ELEC 390 – Lab 04

Department of Electrical and Computer Engineering
Queen's University

Composed By
Nicholas Seegobin (20246787)
Zeerak Asim (20237955)
Lauren Steel (20218337)
Saman Saeidi (20217992)

Section 03
Date of Submission
2023 March 9th

Question 1)

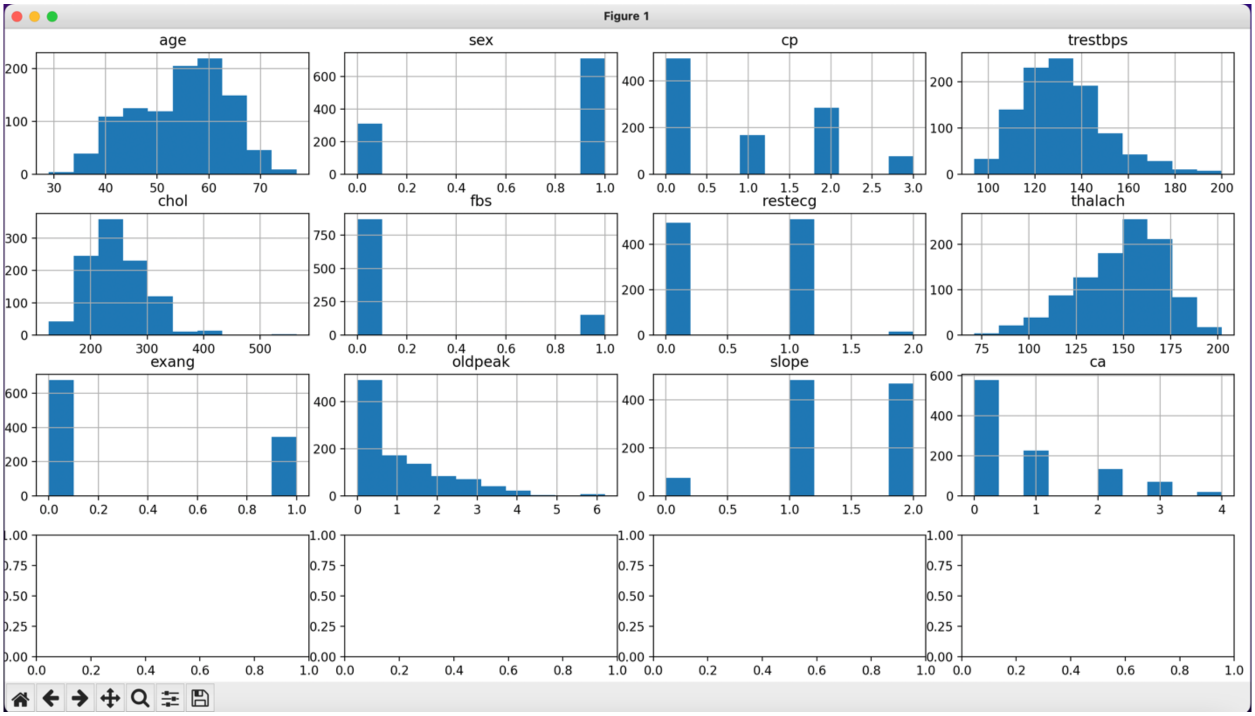
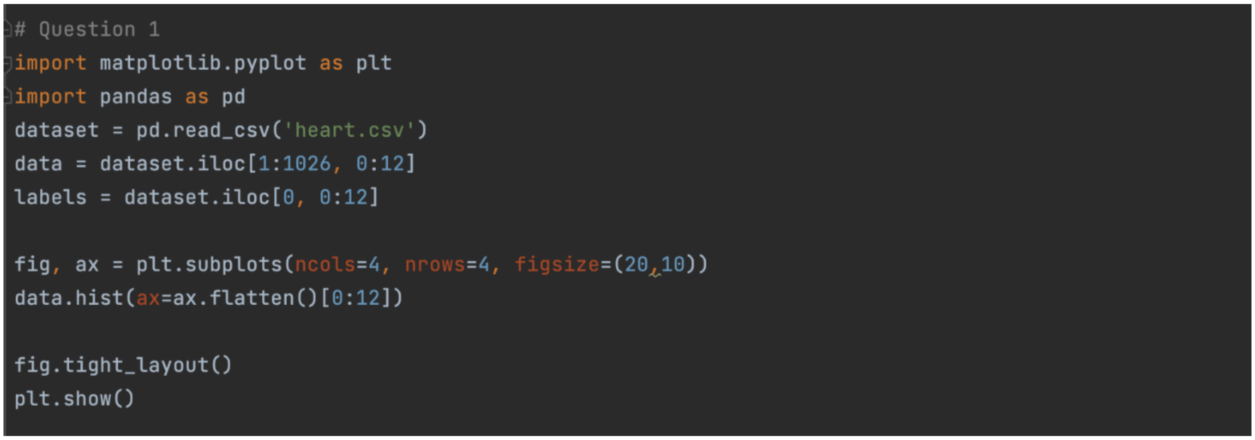


