Q2 Readme：

Introduction:

In Q2, we can divide our work into two parts:

- Two main features extracting from the origin dataset
- Write the grid-based outlier detection by myself.

The main methods:

1. Two main feature extractors
   Using PCA to reduce dimension, and Explained variance percentage = 0.98, which means these two dimensions are excellent representative of the original data.
2. Distance-Based outlier detection: a grid-based method
   a) Divide the dataset into cells with length (K is the dimensionality, D is the distance)
   b) Define Layer-1 neighbors – all the intermediate neighbor cells. The maximum distance between a cell and its neighbor cells is D
   c) Define Layer-2 neighbors – the cells within 3 cell of a certain cell. The minimum distance between a cell and the cells outside of Layer-2 neighbors is D
   d) Criteria:
      - Search a cell internally. If there are M objects inside, all the objects in this cell are not outlier
      - Search its layer-1 neighbors. For a cell which is dense (more than M objects in one cell), all the layer-1 neighbors belonging to this cell are not outliers. Besides, if there are M objects inside a cell and its layer-1 neighbors, all the objects in this cell are not outlier
      - Search its layer-2 neighbors. If there are less than M objects inside a cell, its layer-1 neighbor cells, and its layer-2 neighbor cells, all the objects in this cell are outlier.
      - Otherwise, the objects in this cell could be outlier, and then need to calculate the distance between the objects in this cell and the objects in the cells in the layer-2 neighbor cells to see whether the total points within D distance is more than M or not.

Result:

summary: the # of outlier detected in grid-search increase with M and decrease with D.

(1) when M=100:
    D=0.1:7134
    D=0.5:3730
    D=2:66
    D=5:0

(2) when D=2:
    M=100:66
    M=200:362
    M=1500:2712

and the corresponding result listed below.