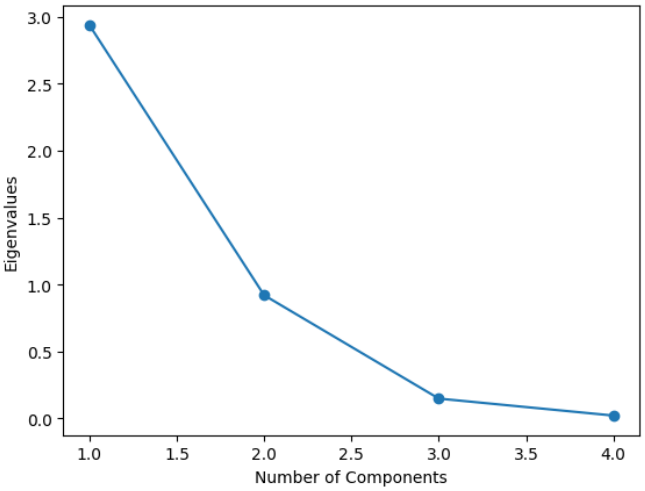
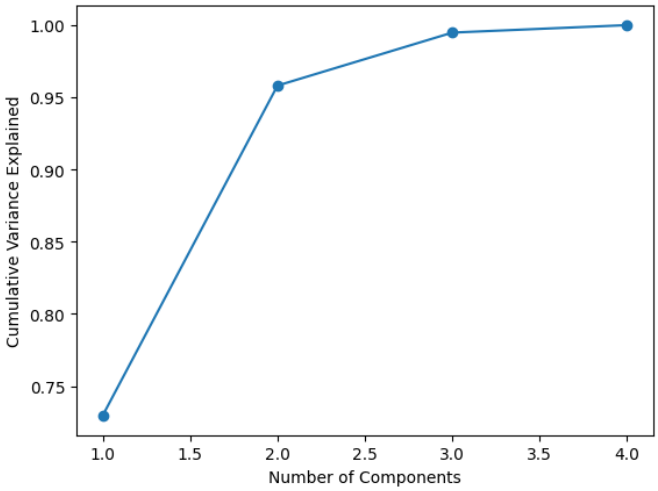
Assignment 2

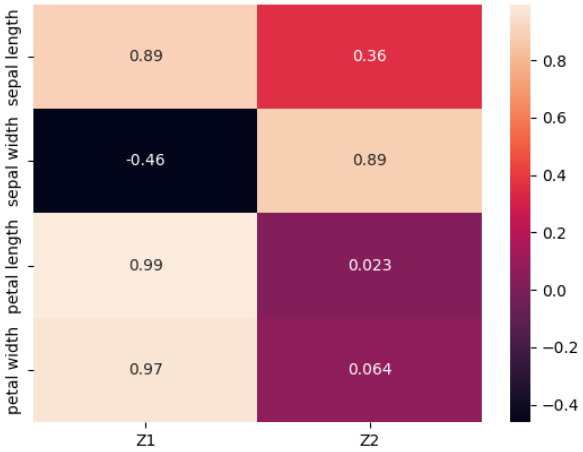
a)





In the first image, we can see that eigenvalue for component 1 is close to 3 and eigenvalue for component 2 is close 1, while eigenvalues for other components are very small. On the right image, we can also observe that the first two component account for most of the variance, close to 95%. So based on that, we can say 2 components are enough to adequately describe the data.

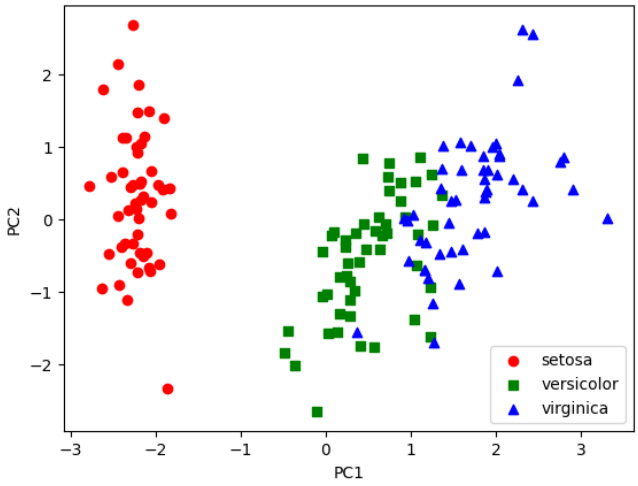
Below, we can see the correlation between principal components and original variables:

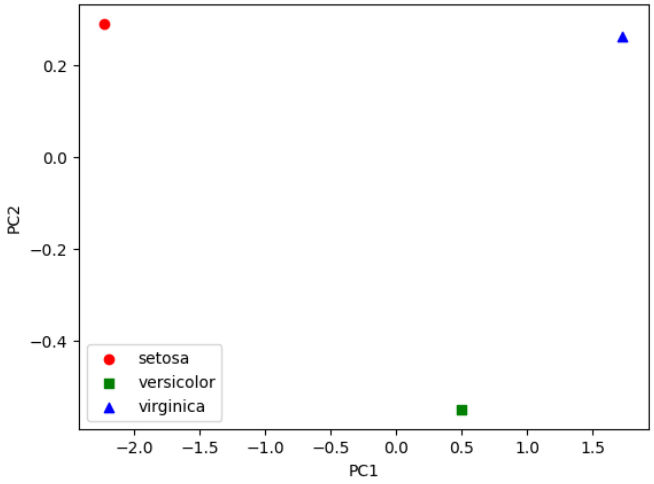


Z1 has high positive loadings on sepal length (0.89) and petal length (0.99) and a high negative loading on sepal width (-0.46), while also having a high positive loading on petal width (0.97). The fact that Z1 has a high positive loading with petal width makes interpretation difficult, but we can generally say that Z1 represents a combination of features related to the length of the flower parts.

On the other hand, Z2 has high positive loadings on sepal width (0.89) and a very small positive loading on petal length (0.023) and petal width (0.064), while having a moderate positive loading on sepal length (0.36). Again, the fact that Z2 has a very low correlation with petal width makes interpretation difficult, but we can generally say that Z2 represents a combination of features related to the width of the flower parts.

b)





We were required to build the graph on the right, which depicts the average principal component scores for each of the different types of iris for the first wo principal components. In addition to that, we also decided to include the image on the left, which is also very useful since it allows us to see how all the data points are organized in reduced feature space. In that image, we can see that 3 clusters can be said to be well separated, although clusters of **viriginica** and **versicolor** iris have some overlapping samples.

By looking on the image on the right, we can say that average principal component scores for each of the different types of irises are very different from each other. Since we interpreted PC1 as generally related to length of the flower parts, and PC2 as generally related to width of the flower parts, we can say that **virginica** has similar width with **setosa,** but it has much higher length than **setosa**. **Versicolor** has much smaller width compared two other two types of irises, however it has relatively higher length compared to **setosa**.

This can also be (partially) verified also by looking at the mean pedal/sepal length and width values of each iris type from the original data.

Species	sepal length	sepal width	petal length	petal width
1	5.006	3.428	1.462	0.246
2	5.936	2.770	4.260	1.326
3	6.588	2.974	5.552	2.026

Here species 1 is setosa, species 2 is versicolor and species 3 is viriginica.

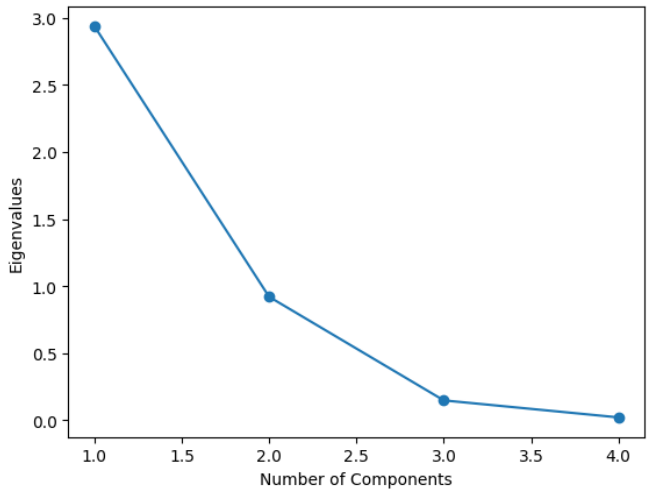
Appendix

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
from sklearn.decomposition import PCA
from sklearn.preprocessing import StandardScaler

data = pd.read_excel("data/IRIS.xlsx")

data.head()

data.columns = ["Species","sepal length","sepal width","petal length","petal
width"]

data.head()
```

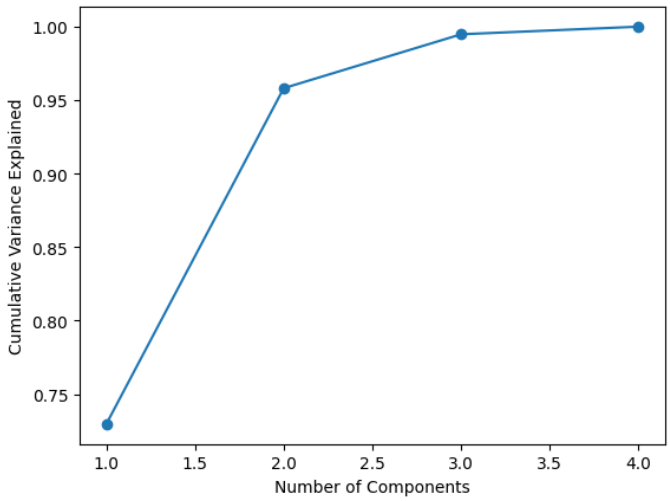
data.groupby(by = ['Species']).mean()

Taks 1

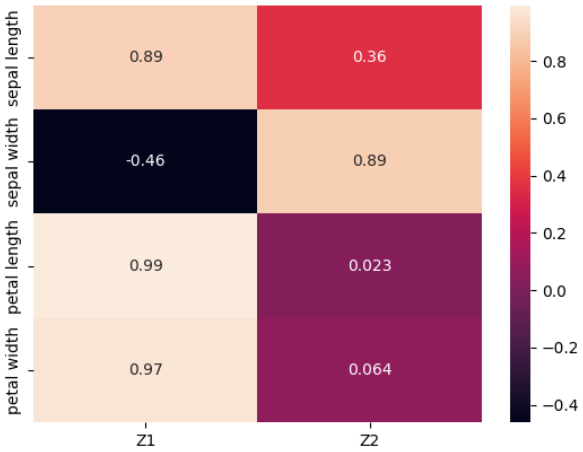
```
X = data.iloc[:,1:]
y = data.iloc[:,0]