Figure 1: Plots generated by python code for question 1.



```python
# Question 1
import matplotlib.pyplot as plt
import pandas as pd
dataset = pd.read_csv('heart.csv')
data = dataset.iloc[1:1026, 0:12]
labels = dataset.iloc[0, 0:12]

fig, ax = plt.subplots(ncols=4, nrows=4, figsize=(20,10))
data.hist(ax=ax.flatten()[0:12])

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

Figure 2: Python code written for question 1.

Question 2)

A. Based on the histograms created, we can see that the majority of patients are older than 40.
B. The highest likelihood for a patient selected at random is that they are 60 years old.
C. We can infer that the majority of patients have cholesterol in the 250 range and that the distribution is parabolic between 50 and roughly 350-400.
D. The binary features of the Heart Disease Dataset are sex, fbs, and exang.

Question 3)

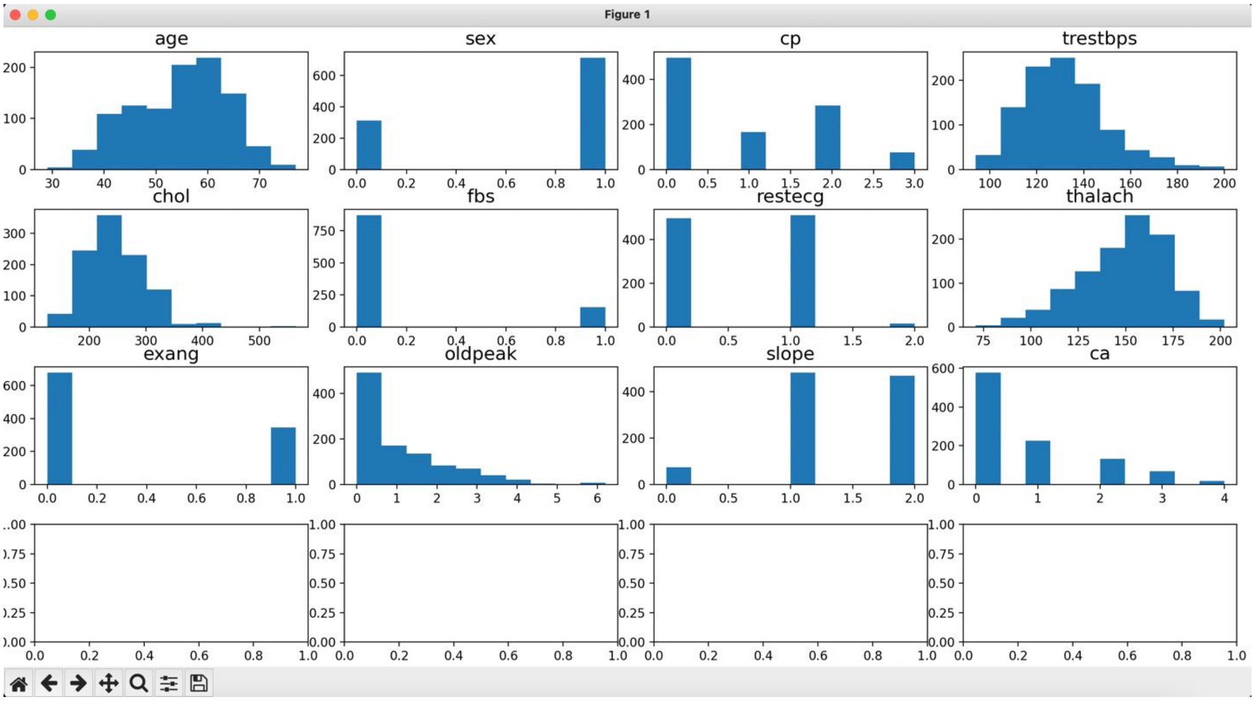
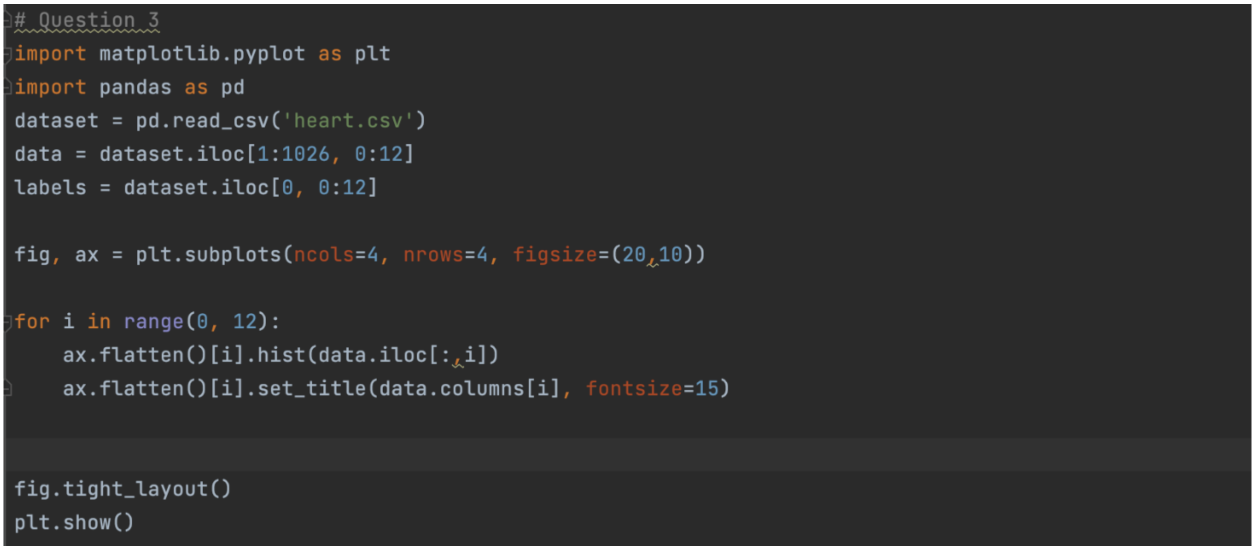


Figure 3: Plots generated by python code for question 3.



```
# Question 3
import matplotlib.pyplot as plt
import pandas as pd
dataset = pd.read_csv('heart.csv')
data = dataset.iloc[1:1026, 0:12]
labels = dataset.iloc[0, 0:12]

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

for i in range(0, 12):
    ax.flatten()[i].hist(data.iloc[:,i])
    ax.flatten()[i].set_title(data.columns[i], fontsize=15)


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

Figure 4: Python code written for question 3.

Question 4)

