In [365]:
```
"""
HOMEWORK 5 SOLUTION CREATED BY
OLABODE ALAMU
1498663
GUIDE TO ENGINEERING DATA SCIENCE

"""
```

Out[365]:
```
'\nHOMEWORK 5 SOLUTION CREATED BY\nOLABODE ALAMU\n1498663\nGUIDE TO ENGINEERI
NG DATA SCIENCE\n\n'
```

In [212]:
```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
```

In [213]:
```
# Import the dataset
#head_column = ['Date','Q-E','ZN-E','PH-E','DBO-E','DQO-E','SS-E','SSV-E','SED
-E','COND-E','PH-P','DBO-P','SS-P','SSV-P','SED-P','COND-P','PH-D','DBO-D','DQ
O-D','SS-D','SSV-D','SED-D','COND-D','PH-S','DBO-S','DQO-S','SS-S','SSV-S','SE
D-S','COND-S','RD-DBO-P','RD-SS-P','RD-SED-P','RD-DBO-S','RD-DQO-S','RD-DBO-
G','RD-DQO-G','RD-SS-G','RD-SED-G']

Water = pd.read_csv('water-treatment.data', header = None, na_values = '?', in
dex_col = 0)
```

In [324]:
```
# Prints out the top 5 rows of the dataset
print(Water.head())
```

```
               1    2    3      4      5      6     7    8     9    10  ...
  \
0                                                                      ...

D-1/3/90  44101.0  1.5  7.8    NaN  407.0  166.0  66.3  4.5  2110  7.9  ...

D-2/3/90  39024.0  3.0  7.7    NaN  443.0  214.0  69.2  6.5  2660  7.7  ...

D-4/3/90  32229.0  5.0  7.6    NaN  528.0  186.0  69.9  3.4  1666  7.7  ...

D-5/3/90  35023.0  3.5  7.9  205.0  588.0  192.0  65.6  4.5  2430  7.8  ...

D-6/3/90  36924.0  1.5  8.0  242.0  496.0  176.0  64.8  4.0  2110  7.9  ...


              29    30    31    32    33    34    35    36    37     38
0
D-1/3/90  2000.0   NaN  58.8  95.5   NaN  70.0   NaN  79.4  87.3   99.6
D-2/3/90  2590.0   NaN  60.7  94.8   NaN  80.8   NaN  79.5  92.1  100.0
D-4/3/90  1888.0   NaN  58.2  95.6   NaN  52.9   NaN  75.8  88.7   98.5
D-5/3/90  1840.0  33.1  64.2  95.3  87.3  72.3  90.2  82.3  89.6  100.0
D-6/3/90  2120.0   NaN  62.7  95.6   NaN  71.0  92.1  78.2  87.5   99.5

[5 rows x 38 columns]
```

In [325]:
```
print(Water.shape) # Shows the shape of the dataframe
```

```
(527, 38)
```

In [215]:
```
X = Water.columns # This shows all the columns present in the dataset
```

In [216]:
```
print(X)
```

```
Int64Index([ 1,  2,  3,  4,  5,  6,  7,  8,  9, 10, 11, 12, 13, 14, 15, 16, 1
7,
            18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 3
4,
            35, 36, 37, 38],
           dtype='int64')
```

