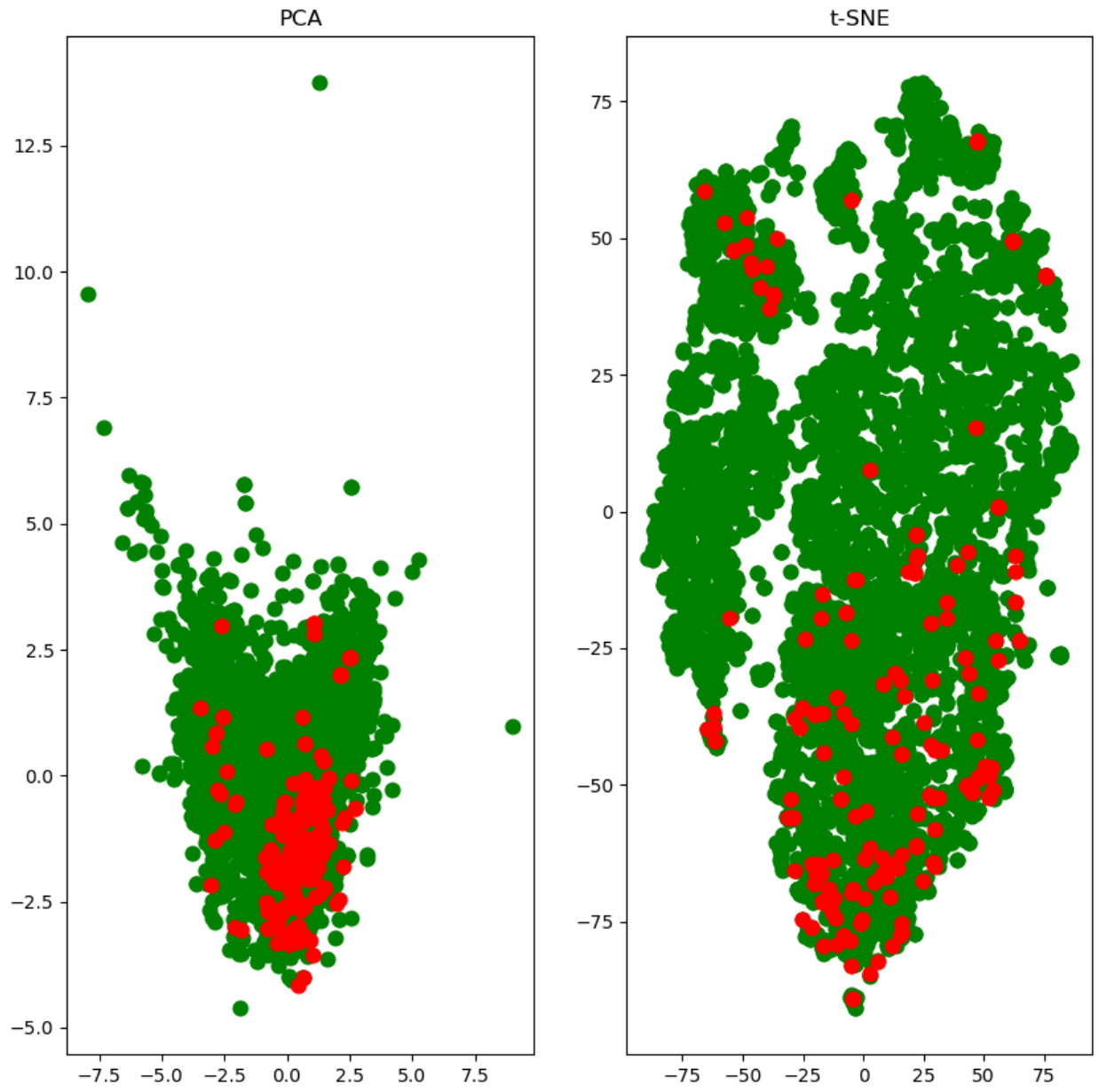
Question 7

A. The scatter plot between thalach and age shows a negative correlation. As age increases, the maximum heart rate tends to decrease. Therefore, thalach and age have a negative correlation.

B. The scatter plot between thalach and chol does not show a strong correlation. There is a weak positive correlation between the two variables, which means that as one variable increases, the other variable tends to increase slightly, but the correlation is not strong enough to make any strong conclusions. Therefore, thalach and chol do not have a strong correlation.