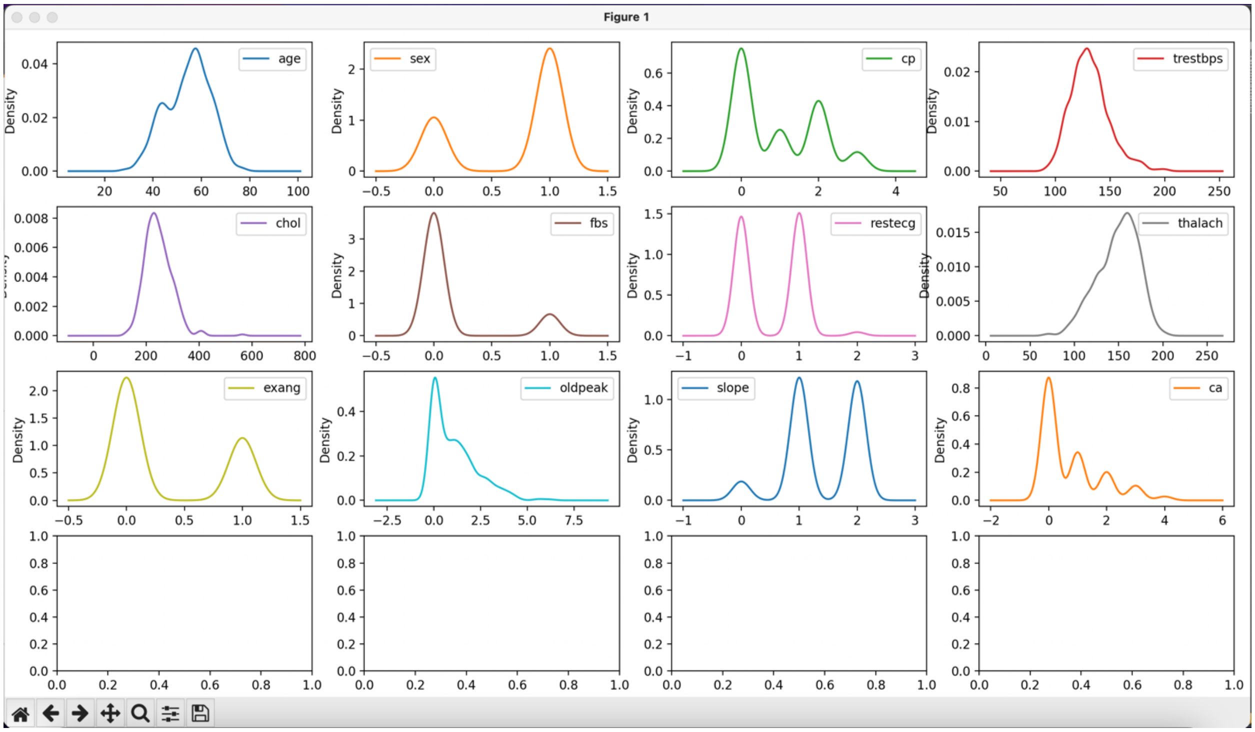
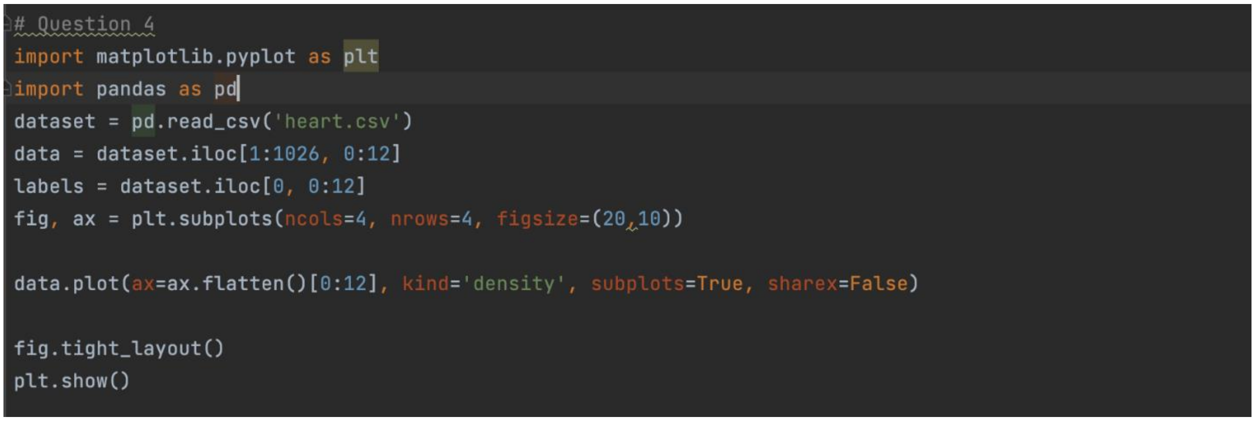