In [326]:
```
# This for loop was created to iterate through the difference columns of the d
ataset, split out each column
# and calculate the different statistical parameters for each column and print
 them out
for column in Water.columns:
    Maximum = Water[i].max()
    Minimum = Water[i].min()
    Mean_value = Water[i].mean()
    Median_value = Water[i].median()
    Standard_dev = Water[i].std()
    print('Column ',i,'\nMaximum value is ', Maximum, '\nMinimum value is ', M
inimum,
          '\nMean value is ', Mean_value, '\nMedian is ', Median_value,
          '\nStandard deviation is ', Standard_dev)
```

```
Column  37
Maximum value is  99.4
Minimum value is  10.3
Mean value is  88.9633911368016
Median is  90.7
Standard deviation is  8.221612495740384
Column  37
Maximum value is  99.4
Minimum value is  10.3
Mean value is  88.9633911368016
Median is  90.7
Standard deviation is  8.221612495740384
Column  37
Maximum value is  99.4
Minimum value is  10.3
Mean value is  88.9633911368016
Median is  90.7
Standard deviation is  8.221612495740384
Column  37
Maximum value is  99.4
Minimum value is  10.3
Mean value is  88.9633911368016
Median is  90.7
Standard deviation is  8.221612495740384
Column  37
Maximum value is  99.4
Minimum value is  10.3
Mean value is  88.9633911368016
Median is  90.7
Standard deviation is  8.221612495740384
Column  37
Maximum value is  99.4
Minimum value is  10.3
Mean value is  88.9633911368016
Median is  90.7
Standard deviation is  8.221612495740384
Column  37
Maximum value is  99.4
Minimum value is  10.3
Mean value is  88.9633911368016
Median is  90.7
Standard deviation is  8.221612495740384
Column  37
Maximum value is  99.4
Minimum value is  10.3
Mean value is  88.9633911368016
Median is  90.7
Standard deviation is  8.221612495740384
Column  37
Maximum value is  99.4
Minimum value is  10.3
Mean value is  88.9633911368016
Median is  90.7
Standard deviation is  8.221612495740384
Column  37
Maximum value is  99.4
Minimum value is  10.3
```

```
Mean value is  88.9633911368016
Median is  90.7
Standard deviation is  8.221612495740384
Column  37
Maximum value is  99.4
Minimum value is  10.3
Mean value is  88.9633911368016
Median is  90.7
Standard deviation is  8.221612495740384
Column  37
Maximum value is  99.4
Minimum value is  10.3
Mean value is  88.9633911368016
Median is  90.7
Standard deviation is  8.221612495740384
Column  37
Maximum value is  99.4
Minimum value is  10.3
Mean value is  88.9633911368016
Median is  90.7
Standard deviation is  8.221612495740384
Column  37
Maximum value is  99.4
Minimum value is  10.3
Mean value is  88.9633911368016
Median is  90.7
Standard deviation is  8.221612495740384
Column  37
Maximum value is  99.4
Minimum value is  10.3
Mean value is  88.9633911368016
Median is  90.7
Standard deviation is  8.221612495740384
Column  37
Maximum value is  99.4
Minimum value is  10.3
Mean value is  88.9633911368016
Median is  90.7
Standard deviation is  8.221612495740384
Column  37
Maximum value is  99.4
Minimum value is  10.3
Mean value is  88.9633911368016
Median is  90.7
Standard deviation is  8.221612495740384
Column  37
Maximum value is  99.4
Minimum value is  10.3
Mean value is  88.9633911368016
Median is  90.7
Standard deviation is  8.221612495740384
Column  37
Maximum value is  99.4
Minimum value is  10.3
Mean value is  88.9633911368016
Median is  90.7
Standard deviation is  8.221612495740384
```

```
Column  37
Maximum value is  99.4
Minimum value is  10.3
Mean value is  88.9633911368016
Median is  90.7
Standard deviation is  8.221612495740384
Column  37
Maximum value is  99.4
Minimum value is  10.3
Mean value is  88.9633911368016
Median is  90.7
Standard deviation is  8.221612495740384
Column  37
Maximum value is  99.4
Minimum value is  10.3
Mean value is  88.9633911368016
Median is  90.7
Standard deviation is  8.221612495740384
Column  37
Maximum value is  99.4
Minimum value is  10.3
Mean value is  88.9633911368016
Median is  90.7
Standard deviation is  8.221612495740384
Column  37
Maximum value is  99.4
Minimum value is  10.3
Mean value is  88.9633911368016
Median is  90.7
Standard deviation is  8.221612495740384
Column  37
Maximum value is  99.4
Minimum value is  10.3
Mean value is  88.9633911368016
Median is  90.7
Standard deviation is  8.221612495740384
Column  37
Maximum value is  99.4
Minimum value is  10.3
Mean value is  88.9633911368016
Median is  90.7
Standard deviation is  8.221612495740384
Column  37
Maximum value is  99.4
Minimum value is  10.3
Mean value is  88.9633911368016
Median is  90.7
Standard deviation is  8.221612495740384
Column  37
Maximum value is  99.4
Minimum value is  10.3
Mean value is  88.9633911368016
Median is  90.7
Standard deviation is  8.221612495740384
Column  37
Maximum value is  99.4
Minimum value is  10.3
```

```
Mean value is  88.9633911368016
Median is  90.7
Standard deviation is  8.221612495740384
Column  37
Maximum value is  99.4
Minimum value is  10.3
Mean value is  88.9633911368016
Median is  90.7
Standard deviation is  8.221612495740384
Column  37
Maximum value is  99.4
Minimum value is  10.3
Mean value is  88.9633911368016
Median is  90.7
Standard deviation is  8.221612495740384
Column  37
Maximum value is  99.4
Minimum value is  10.3
Mean value is  88.9633911368016
Median is  90.7
Standard deviation is  8.221612495740384
Column  37
Maximum value is  99.4
Minimum value is  10.3
Mean value is  88.9633911368016
Median is  90.7
Standard deviation is  8.221612495740384
Column  37
Maximum value is  99.4
Minimum value is  10.3
Mean value is  88.9633911368016
Median is  90.7
Standard deviation is  8.221612495740384
Column  37
Maximum value is  99.4
Minimum value is  10.3
Mean value is  88.9633911368016
Median is  90.7
Standard deviation is  8.221612495740384
Column  37
Maximum value is  99.4
Minimum value is  10.3
Mean value is  88.9633911368016
Median is  90.7
Standard deviation is  8.221612495740384
Column  37
Maximum value is  99.4
Minimum value is  10.3
Mean value is  88.9633911368016
Median is  90.7
Standard deviation is  8.221612495740384
Column  37
Maximum value is  99.4
Minimum value is  10.3
Mean value is  88.9633911368016
Median is  90.7
Standard deviation is  8.221612495740384
```

In [327]:
```
# Computes the number of missing values in each column and prints out a statem
ent
Missing_values = Water.isnull().sum()
for column in Water.columns: # iterates through each column
    Missing_value = Missing_values[column]
    print('Column ',column, ' has ',Missing_value, ' number of missing value
s.')
```

```
Column  1  has  18  number of missing values.
Column  2  has  3  number of missing values.
Column  3  has  0  number of missing values.
Column  4  has  23  number of missing values.
Column  5  has  6  number of missing values.
Column  6  has  1  number of missing values.
Column  7  has  11  number of missing values.
Column  8  has  25  number of missing values.
Column  9  has  0  number of missing values.
Column  10  has  0  number of missing values.
Column  11  has  40  number of missing values.
Column  12  has  0  number of missing values.
Column  13  has  11  number of missing values.
Column  14  has  24  number of missing values.
Column  15  has  0  number of missing values.
Column  16  has  0  number of missing values.
Column  17  has  28  number of missing values.
Column  18  has  9  number of missing values.
Column  19  has  2  number of missing values.
Column  20  has  13  number of missing values.
Column  21  has  25  number of missing values.
Column  22  has  0  number of missing values.
Column  23  has  1  number of missing values.
Column  24  has  23  number of missing values.
Column  25  has  18  number of missing values.
Column  26  has  5  number of missing values.
Column  27  has  17  number of missing values.
Column  28  has  28  number of missing values.
Column  29  has  1  number of missing values.
Column  30  has  62  number of missing values.
Column  31  has  4  number of missing values.
Column  32  has  27  number of missing values.
Column  33  has  40  number of missing values.
Column  34  has  26  number of missing values.
Column  35  has  36  number of missing values.
Column  36  has  25  number of missing values.
Column  37  has  8  number of missing values.
Column  38  has  31  number of missing values.
```