Question 8

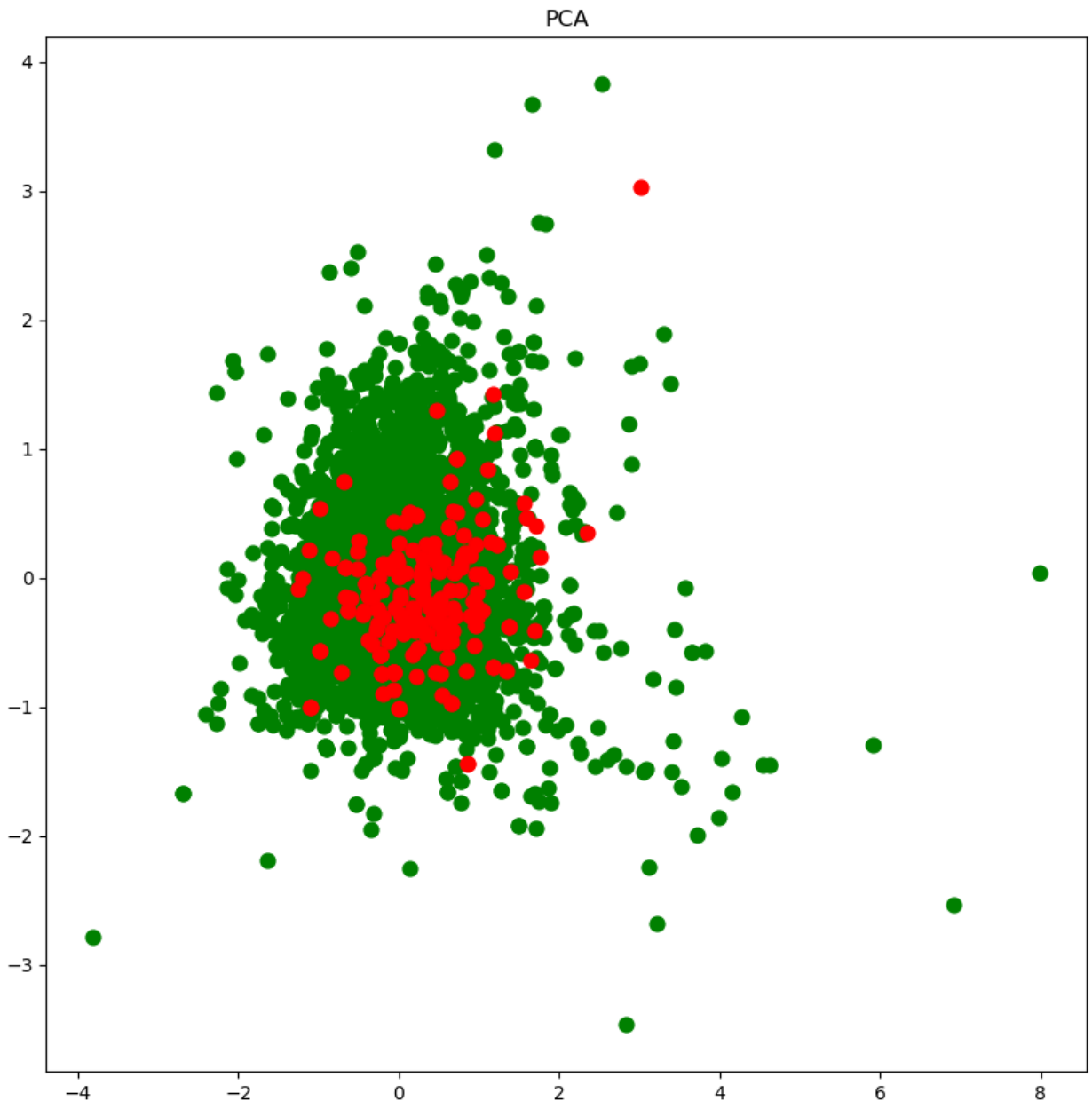
```
from sklearn.decomposition import PCA
from sklearn.manifold import TSNE
from sklearn.preprocessing import StandardScaler
data_set = pd.read_csv('winequalityN.csv')

labels = data_set.iloc[:,12]
data = data_set.iloc[:,1:12]

for i in range(len(labels)):
    if labels[i] >7:
        labels[i] = 1
    else:
        labels[i] = 0

pca = PCA(n_components=2)
tsne = TSNE(n_components=2, perplexity=30)
sc = StandardScaler()

data = sc.fit_transform(data)
pca_data = pca.fit_transform(data)
tsne_data = tsne.fit_transform(data)

fig, (ax1,ax2) = plt.subplots(1, 2, figsize=(10,10))
colors = ['green','red']
my_legends = [0, 1]

for i in range(len(my_legends)):
    ax1.scatter(pca_data[labels == i, 0], pca_data[labels == i, 1],
c=colors[i], s =60)
    ax2.scatter(tsne_data[labels == i, 0], tsne_data[labels == i, 1],
c=colors[i], s =60)

ax1.set_title('PCA')
ax2.set_title('t-SNE')
plt.show()
```



Question 9

```
data_set = pd.read_csv('winequalityN.csv')

labels = data_set.iloc[:,12]
data = data_set.iloc[:,1:12]

for i in range(len(labels)):
    if labels[i] >7:
        labels[i] = 1
    else:
```

```
        labels[i] = 0

pca = PCA(n_components=11)

sc = StandardScaler()

data = sc.fit_transform(data)
pca_data = pca.fit_transform(data)



fig, ax = plt.subplots(figsize=(10,10))
colors = ['green','red']
my_legends = [0, 1]

for i in range(len(my_legends)):
    ax.scatter(pca_data[labels == i, 7], pca_data[labels == i, 8],
c=colors[i], s =60)



ax.set_title('PCA')

plt.show()
```



Question 10

The scenario that carries more information from the wine quality dataset is the scenario in question 8, where we compare the first and second dimensions as they provide more information on the variance in the data set than the 8th and 9th dimensions.