Figure 5: Plots generated by python code for question 4.



```python
# Question 4
import matplotlib.pyplot as plt
import pandas as pd
dataset = pd.read_csv('heart.csv')
data = dataset.iloc[1:1026, 0:12]
labels = dataset.iloc[0, 0:12]
fig, ax = plt.subplots(ncols=4, nrows=4, figsize=(20,10))

data.plot(ax=ax.flatten()[0:12], kind='density', subplots=True, sharex=False)

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

Figure 6: Python code written for question 4.

Question 5)

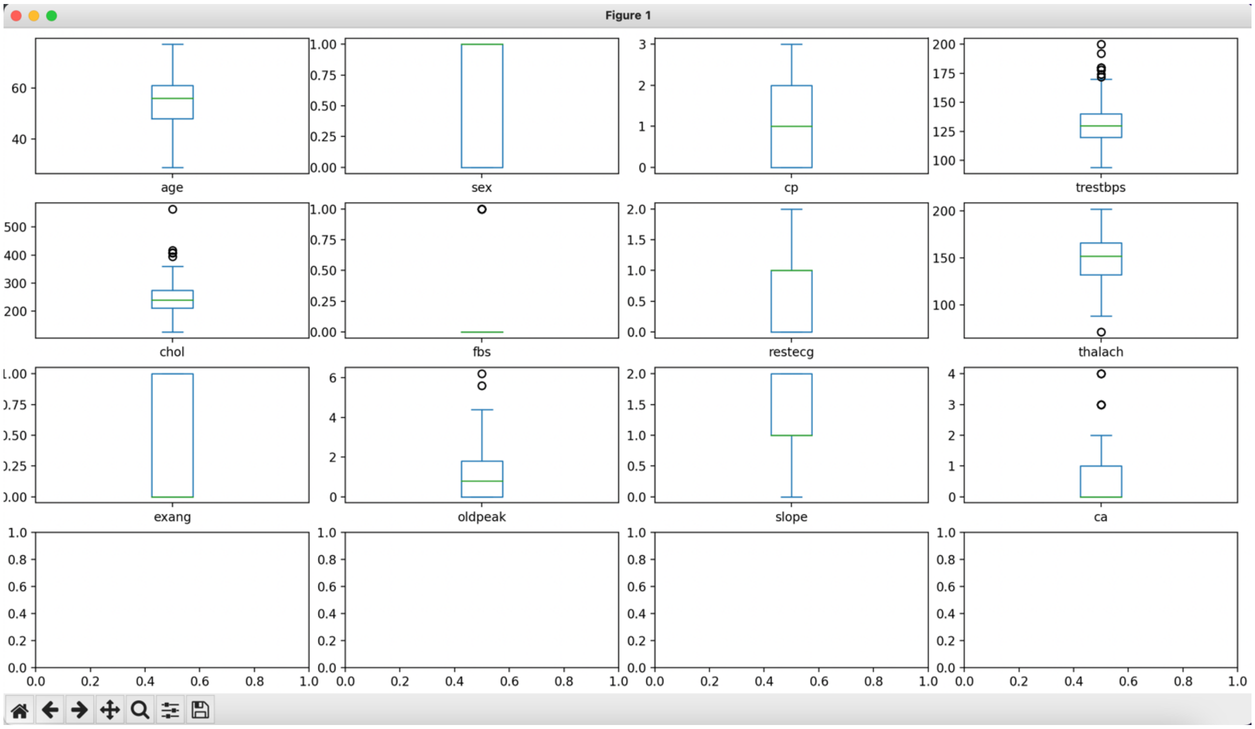
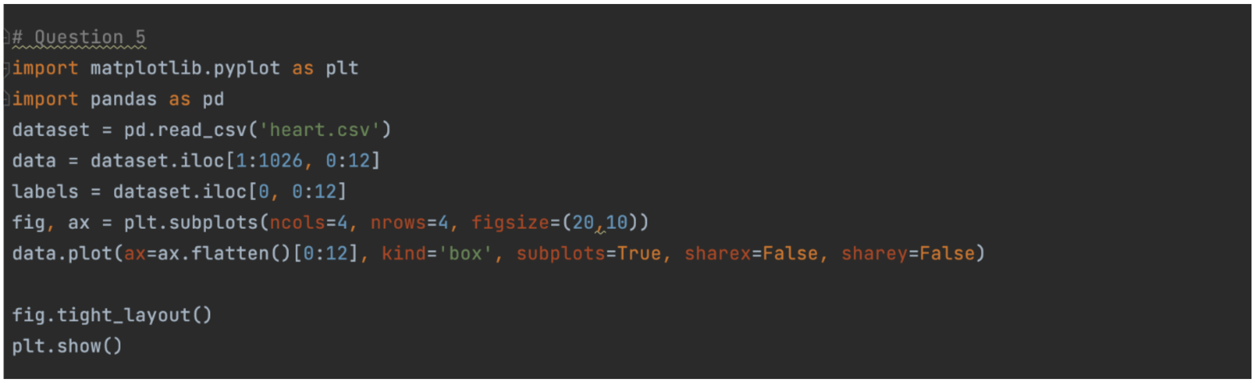


Figure 7: Plots generated by python code for question 5.



```
# Question 5
import matplotlib.pyplot as plt
import pandas as pd
dataset = pd.read_csv('heart.csv')
data = dataset.iloc[1:1026, 0:12]
labels = dataset.iloc[0, 0:12]
fig, ax = plt.subplots(ncols=4, nrows=4, figsize=(20,10))
data.plot(ax=ax.flatten()[0:12], kind='box', subplots=True, sharex=False, sharey=False)

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

Figure 8: Python code written for question 5.

Question 6)

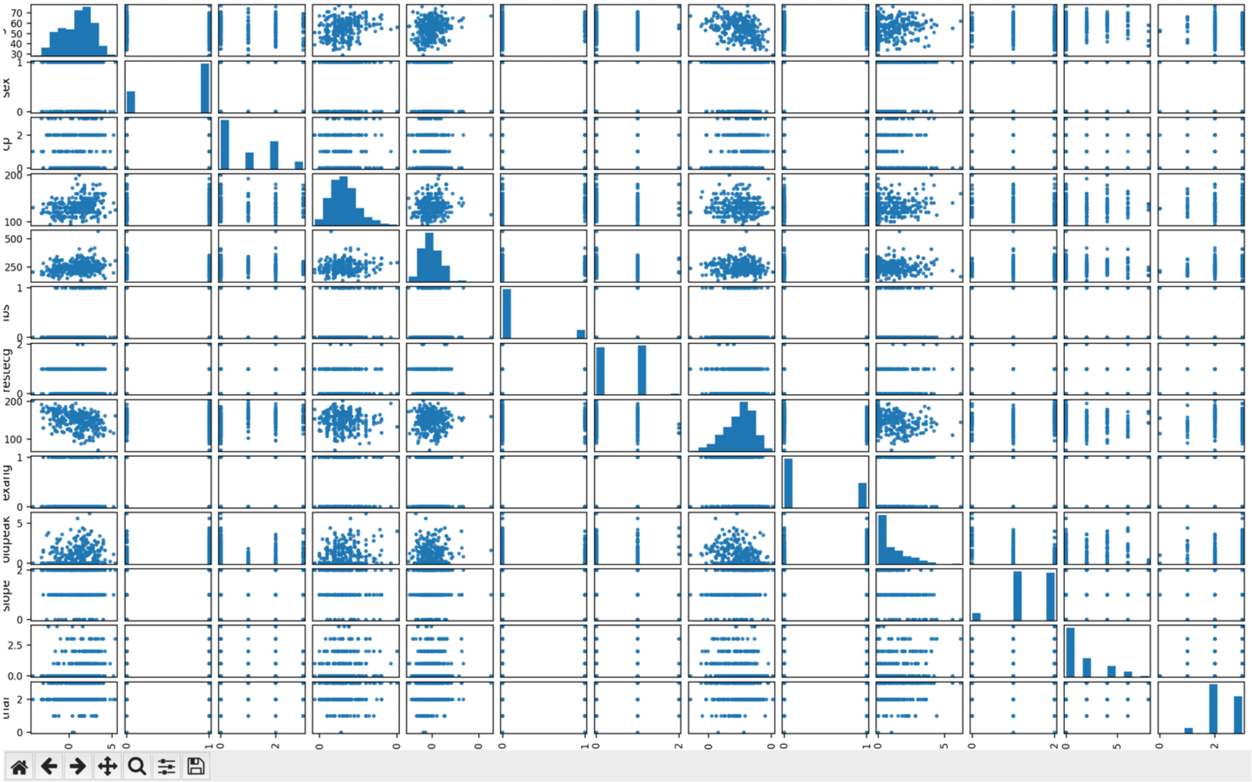
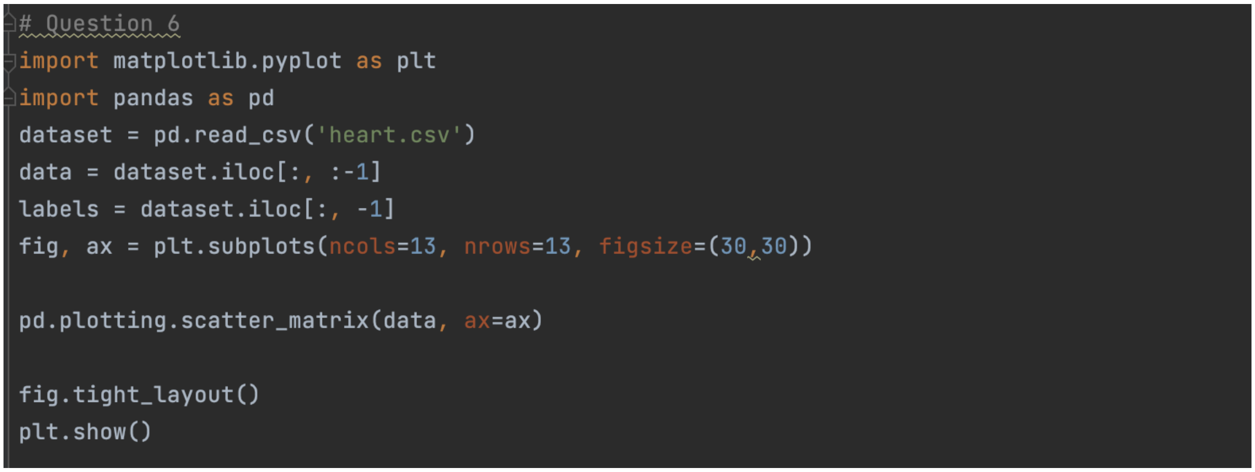