In [328]:
```
# Import missing value using the median
# Step 1: Computes the medians across all the columns and passes them into a l
ist
Value_list = [] # Creates an empty list

for column in Water.columns:
    Median_value = Water[column].median()
    Value_list.append(Median_value)
print(Value_list)

# Step 2 Transform the list of median values into a dictionary where the colum
n number is the key and the median
# values are the value in the key: value pair of the dictionary
Median_dict = {}
for column in Water.columns:
    Median_dict[column]= Value_list[column-1]

print(Median_dict)

# Step 3 Fill the missing value with the medians in the dictionary

Water_cleaned = Water.fillna(value = Median_dict)

print(Water_cleaned)
```

```
[35990.0, 1.5, 7.8, 182.5, 400.0, 196.0, 64.3, 4.5, 1406.0, 7.8, 197.0, 220.
0, 62.8, 4.5, 1420.0, 7.8, 119.0, 274.0, 90.0, 74.5, 0.3, 1428.0, 7.7, 18.0,
 84.0, 19.0, 81.15, 0.01, 1432.0, 39.6, 59.5, 93.3, 85.4, 69.9, 90.2, 79.15,
 90.7, 99.7]
{1: 35990.0, 2: 1.5, 3: 7.8, 4: 182.5, 5: 400.0, 6: 196.0, 7: 64.3, 8: 4.5,
 9: 1406.0, 10: 7.8, 11: 197.0, 12: 220.0, 13: 62.8, 14: 4.5, 15: 1420.0, 16:
7.8, 17: 119.0, 18: 274.0, 19: 90.0, 20: 74.5, 21: 0.3, 22: 1428.0, 23: 7.7,
 24: 18.0, 25: 84.0, 26: 19.0, 27: 81.15, 28: 0.01, 29: 1432.0, 30: 39.6, 31:
59.5, 32: 93.3, 33: 85.4, 34: 69.9, 35: 90.2, 36: 79.15, 37: 90.7, 38: 99.7}
                1     2    3      4      5      6     7    8     9    10  \
0
D-1/3/90    44101.0  1.50  7.8  182.5  407.0  166.0  66.3  4.5  2110  7.9
D-2/3/90    39024.0  3.00  7.7  182.5  443.0  214.0  69.2  6.5  2660  7.7
D-4/3/90    32229.0  5.00  7.6  182.5  528.0  186.0  69.9  3.4  1666  7.7
D-5/3/90    35023.0  3.50  7.9  205.0  588.0  192.0  65.6  4.5  2430  7.8
D-6/3/90    36924.0  1.50  8.0  242.0  496.0  176.0  64.8  4.0  2110  7.9
D-7/3/90    38572.0  3.00  7.8  202.0  372.0  186.0  68.8  4.5  1644  7.8
D-8/3/90    41115.0  6.00  7.8  182.5  552.0  262.0  64.1  5.0  1603  7.8
D-9/3/90    36107.0  5.00  7.7  215.0  489.0  334.0  40.7  6.0  1613  7.6
D-11/3/90   29156.0  2.50  7.7  206.0  451.0  194.0  69.1  4.5  1249  7.7
D-12/3/90   39246.0  2.00  7.8  172.0  506.0  200.0  69.0  5.0  1865  7.8
D-13/3/90   42393.0  0.70  7.9  189.0  478.0  230.0  67.0  5.5  1410  8.1
D-14/3/90   42857.0  1.50  7.7  238.0  319.0  292.0  33.8  3.5  1261  7.6
D-15/3/90   42911.0  0.70  7.6  114.0  252.0  116.0  58.6  1.2  1238  7.9
D-16/3/90   40376.0  1.50  8.1  204.0  333.0  174.0  67.8  3.0  2390  7.8
D-18/3/90   40923.0  3.50  7.6  146.0  329.0  188.0  57.4  2.5  1300  7.6
D-19/3/90   43830.0  1.50  7.8  177.0  512.0  214.0  58.9  5.5  1605  7.7
D-20/3/90   39165.0  1.20  7.4  250.0  447.0  252.0  61.1  7.0  1533  7.4
D-21/3/90   35791.0  1.20  7.8  277.0  466.0  246.0  63.4  4.0  1556  7.7
D-22/3/90   37419.0  1.20  7.6  219.0  446.0  222.0  61.3  5.5  1600  7.7
D-23/3/90   40983.0  3.00  7.6  182.0  431.0  214.0  57.0  7.0  1591  7.5
D-25/3/90   42217.0  8.50  7.5  138.0  333.0  240.0  55.0  3.8  1087  7.5
D-26/3/90   47665.0  1.20  7.7  156.0  405.0  200.0  74.0  4.0  1856  7.6
D-27/3/90   44314.0  3.00  7.8  155.0  389.0  308.0  49.4  6.0  1927  7.7
D-28/3/90   40841.0  1.00  7.6  179.0  389.0  168.0  69.0  3.5  1240  7.8
D-29/3/90   41157.0  3.00  8.0  145.0  398.0  192.0  66.7  4.5  2240  8.0
D-30/3/90   40078.0  1.40  7.9  198.0  464.0  228.0  64.9  4.6  1431  7.6
D-1/2/90    44365.0  7.50  7.9  182.5  365.0  212.0  62.3  3.5  1339  7.9
D-2/2/90    43080.0  4.25  7.8   95.0  349.0  136.0  76.5  2.5  1063  7.8
D-4/2/90    29414.0  3.00  7.6  160.0  374.0  168.0  69.0  3.1  1042  7.6
D-5/2/90    37312.0  1.00  8.1  205.0  492.0  192.0  70.8  4.0  1454  8.1
...             ...   ...  ...    ...    ...    ...   ...  ...   ...  ...
D-25/10/91  35400.0  0.70  7.6  156.0  364.0  194.0  63.9  5.5  1680  7.6
D-26/10/91  30964.0  3.30  7.7  220.0  540.0  184.0  62.0  3.5  1445  7.7
D-27/10/91  35573.0  7.30  7.6  176.0  333.0  178.0  64.0  3.5  1627  7.7
D-29/10/91  29801.0  1.60  7.7  172.0  400.0  136.0  70.1  1.5  1402  7.7
D-30/10/91  31524.0  1.60  7.9  182.5  478.0  204.0  64.7  6.0  1798  7.9
D-1/8/91    29834.0  3.00  7.4  160.0  348.0  194.0  61.9  3.0  1720  7.5
D-2/8/91    28492.0  2.60  7.5  124.0  281.0  172.0  66.3  3.0  1520  7.5
D-4/8/91    24978.0  0.50  7.3  146.0  288.0  124.0  67.7  2.0  1210  7.4
D-5/8/91    29719.0  0.20  7.6  133.0  284.0  186.0  71.0  5.0  1114  7.6
D-6/8/91    29741.0  0.45  7.9  151.0  316.0  196.0  64.3  2.5   948  7.8
D-7/8/91    29027.0  0.40  7.6  136.0  328.0  186.0  67.7  3.0   899  7.6
D-8/8/91    30211.0  0.50  7.6  114.0  521.0  506.0  44.3  7.5   866  7.5
D-9/8/91    30848.0  0.20  7.7  142.0  376.0  144.0  70.8  3.0   940  7.6
D-11/8/91   17527.0  0.55  7.5  150.0  171.0  172.0  37.2  1.4   732  7.5
D-12/8/91   33331.0  0.23  7.6   92.0  233.0  234.0  37.6  1.4   829  7.6
```