X.head()

standard_scaler = StandardScaler()
X_normalized = pd.DataFrame(standard_scaler.fit_transform(X), columns =
X.columns)
X_normalized.head()



pca = PCA()
pca.fit(X_normalized)
eigenvalues = pca.explained_variance_
cumulative_variance_explained = np.cumsum(pca.explained_variance_ratio_)

plt.plot(range(1, len(eigenvalues)+1), eigenvalues, '-o')
plt.xlabel('Number of Components')
plt.ylabel('Eigenvalues')
plt.show()
```



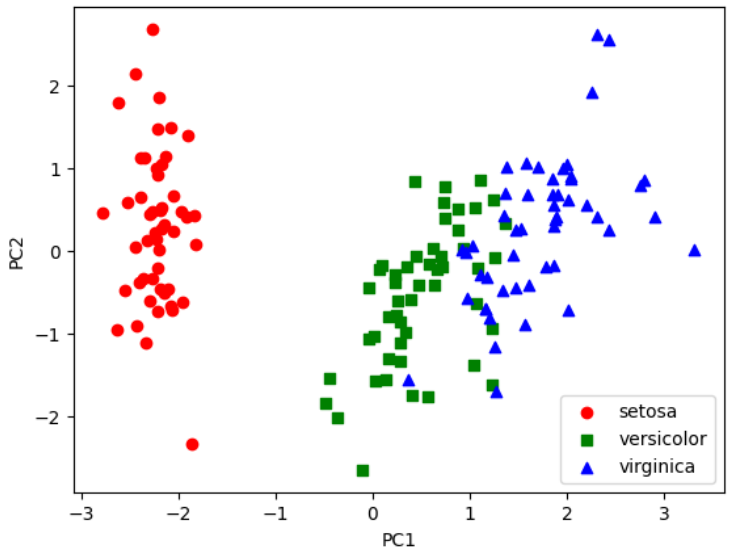

```
# Plot the cumulative variance explained as a function of the number of components
plt.plot(range(1, len(cumulative_variance_explained)+1), cumulative_variance_explained, '-o')
plt.xlabel('Number of Components')
plt.ylabel('Cumulative Variance Explained')
plt.show()
```




```
pca = PCA(n_components = 2)
pca.fit(X_normalized)

PCA(n_components=2)

loadings = pca.components_.T * np.sqrt(pca.explained_variance_)
loading_matrix = pd.DataFrame(loadings, columns=['Z1', 'Z2'], index = data.columns[1:])

import seaborn as sn

sn.heatmap(loading_matrix, annot = True)

plt.savefig("correlation_graph.png")

plt.show()
```




```
# Separate the dataset into three subsets, one for each type of iris
setosa_pca =  Z[data['Species'] == 1]
versicolor_pca = Z[data['Species'] == 2]
virginica_pca = Z[data['Species'] == 3]

plt.scatter(setosa_pca.iloc[:,0], setosa_pca.iloc[:,1], c='red', marker='o',
label='setosa')
plt.scatter(versicolor_pca.iloc[:,0], versicolor_pca.iloc[:,1], c='green',
marker='s', label='versicolor')
plt.scatter(virginica_pca.iloc[:,0], virginica_pca.iloc[:,1], c='blue',
marker='^', label='virginica')
plt.xlabel('PC1')
plt.ylabel('PC2')
plt.legend()
plt.show()
```




```
# Calculate the average principal component scores for each type of iris
setosa_pca_avg = setosa_pca.mean(axis=0)
```

```
versicolor_pca_avg = versicolor_pca.mean(axis=0)
virginica_pca_avg = virginica_pca.mean(axis=0)

# Plot the average principal component scores for each type of iris
plt.scatter(setosa_pca_avg[0], setosa_pca_avg[1], c='red', marker='o',
label='setosa')
plt.scatter(versicolor_pca_avg[0], versicolor_pca_avg[1], c='green',
marker='s', label='versicolor')
plt.scatter(virginica_pca_avg[0], virginica_pca_avg[1], c='blue', marker='^',
label='virginica')
plt.xlabel('PC1')
plt.ylabel('PC2')
plt.legend()
plt.show()
```