Figure 9: Plots generated by python code for question 6.



```
# Question 6
import matplotlib.pyplot as plt
import pandas as pd
dataset = pd.read_csv('heart.csv')
data = dataset.iloc[:, :-1]
labels = dataset.iloc[:, -1]
fig, ax = plt.subplots(ncols=13, nrows=13, figsize=(30,30))

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

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

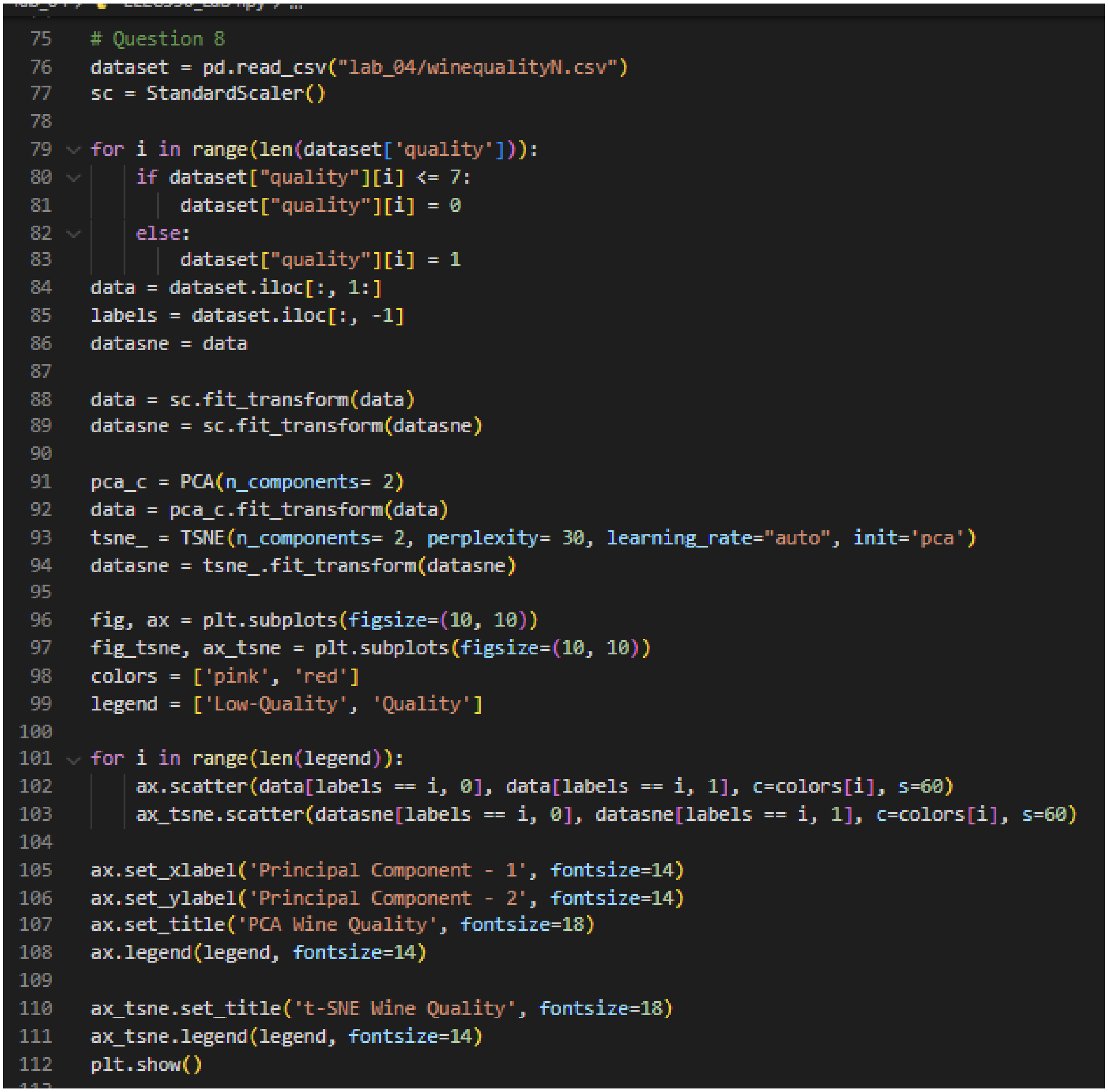
Figure 10: Plots generated by python code for question 6.

Question 7)

A. Based on the scatter matrix plot thalach (max heart rate) and age have a negative correlation. Overall, as age increases, heartrate decreases.

B. Based on the scatter matrix plot thalach and chol do not have strong correlation. As seen in Figure 9 the plots depicting the relation between them demonstrate no significant correlation. There is no visible pattern in their relationship.

Question 8)