D-13/8/91	27998.0	0.62	7.5	138.0	268.0	154.0	66.2	1.7	890	7.5
D-14/8/91	32845.0	0.23	7.6	84.0	251.0	98.0	71.4	2.0	866	7.6
D-16/8/91	27933.0	0.20	7.6	158.0	375.0	178.0	60.7	3.5	1049	7.7
D-18/8/91	27527.0	0.20	7.3	191.0	240.0	166.0	73.5	3.0	1072	7.4
D-19/8/91	32363.0	0.10	7.6	159.0	310.0	146.0	68.5	1.6	1096	7.6
D-20/8/91	31437.0	0.47	7.6	132.0	304.0	148.0	64.9	2.0	939	7.7
D-21/8/91	31914.0	2.00	7.7	127.0	274.0	144.0	72.2	2.0	1031	7.6
D-22/8/91	28088.0	0.20	7.5	153.0	307.0	124.0	82.3	2.5	1044	7.6
D-23/8/91	27838.0	0.13	7.6	179.0	265.0	128.0	71.9	1.8	992	7.6
D-25/8/91	29271.0	0.36	7.5	99.0	585.0	140.0	71.4	4.5	962	7.6
D-26/8/91	32723.0	0.16	7.7	93.0	252.0	176.0	56.8	2.3	894	7.7
D-27/8/91	33535.0	0.32	7.8	192.0	346.0	172.0	68.6	4.0	988	7.8
D-28/8/91	32922.0	0.30	7.4	139.0	367.0	180.0	64.4	3.0	1060	7.5
D-29/8/91	32190.0	0.30	7.3	200.0	545.0	258.0	65.1	4.0	1260	7.4
D-30/8/91	30488.0	0.21	7.5	152.0	300.0	132.0	69.7	4.5	1073	7.4

