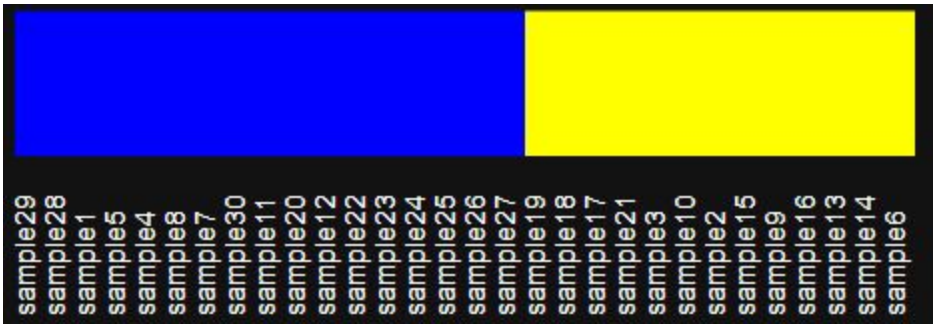
Assignment 4

İlker SIĞIRCI
2171940

Steps for Goal 1:

In this step, I have used the webpage that is given in the assignment text which is “http://mev.tm4.org”. In that webpage, I uploaded hw4dataset.tsv file. After that, I chose to cluster the groups in the dataset with k-means clustering.



Diseased (m = 17) = [1, 4, 5, 7, 8, 11, 12, 20, 22, 23, 24, 25, 26, 27, 28, 29, 30]

Healthy (n = 13) = [2, 3, 6, 9, 10, 13, 14, 15, 16, 17, 18, 19, 21]

Steps for Goal 2:

In this step, I have written a simple python script. In this script, first I have divided the dataset as “healthy” and “diseased” as I found in Goal 1. Then, I sum the row values in these two different dataset. Finally, I have substracted those rows on two different dataset. Specifically, I did healthy_df - diseased_df.

That means, the result of most negative ones belongs to diseased dataset while the most positive one belongs to the healthy one.

207430_s_at	-114173.7
210297_s_at	-79187.4
205623_at	-70827.7
204151_x_at	-67523.1
214303_x_at	-66676.2
214385_s_at	-60824.3
209699_x_at	-56351.2
201884_at	-48080.5
201891_s_at	-47414.8
204351_at	-47322.8

Diseased

215963_x_at	31565.6
213477_x_at	34590.0
212790_x_at	41285.3
201257_x_at	42029.2
206559_x_at	50763.5
210646_x_at	73386.3
203021_at	96562.3
204892_x_at	166822.9
220542_s_at	237036.4
205725_at	361614.0

Healthy

Code:

```
import numpy as np
import pandas as pd

df = pd.read_csv("hw4dataset.tsv", header=None, sep='\t')
df = df.rename(columns=df.iloc[0]).drop(df.index[0])
diseased = [1, 4, 5, 7, 8, 11, 12, 20, 22, 23, 24, 25, 26, 27, 28, 29, 30]
healthy  = [2, 3, 6, 9, 10, 13, 14, 15, 16, 17, 18, 19, 21]
diseased = ["sample"+str(x) for x in diseased]
healthy  = ["sample"+str(x) for x in healthy]
diseased_df = df[diseased]
healthy_df = df[healthy]
diseased_sum = diseased_df.astype(float).sum(axis = 1)
healthy_sum  = healthy_df.astype(float).sum(axis = 1)
df_add = healthy_sum.sub(diseased_sum)
df_add  = pd.concat([df["ID"],df_add], axis=1)
df_add.sort_values(by=[0], inplace=True)

print(df_add.iloc[:10])     # Top 10 Diseased
print(df_add.iloc[-10:])    # Top 10 Healthy
```