```python
# Question 8
dataset = pd.read_csv("lab_04/winequalityN.csv")
sc = StandardScaler()

for i in range(len(dataset['quality'])):
    if dataset["quality"][i] <= 7:
        dataset["quality"][i] = 0
    else:
        dataset["quality"][i] = 1
data = dataset.iloc[:, 1:]
labels = dataset.iloc[:, -1]
datasne = data

data = sc.fit_transform(data)
datasne = sc.fit_transform(datasne)

pca_c = PCA(n_components= 2)
data = pca_c.fit_transform(data)
tsne_ = TSNE(n_components= 2, perplexity= 30, learning_rate="auto", init='pca')
datasne = tsne_.fit_transform(datasne)

fig, ax = plt.subplots(figsize=(10, 10))
fig_tsne, ax_tsne = plt.subplots(figsize=(10, 10))
colors = ['pink', 'red']
legend = ['Low-Quality', 'Quality']

for i in range(len(legend)):
    ax.scatter(data[labels == i, 0], data[labels == i, 1], c=colors[i], s=60)
    ax_tsne.scatter(datasne[labels == i, 0], datasne[labels == i, 1], c=colors[i], s=60)

ax.set_xlabel('Principal Component - 1', fontsize=14)
ax.set_ylabel('Principal Component - 2', fontsize=14)
ax.set_title('PCA Wine Quality', fontsize=18)
ax.legend(legend, fontsize=14)