	...	29	30	31	32	33	34	35	36	37	\
0	...										
D-1/3/90	...	2000.0	39.6	58.8	95.5	85.4	70.0	90.2	79.40	87.3	
D-2/3/90	...	2590.0	39.6	60.7	94.8	85.4	80.8	90.2	79.50	92.1	
D-4/3/90	...	1888.0	39.6	58.2	95.6	85.4	52.9	90.2	75.80	88.7	
D-5/3/90	...	1840.0	33.1	64.2	95.3	87.3	72.3	90.2	82.30	89.6	
D-6/3/90	...	2120.0	39.6	62.7	95.6	85.4	71.0	92.1	78.20	87.5	
D-7/3/90	...	1764.0	39.6	59.7	96.5	86.7	78.3	90.1	73.10	84.9	
D-8/3/90	...	1703.0	39.6	61.9	93.8	89.1	79.8	90.2	86.20	90.1	
D-9/3/90	...	1606.0	39.6	70.4	95.6	90.6	53.7	92.1	66.90	94.6	
D-11/3/90	...	1338.0	46.1	43.6	92.5	85.6	58.2	92.2	73.80	90.2	
D-12/3/90	...	1616.0	21.2	59.7	90.8	88.4	66.1	89.0	69.00	86.5	
D-13/3/90	...	1575.0	0.6	45.8	92.0	11.6	25.7	19.6	36.00	43.0	
D-14/3/90	...	1304.0	31.8	61.2	92.9	85.4	69.9	90.2	79.15	18.5	
D-15/3/90	...	1221.0	46.6	48.5	91.7	85.4	20.4	26.3	31.70	10.3	
D-16/3/90	...	2550.0	41.1	50.0	84.0	76.5	52.9	84.3	54.10	43.7	
D-18/3/90	...	1545.0	32.7	33.3	90.0	82.6	61.3	87.0	71.40	78.2	
D-19/3/90	...	1110.0	28.0	72.7	92.7	78.8	36.6	85.9	60.40	90.7	
D-20/3/90	...	1402.0	49.8	58.3	92.3	89.9	96.8	94.4	98.10	92.1	
D-21/3/90	...	1606.0	39.6	39.6	78.3	85.5	70.4	91.3	73.40	89.4	
D-22/3/90	...	1780.0	35.3	57.5	92.0	75.6	49.3	80.8	60.80	76.1	
D-23/3/90	...	1597.0	20.1	64.5	93.6	86.9	68.1	87.4	72.20	88.3	
D-25/3/90	...	1223.0	29.4	55.4	91.3	85.2	68.0	88.4	81.40	92.9	
D-26/3/90	...	1706.0	28.1	50.0	91.4	85.2	76.5	87.8	82.50	88.5	
D-27/3/90	...	1869.0	52.0	64.9	90.8	87.6	71.2	90.3	77.60	92.5	
D-28/3/90	...	1416.0	26.7	66.2	95.0	89.2	72.5	91.1	78.70	88.1	
D-29/3/90	...	2290.0	34.3	65.0	94.2	89.3	69.7	89.7	78.10	89.1	
D-30/3/90	...	1475.0	27.2	67.6	97.3	90.4	68.0	91.4	78.00	90.4	
D-1/2/90	...	1377.0	39.6	50.0	89.4	85.4	76.6	90.2	79.50	90.6	
D-2/2/90	...	1220.0	6.8	47.9	80.0	84.6	78.9	80.0	80.80	82.4	
D-4/2/90	...	1087.0	42.7	54.5	95.7	84.9	73.6	88.1	78.90	83.3	
D-5/2/90	...	1275.0	39.6	33.0	94.5	85.3	73.7	84.4	76.80	80.7	
...	...	...	...	...	...	...	...	...	...	...	
D-25/10/91	...	1840.0	47.3	61.3	94.0	76.4	69.9	86.5	82.40	90.7	
D-26/10/91	...	1337.0	39.6	38.6	93.3	85.4	87.0	92.7	95.00	91.8	
D-27/10/91	...	1799.0	39.6	40.4	95.0	85.4	72.9	90.9	79.90	90.7	
D-29/10/91	...	1468.0	32.4	40.4	88.0	87.8	77.6	91.3	85.30	83.8	
D-30/10/91	...	1568.0	39.6	43.9	65.3	85.4	75.3	90.2	81.20	89.7	
D-1/8/91	...	1772.0	29.1	53.5	92.0	85.7	64.2	90.6	72.70	85.6	
D-2/8/91	...	1549.0	34.2	47.7	99.0	84.4	66.1	90.3	73.30	88.4	
D-4/8/91	...	1259.0	39.6	52.3	92.3	75.3	69.9	86.3	79.15	76.6	

```
D-5/8/91    ...    1100.0  55.1  73.2  96.0  82.0  62.5  91.7  78.90  88.7
D-6/8/91    ...     951.0  44.8  64.5  95.0  82.2  80.0  89.4  86.10  88.8
D-7/8/91    ...     898.0  39.6  64.1  94.0  84.9  60.4  91.9  76.80  87.6
D-8/8/91    ...     884.0  48.7  86.9  98.8  81.0  67.9  90.4  91.60  96.0
D-9/8/91    ...     947.0  39.6  47.6  97.3  85.4  77.6  93.7  84.80  90.3
D-11/8/91   ...     728.0  62.8  77.7  90.0  73.8  65.5  92.7  77.20  90.7
D-12/8/91   ...     929.0  36.9  43.6  86.7  87.7  71.9  91.3  79.80  92.3
D-13/8/91   ...     858.0  38.1  41.6  86.7  87.7  79.0  94.2  87.70  91.6
D-14/8/91   ...     879.0  50.9  36.5  80.0  87.0  69.6  91.7  80.50  82.7
D-16/8/91   ...     828.0  68.0  66.7  96.7  81.6  41.8  94.3  72.50  83.1
D-18/8/91   ...     999.0  38.5  54.5  88.0  90.0  69.9  95.8  81.70  93.4
D-19/8/91   ...    1083.0  25.2  61.4  91.2  78.6  65.1  86.8  81.00  89.0
D-20/8/91   ...    1012.0  45.6  60.3  94.4  82.5  72.9  89.4  86.20  91.2
D-21/8/91   ...    1053.0  39.6  59.5  93.3  85.4  77.7  90.2  79.15  90.7
D-22/8/91   ...    1038.0  40.5  54.4  94.0  89.7  75.5  93.5  85.00  90.3
D-23/8/91   ...    1044.0  13.7  45.0  95.0  87.5  71.3  93.9  79.60  89.1
D-25/8/91   ...     968.0  40.8  71.6  93.3  59.0  26.4  74.7  83.80  81.4
D-26/8/91   ...     942.0  39.6  62.3  93.3  69.8  75.9  79.6  78.60  96.6
D-27/8/91   ...     950.0  39.6  58.3  97.8  83.0  59.1  91.1  74.60  90.7
D-28/8/91   ...    1136.0  39.6  65.0  97.1  76.2  66.4  82.0  77.10  88.9
D-29/8/91   ...    1326.0  39.8  65.9  97.1  81.7  70.9  89.5  87.00  89.5
D-30/8/91   ...    1224.0  39.6  69.5  93.3  81.7  76.4  90.2  81.70  86.4

              38
0
D-1/3/90     99.6
D-2/3/90    100.0
D-4/3/90     98.5
D-5/3/90    100.0
D-6/3/90     99.5
D-7/3/90    100.0
D-8/3/90     99.0
D-9/3/90    100.0
D-11/3/90    99.4
D-12/3/90    99.6
D-13/3/90    36.4
D-14/3/90    42.9
D-15/3/90    95.4
D-16/3/90   100.0
D-18/3/90    99.2
D-19/3/90   100.0
D-20/3/90   100.0
D-21/3/90    99.4
D-22/3/90    99.6
D-23/3/90   100.0
D-25/3/90   100.0
D-26/3/90    99.8
D-27/3/90   100.0
D-28/3/90   100.0
D-29/3/90   100.0
D-30/3/90   100.0
D-1/2/90     98.6
D-2/2/90    100.0
D-4/2/90    100.0
D-5/2/90    100.0
...           ...
D-25/10/91   99.8
```

```
D-26/10/91   95.7
D-27/10/91   98.6
D-29/10/91   96.7
D-30/10/91   99.2
D-1/8/91     99.3
D-2/8/91     99.3
D-4/8/91     99.0
D-5/8/91     99.6
D-6/8/91    100.0
D-7/8/91     99.3
D-8/8/91     99.7
D-9/8/91     99.3
D-11/8/91    98.6
D-12/8/91    99.3
D-13/8/91    98.8
D-14/8/91    99.0
D-16/8/91    98.6
D-18/8/91    99.3
D-19/8/91    99.4
D-20/8/91    99.5
D-21/8/91    99.7
D-22/8/91   100.0
D-23/8/91   100.0
D-25/8/91    99.7
D-26/8/91    99.6
D-27/8/91   100.0
D-28/8/91    99.0
D-29/8/91    99.8
D-30/8/91    99.7

[527 rows x 38 columns]
```

