PROBLEM 3

Importing the required packages

```
import numpy as np
import pandas as pd
```

(a) Generating Data with 200 samples and 10 features and Performing PCA

```
np.random.seed(0)
num_samples = 200
num_features = 10
X = np.random.normal(size = (num_samples, num_features))

# Calculate the mean of the data
mean_vec = np.mean(X, axis=0)

# Center the data by subtracting the mean
X_centered = X - mean_vec

# Perform SVD on the centered data
U, S, Vt = np.linalg.svd(X_centered)

# Calculate the principal components
principal_components = Vt.T[:, :2]
```

(b) Calculate the Sum of Distances

```
# Project the centered data onto the plane formed by the first 2
principal components
projected_data = X_centered.dot(principal_components)

# Calculate distances between samples and the plane
distances = np.linalg.norm(X_centered -
projected_data.dot(principal_components.T), axis=1)

print("Sum of the distances between samples and plane:",
sum(distances))
```

```
Sum of the distances between samples and plane: 516.8355639975055
```

(c) Generate Random Planes and Calculate Distances

```
# Function to calculate sum of distances between samples and a given
plane
def sum_distances_to_plane(plane_normal, plane_point, samples):
    distances = np.abs(np.dot(samples - plane_point, plane_normal))
    return np.sum(distances)
```

```
# Generate 50 random planes and calculate their sum distances
num_random_planes = 50
sum_distances_random_planes = []

for _ in range(num_random_planes):
    # Generate a random plane by selecting two random data points
    random_indices = np.random.choice(num_samples, 2, replace=False)
    plane_point = X[random_indices[0]]
    plane_normal = X[random_indices[1]] - plane_point
    plane_normal /= np.linalg.norm(plane_normal)

    # Calculate sum of distances
    sum_distance = sum_distances_to_plane(plane_normal, plane_point,
X_centered)
    sum_distances_random_planes.append(sum_distance)

# Calculate sum of distances for the principal components plane
sum_distance_principal_components =
sum_distances_to_plane(principal_components[:, 0], mean_vec,
X_centered)

print("Sum of distances for random planes:",
sum_distances_random_planes)
print("Sum of distances for principal components plane:",
sum_distance_principal_components)
print("Is the sum of distances least for the principal components
plane?",
      sum_distance_principal_components <=
min(sum_distances_random_planes))
```

```
Sum of distances for random planes: [532.4151425093023,
504.54980425868644, 383.24388690332296, 569.3151214799045,
347.8249958494118, 463.83052069042196, 495.74597617112954,
406.91416471828677, 369.4232201230117, 313.645521812919,
708.3188777736088, 502.3677065704577, 682.212380523835,
513.0834713308132, 306.97550172055793, 568.5341764779935,
545.0026515857459, 563.6499801014721, 445.6141961483419,
192.7526016473585, 723.4837287706162, 563.5268825470104,
251.32073243060225, 254.375835095368, 287.8296837327546,
578.7507338852724, 261.4196006115002, 451.3739255806056,
317.320929203768, 310.58919427183474, 580.0401418572764,
842.5976986582161, 318.06469374394146, 640.9027759548535,
439.77815638866116, 765.5999913716325, 414.1323930231304,
454.35926577271465, 309.8732762784041, 689.4453784896366,
372.8319253106386, 567.2659974546001, 261.2452741489482,
199.94319239805793, 432.2276505588093, 426.5191597484502,
428.81335915460187, 559.1882378712977, 327.99835859539667,
372.0734546668329]
Sum of distances for principal components plane: 179.66979828741438
Is the sum of distances least for the principal components plane? True
```

Hence it is verified that the distance is minimum for the plane obtained by the PCA