ax_tsne.set_title('t-SNE Wine Quality', fontsize=18)
ax_tsne.legend(legend, fontsize=14)
plt.show()
```

Figure 11: Python code written for question 8.

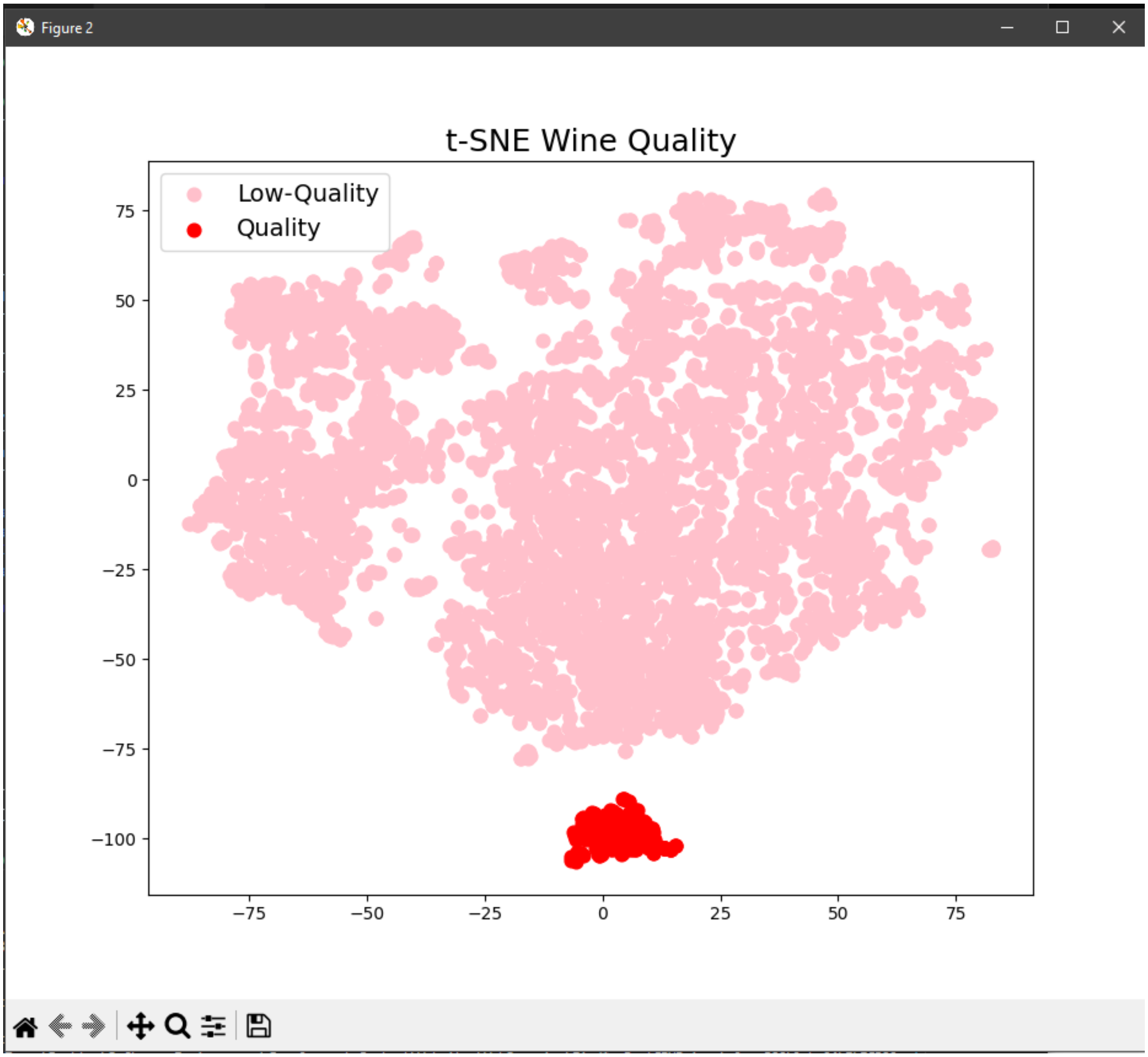


Figure 12: Scatter plot for t-SNE wine quality, generated by python code for question 8.

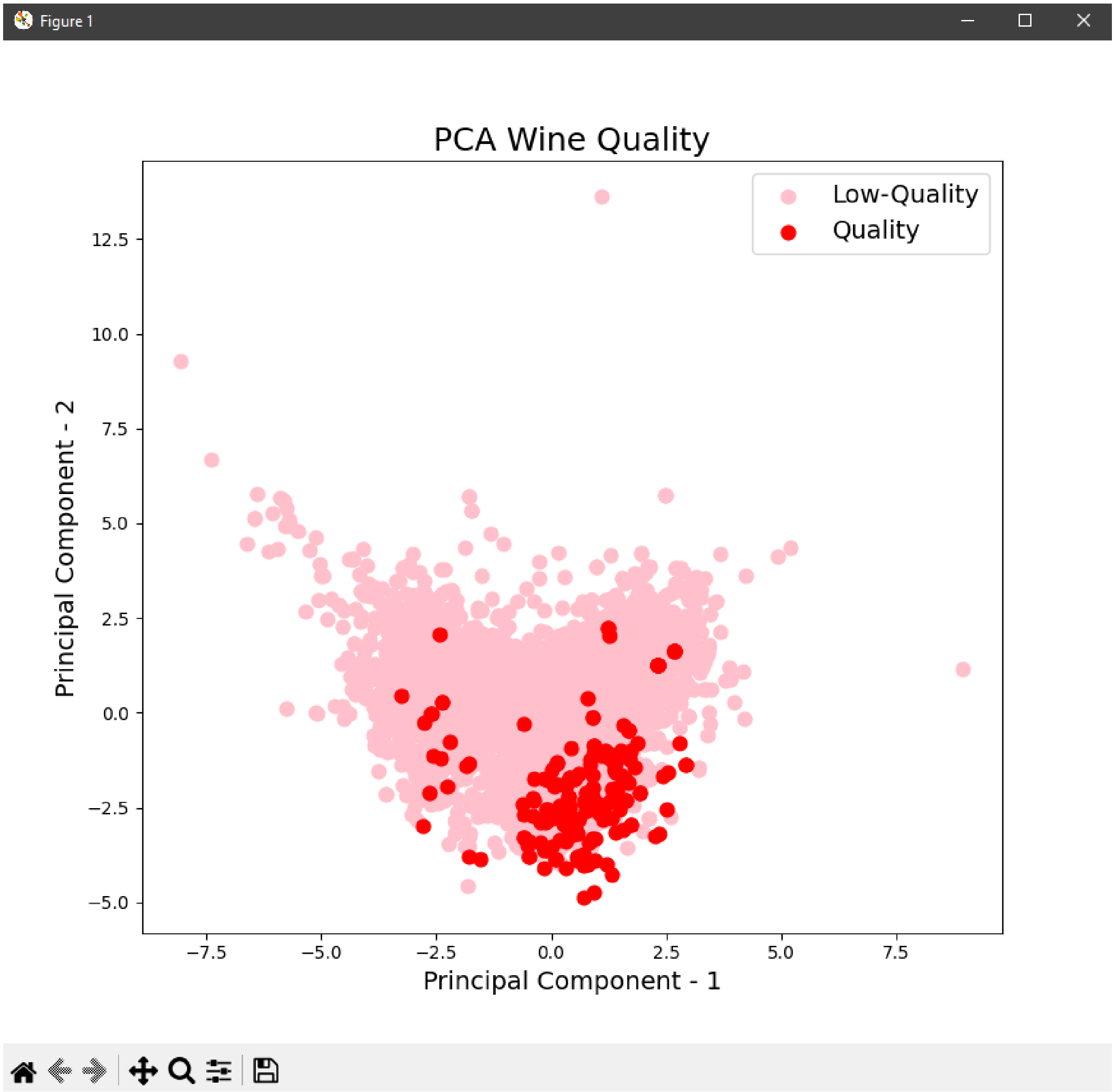
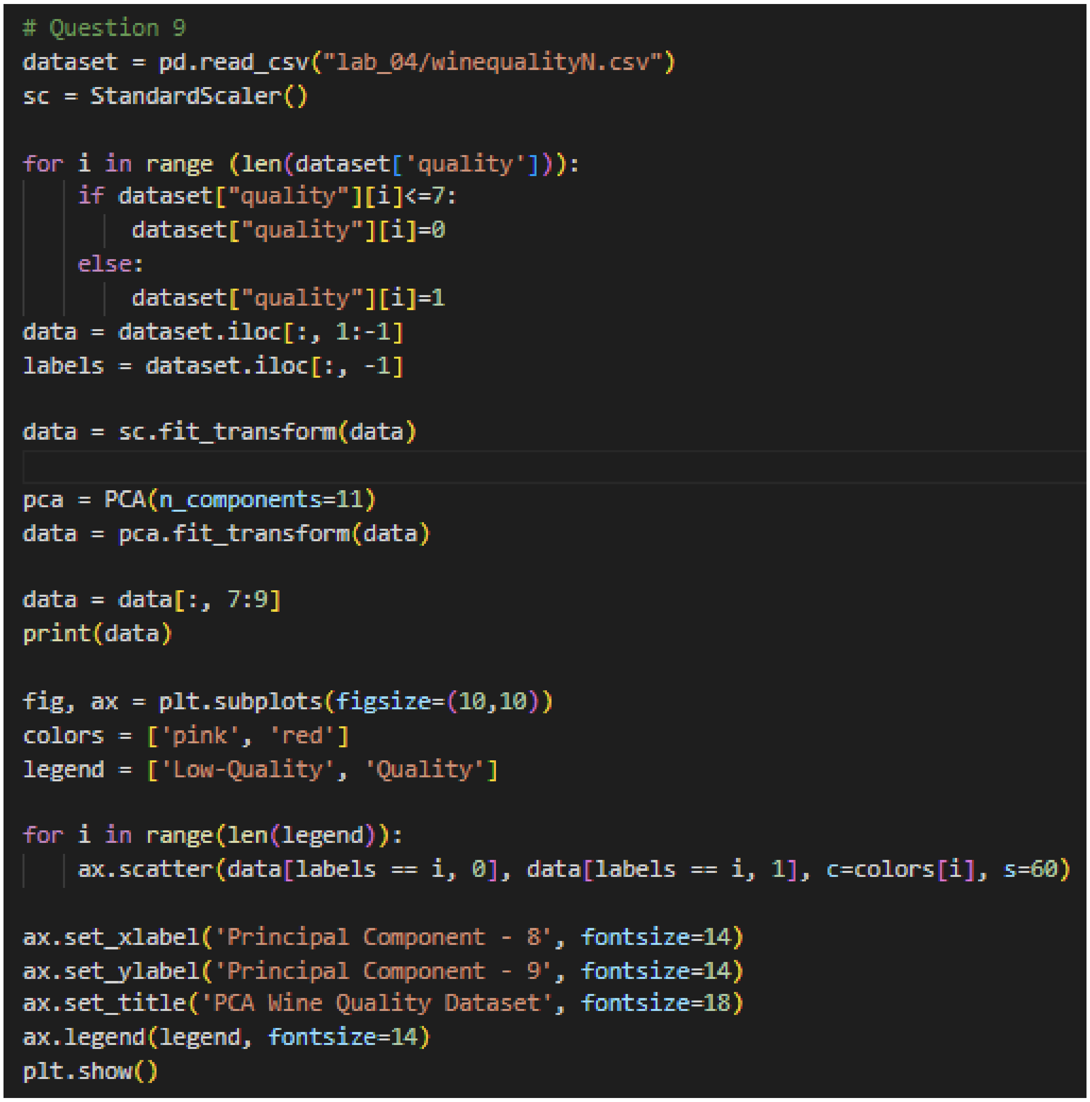


Figure 13: Scatter plot for PCA wine quality, generated by python code for question 8.

Question 9)