In [329]:
```
print(Water_cleaned)
```

	1	2	3	4	5	6	7	8	9	10	\
0											
D-1/3/90	44101.0	1.50	7.8	182.5	407.0	166.0	66.3	4.5	2110	7.9	
D-2/3/90	39024.0	3.00	7.7	182.5	443.0	214.0	69.2	6.5	2660	7.7	
D-4/3/90	32229.0	5.00	7.6	182.5	528.0	186.0	69.9	3.4	1666	7.7	
D-5/3/90	35023.0	3.50	7.9	205.0	588.0	192.0	65.6	4.5	2430	7.8	
D-6/3/90	36924.0	1.50	8.0	242.0	496.0	176.0	64.8	4.0	2110	7.9	
D-7/3/90	38572.0	3.00	7.8	202.0	372.0	186.0	68.8	4.5	1644	7.8	
D-8/3/90	41115.0	6.00	7.8	182.5	552.0	262.0	64.1	5.0	1603	7.8	
D-9/3/90	36107.0	5.00	7.7	215.0	489.0	334.0	40.7	6.0	1613	7.6	
D-11/3/90	29156.0	2.50	7.7	206.0	451.0	194.0	69.1	4.5	1249	7.7	
D-12/3/90	39246.0	2.00	7.8	172.0	506.0	200.0	69.0	5.0	1865	7.8	
D-13/3/90	42393.0	0.70	7.9	189.0	478.0	230.0	67.0	5.5	1410	8.1	
D-14/3/90	42857.0	1.50	7.7	238.0	319.0	292.0	33.8	3.5	1261	7.6	
D-15/3/90	42911.0	0.70	7.6	114.0	252.0	116.0	58.6	1.2	1238	7.9	
D-16/3/90	40376.0	1.50	8.1	204.0	333.0	174.0	67.8	3.0	2390	7.8	
D-18/3/90	40923.0	3.50	7.6	146.0	329.0	188.0	57.4	2.5	1300	7.6	
D-19/3/90	43830.0	1.50	7.8	177.0	512.0	214.0	58.9	5.5	1605	7.7	
D-20/3/90	39165.0	1.20	7.4	250.0	447.0	252.0	61.1	7.0	1533	7.4	
D-21/3/90	35791.0	1.20	7.8	277.0	466.0	246.0	63.4	4.0	1556	7.7	
D-22/3/90	37419.0	1.20	7.6	219.0	446.0	222.0	61.3	5.5	1600	7.7	
D-23/3/90	40983.0	3.00	7.6	182.0	431.0	214.0	57.0	7.0	1591	7.5	
D-25/3/90	42217.0	8.50	7.5	138.0	333.0	240.0	55.0	3.8	1087	7.5	
D-26/3/90	47665.0	1.20	7.7	156.0	405.0	200.0	74.0	4.0	1856	7.6	
D-27/3/90	44314.0	3.00	7.8	155.0	389.0	308.0	49.4	6.0	1927	7.7	
D-28/3/90	40841.0	1.00	7.6	179.0	389.0	168.0	69.0	3.5	1240	7.8	
D-29/3/90	41157.0	3.00	8.0	145.0	398.0	192.0	66.7	4.5	2240	8.0	
D-30/3/90	40078.0	1.40	7.9	198.0	464.0	228.0	64.9	4.6	1431	7.6	
D-1/2/90	44365.0	7.50	7.9	182.5	365.0	212.0	62.3	3.5	1339	7.9	
D-2/2/90	43080.0	4.25	7.8	95.0	349.0	136.0	76.5	2.5	1063	7.8	
D-4/2/90	29414.0	3.00	7.6	160.0	374.0	168.0	69.0	3.1	1042	7.6	
D-5/2/90	37312.0	1.00	8.1	205.0	492.0	192.0	70.8	4.0	1454	8.1	
...	...	...	...	...	...	...	...	...	...	...	
D-25/10/91	35400.0	0.70	7.6	156.0	364.0	194.0	63.9	5.5	1680	7.6	
D-26/10/91	30964.0	3.30	7.7	220.0	540.0	184.0	62.0	3.5	1445	7.7	
D-27/10/91	35573.0	7.30	7.6	176.0	333.0	178.0	64.0	3.5	1627	7.7	
D-29/10/91	29801.0	1.60	7.7	172.0	400.0	136.0	70.1	1.5	1402	7.7	
D-30/10/91	31524.0	1.60	7.9	182.5	478.0	204.0	64.7	6.0	1798	7.9	
D-1/8/91	29834.0	3.00	7.4	160.0	348.0	194.0	61.9	3.0	1720	7.5	
D-2/8/91	28492.0	2.60	7.5	124.0	281.0	172.0	66.3	3.0	1520	7.5	
D-4/8/91	24978.0	0.50	7.3	146.0	288.0	124.0	67.7	2.0	1210	7.4	
D-5/8/91	29719.0	0.20	7.6	133.0	284.0	186.0	71.0	5.0	1114	7.6	
D-6/8/91	29741.0	0.45	7.9	151.0	316.0	196.0	64.3	2.5	948	7.8	
D-7/8/91	29027.0	0.40	7.6	136.0	328.0	186.0	67.7	3.0	899	7.6	
D-8/8/91	30211.0	0.50	7.6	114.0	521.0	506.0	44.3	7.5	866	7.5	
D-9/8/91	30848.0	0.20	7.7	142.0	376.0	144.0	70.8	3.0	940	7.6	
D-11/8/91	17527.0	0.55	7.5	150.0	171.0	172.0	37.2	1.4	732	7.5	
D-12/8/91	33331.0	0.23	7.6	92.0	233.0	234.0	37.6	1.4	829	7.6	
D-13/8/91	27998.0	0.62	7.5	138.0	268.0	154.0	66.2	1.7	890	7.5	
D-14/8/91	32845.0	0.23	7.6	84.0	251.0	98.0	71.4	2.0	866	7.6	
D-16/8/91	27933.0	0.20	7.6	158.0	375.0	178.0	60.7	3.5	1049	7.7	
D-18/8/91	27527.0	0.20	7.3	191.0	240.0	166.0	73.5	3.0	1072	7.4	
D-19/8/91	32363.0	0.10	7.6	159.0	310.0	146.0	68.5	1.6	1096	7.6	
D-20/8/91	31437.0	0.47	7.6	132.0	304.0	148.0	64.9	2.0	939	7.7	
D-21/8/91	31914.0	2.00	7.7	127.0	274.0	144.0	72.2	2.0	1031	7.6	
D-22/8/91	28088.0	0.20	7.5	153.0	307.0	124.0	82.3	2.5	1044	7.6	
D-23/8/91	27838.0	0.13	7.6	179.0	265.0	128.0	71.9	1.8	992	7.6	

