Assignment 1

1.

x1 and x2 correlation = 0.5297480198963057
x1 and x3 correlation = 0.3144432542848068
x2 and x3 correlation = -0.12945492014724624

Code(python):
./corr.py

2.
Sample 1 k1 = 1000000
Sample 1 k2 = 1000000
Sample 2 k1 = 800000
Sample 2 k2 = 800000
Code(python)
./KNN.py

3.
a) 100587, 101964, 120967, 100587, 101964, 120967
b) 120967
c) 0.67
d) When using k-means for finding out lings in a datasets it is important to choose an appropriate number of clusters. Too few and we will not catch the corrupted values. As we are picking the values furthest from the centroids as anomaly we also need to know how big proportionally the corrupted values are if we choose a too small number we will miss corrupted values if it is too big we will pick falsely corrupted values.
e) No 5 gives same result as 6 but 4 detect falsely corrupted values.
code(python)
./kmeans.py