```python
# Question 9
dataset = pd.read_csv("lab_04/winequalityN.csv")
sc = StandardScaler()

for i in range (len(dataset['quality'])):
    if dataset["quality"][i]<=7:
        dataset["quality"][i]=0
    else:
        dataset["quality"][i]=1
data = dataset.iloc[:, 1:-1]
labels = dataset.iloc[:, -1]

data = sc.fit_transform(data)

pca = PCA(n_components=11)
data = pca.fit_transform(data)

data = data[:, 7:9]
print(data)

fig, ax = plt.subplots(figsize=(10,10))
colors = ['pink', 'red']
legend = ['Low-Quality', 'Quality']

for i in range(len(legend)):
    ax.scatter(data[labels == i, 0], data[labels == i, 1], c=colors[i], s=60)

ax.set_xlabel('Principal Component - 8', fontsize=14)
ax.set_ylabel('Principal Component - 9', fontsize=14)
ax.set_title('PCA Wine Quality Dataset', fontsize=18)
ax.legend(legend, fontsize=14)
plt.show()
```

Figure 14: Python code written for question 9.

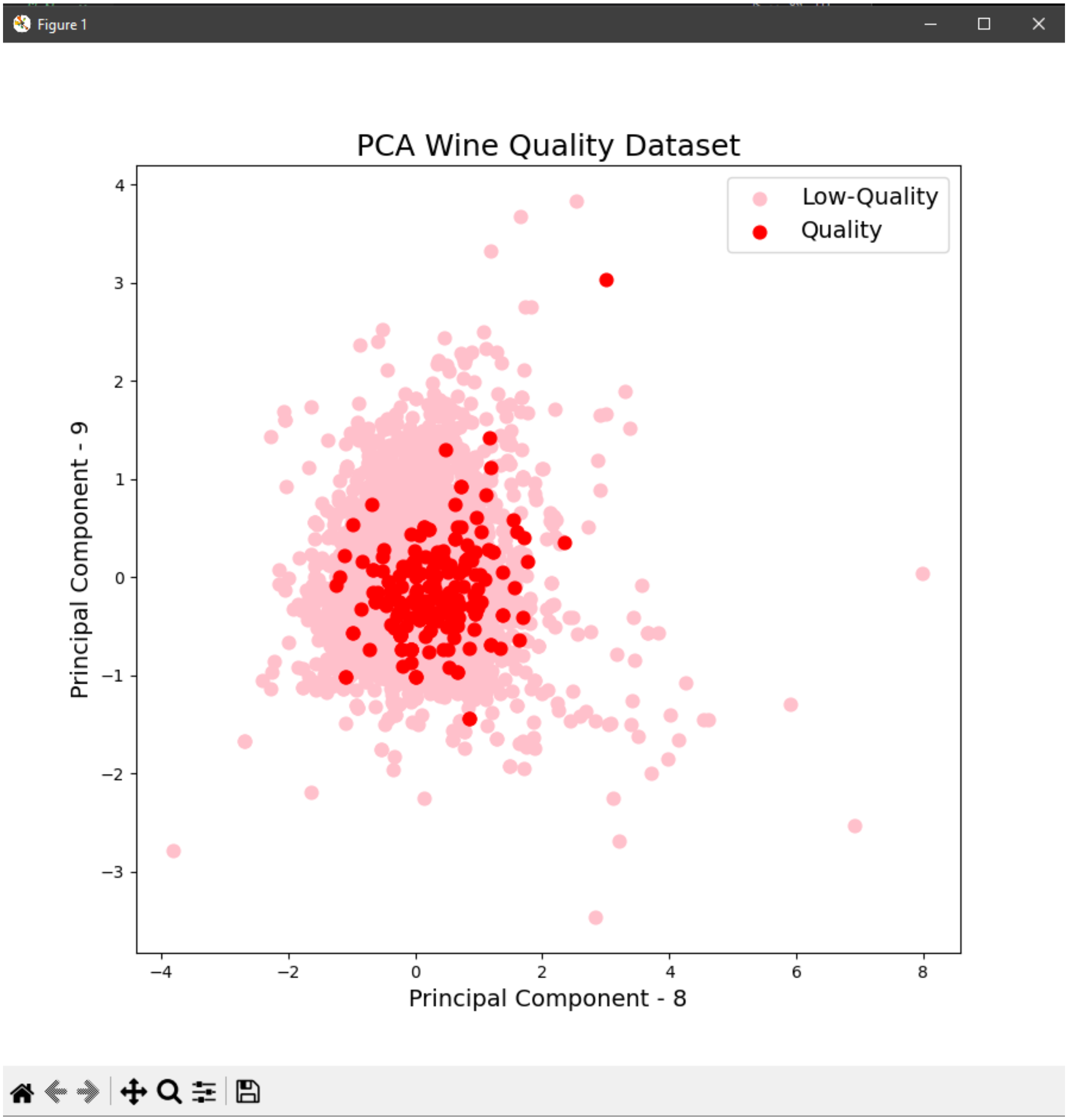


Figure 15: Scatter plot for the PCA wine quality dataset, generated by python code for question 9.

Question 10)

The difference in information between questions 8 and 9 is that question 8 uses all of the datasets versus a small portion of the dataset used in question 9. Generally, in question 8, it would be a much better fit since PCA components 1 and 2 were used, and more of the dataset was used, making it more accurate.