D-25/8/91	29271.0	0.36	7.5	99.0	585.0	140.0	71.4	4.5	962	7.6
D-26/8/91	32723.0	0.16	7.7	93.0	252.0	176.0	56.8	2.3	894	7.7
D-27/8/91	33535.0	0.32	7.8	192.0	346.0	172.0	68.6	4.0	988	7.8
D-28/8/91	32922.0	0.30	7.4	139.0	367.0	180.0	64.4	3.0	1060	7.5
D-29/8/91	32190.0	0.30	7.3	200.0	545.0	258.0	65.1	4.0	1260	7.4
D-30/8/91	30488.0	0.21	7.5	152.0	300.0	132.0	69.7	4.5	1073	7.4

	...	29	30	31	32	33	34	35	36	37	\
0	...										
D-1/3/90	...	2000.0	39.6	58.8	95.5	85.4	70.0	90.2	79.40	87.3	
D-2/3/90	...	2590.0	39.6	60.7	94.8	85.4	80.8	90.2	79.50	92.1	
D-4/3/90	...	1888.0	39.6	58.2	95.6	85.4	52.9	90.2	75.80	88.7	
D-5/3/90	...	1840.0	33.1	64.2	95.3	87.3	72.3	90.2	82.30	89.6	
D-6/3/90	...	2120.0	39.6	62.7	95.6	85.4	71.0	92.1	78.20	87.5	
D-7/3/90	...	1764.0	39.6	59.7	96.5	86.7	78.3	90.1	73.10	84.9	
D-8/3/90	...	1703.0	39.6	61.9	93.8	89.1	79.8	90.2	86.20	90.1	
D-9/3/90	...	1606.0	39.6	70.4	95.6	90.6	53.7	92.1	66.90	94.6	
D-11/3/90	...	1338.0	46.1	43.6	92.5	85.6	58.2	92.2	73.80	90.2	
D-12/3/90	...	1616.0	21.2	59.7	90.8	88.4	66.1	89.0	69.00	86.5	
D-13/3/90	...	1575.0	0.6	45.8	92.0	11.6	25.7	19.6	36.00	43.0	
D-14/3/90	...	1304.0	31.8	61.2	92.9	85.4	69.9	90.2	79.15	18.5	
D-15/3/90	...	1221.0	46.6	48.5	91.7	85.4	20.4	26.3	31.70	10.3	
D-16/3/90	...	2550.0	41.1	50.0	84.0	76.5	52.9	84.3	54.10	43.7	
D-18/3/90	...	1545.0	32.7	33.3	90.0	82.6	61.3	87.0	71.40	78.2	
D-19/3/90	...	1110.0	28.0	72.7	92.7	78.8	36.6	85.9	60.40	90.7	
D-20/3/90	...	1402.0	49.8	58.3	92.3	89.9	96.8	94.4	98.10	92.1	
D-21/3/90	...	1606.0	39.6	39.6	78.3	85.5	70.4	91.3	73.40	89.4	
D-22/3/90	...	1780.0	35.3	57.5	92.0	75.6	49.3	80.8	60.80	76.1	
D-23/3/90	...	1597.0	20.1	64.5	93.6	86.9	68.1	87.4	72.20	88.3	
D-25/3/90	...	1223.0	29.4	55.4	91.3	85.2	68.0	88.4	81.40	92.9	
D-26/3/90	...	1706.0	28.1	50.0	91.4	85.2	76.5	87.8	82.50	88.5	
D-27/3/90	...	1869.0	52.0	64.9	90.8	87.6	71.2	90.3	77.60	92.5	
D-28/3/90	...	1416.0	26.7	66.2	95.0	89.2	72.5	91.1	78.70	88.1	
D-29/3/90	...	2290.0	34.3	65.0	94.2	89.3	69.7	89.7	78.10	89.1	
D-30/3/90	...	1475.0	27.2	67.6	97.3	90.4	68.0	91.4	78.00	90.4	
D-1/2/90	...	1377.0	39.6	50.0	89.4	85.4	76.6	90.2	79.50	90.6	
D-2/2/90	...	1220.0	6.8	47.9	80.0	84.6	78.9	80.0	80.80	82.4	
D-4/2/90	...	1087.0	42.7	54.5	95.7	84.9	73.6	88.1	78.90	83.3	
D-5/2/90	...	1275.0	39.6	33.0	94.5	85.3	73.7	84.4	76.80	80.7	
...	...	...	...	...	...	...	...	...	...	...	
D-25/10/91	...	1840.0	47.3	61.3	94.0	76.4	69.9	86.5	82.40	90.7	
D-26/10/91	...	1337.0	39.6	38.6	93.3	85.4	87.0	92.7	95.00	91.8	
D-27/10/91	...	1799.0	39.6	40.4	95.0	85.4	72.9	90.9	79.90	90.7	
D-29/10/91	...	1468.0	32.4	40.4	88.0	87.8	77.6	91.3	85.30	83.8	
D-30/10/91	...	1568.0	39.6	43.9	65.3	85.4	75.3	90.2	81.20	89.7	
D-1/8/91	...	1772.0	29.1	53.5	92.0	85.7	64.2	90.6	72.70	85.6	
D-2/8/91	...	1549.0	34.2	47.7	99.0	84.4	66.1	90.3	73.30	88.4	
D-4/8/91	...	1259.0	39.6	52.3	92.3	75.3	69.9	86.3	79.15	76.6	
D-5/8/91	...	1100.0	55.1	73.2	96.0	82.0	62.5	91.7	78.90	88.7	
D-6/8/91	...	951.0	44.8	64.5	95.0	82.2	80.0	89.4	86.10	88.8	
D-7/8/91	...	898.0	39.6	64.1	94.0	84.9	60.4	91.9	76.80	87.6	
D-8/8/91	...	884.0	48.7	86.9	98.8	81.0	67.9	90.4	91.60	96.0	
D-9/8/91	...	947.0	39.6	47.6	97.3	85.4	77.6	93.7	84.80	90.3	
D-11/8/91	...	728.0	62.8	77.7	90.0	73.8	65.5	92.7	77.20	90.7	
D-12/8/91	...	929.0	36.9	43.6	86.7	87.7	71.9	91.3	79.80	92.3	
D-13/8/91	...	858.0	38.1	41.6	86.7	87.7	79.0	94.2	87.70	91.6	
D-14/8/91	...	879.0	50.9	36.5	80.0	87.0	69.6	91.7	80.50	82.7	

```
D-16/8/91   ...     828.0  68.0  66.7  96.7  81.6  41.8  94.3  72.50  83.1
D-18/8/91   ...     999.0  38.5  54.5  88.0  90.0  69.9  95.8  81.70  93.4
D-19/8/91   ...    1083.0  25.2  61.4  91.2  78.6  65.1  86.8  81.00  89.0
D-20/8/91   ...    1012.0  45.6  60.3  94.4  82.5  72.9  89.4  86.20  91.2
D-21/8/91   ...    1053.0  39.6  59.5  93.3  85.4  77.7  90.2  79.15  90.7
D-22/8/91   ...    1038.0  40.5  54.4  94.0  89.7  75.5  93.5  85.00  90.3
D-23/8/91   ...    1044.0  13.7  45.0  95.0  87.5  71.3  93.9  79.60  89.1
D-25/8/91   ...     968.0  40.8  71.6  93.3  59.0  26.4  74.7  83.80  81.4
D-26/8/91   ...     942.0  39.6  62.3  93.3  69.8  75.9  79.6  78.60  96.6
D-27/8/91   ...     950.0  39.6  58.3  97.8  83.0  59.1  91.1  74.60  90.7
D-28/8/91   ...    1136.0  39.6  65.0  97.1  76.2  66.4  82.0  77.10  88.9
D-29/8/91   ...    1326.0  39.8  65.9  97.1  81.7  70.9  89.5  87.00  89.5
D-30/8/91   ...    1224.0  39.6  69.5  93.3  81.7  76.4  90.2  81.70  86.4

              38
0
D-1/3/90     99.6
D-2/3/90    100.0
D-4/3/90     98.5
D-5/3/90    100.0
D-6/3/90     99.5
D-7/3/90    100.0
D-8/3/90     99.0
D-9/3/90    100.0
D-11/3/90    99.4
D-12/3/90    99.6
D-13/3/90    36.4
D-14/3/90    42.9
D-15/3/90    95.4
D-16/3/90   100.0
D-18/3/90    99.2
D-19/3/90   100.0
D-20/3/90   100.0
D-21/3/90    99.4
D-22/3/90    99.6
D-23/3/90   100.0
D-25/3/90   100.0
D-26/3/90    99.8
D-27/3/90   100.0
D-28/3/90   100.0
D-29/3/90   100.0
D-30/3/90   100.0
D-1/2/90     98.6
D-2/2/90    100.0
D-4/2/90    100.0
D-5/2/90    100.0
...            ...
D-25/10/91   99.8
D-26/10/91   95.7
D-27/10/91   98.6
D-29/10/91   96.7
D-30/10/91   99.2
D-1/8/91     99.3
D-2/8/91     99.3
D-4/8/91     99.0
D-5/8/91     99.6
D-6/8/91    100.0
```

```
D-7/8/91     99.3
D-8/8/91     99.7
D-9/8/91     99.3
D-11/8/91    98.6
D-12/8/91    99.3
D-13/8/91    98.8
D-14/8/91    99.0
D-16/8/91    98.6
D-18/8/91    99.3
D-19/8/91    99.4
D-20/8/91    99.5
D-21/8/91    99.7
D-22/8/91   100.0
D-23/8/91   100.0
D-25/8/91    99.7
D-26/8/91    99.6
D-27/8/91   100.0
D-28/8/91    99.0
D-29/8/91    99.8
D-30/8/91    99.7

[527 rows x 38 columns]
```

In [330]:
```
# Scale the variables
from sklearn.preprocessing import StandardScaler
scaled = StandardScaler()
scaled.fit(Water_cleaned)
Water_cleaned_scaled = scaled.transform(Water_cleaned) # Scaled dataset
```

In [331]:
```
print(Water_cleaned_scaled)
```

```
[[ 1.06097737 -0.31176336 -0.04089158 ...,  0.17091754 -0.20723325
   0.11050301]
 [ 0.28219456  0.23571845 -0.44749277 ...,  0.1824499   0.38144415
   0.20304983]
 [-0.76011965  0.96569418 -0.85409395 ..., -0.24424763 -0.03553567
  -0.14400075]
 ...,
 [-0.65381741 -0.7497488  -1.66729633 ..., -0.09432687 -0.01100745
  -0.02831722]
 [-0.76610203 -0.7497488  -2.07389751 ...,  1.04737733  0.06257723
   0.15677642]
 [-1.02717911 -0.7825977  -1.26069514 ...,  0.43616195 -0.31761026
   0.13363972]]
```

In [332]:
```
# Determine the two principal components among the features
from sklearn.decomposition import PCA
pca = PCA(n_components = 2)  # Request that two principal components be calcul
ated from the features
pca.fit(Water_cleaned_scaled)  # fit the dataset
pca_comps = pca.transform(Water_cleaned_scaled) # assign the principal compone
nts to a variable
```

In [333]:
```
print(pca_comps)
```

```
[[ 1.61324758 -0.29293352]
 [ 4.30829529 -0.48676991]
 [ 0.45812616 -1.06688971]
 ...,
 [-3.07358511 -0.47657264]
 [-1.68898355  0.39533681]
 [-2.98845529  0.52221904]]
```

In [334]:
```
print(pca_comps.shape)  # SHows the new shape of the dataframe

# notice that instead of 38 columns, it has been reduced to 2 columns
```

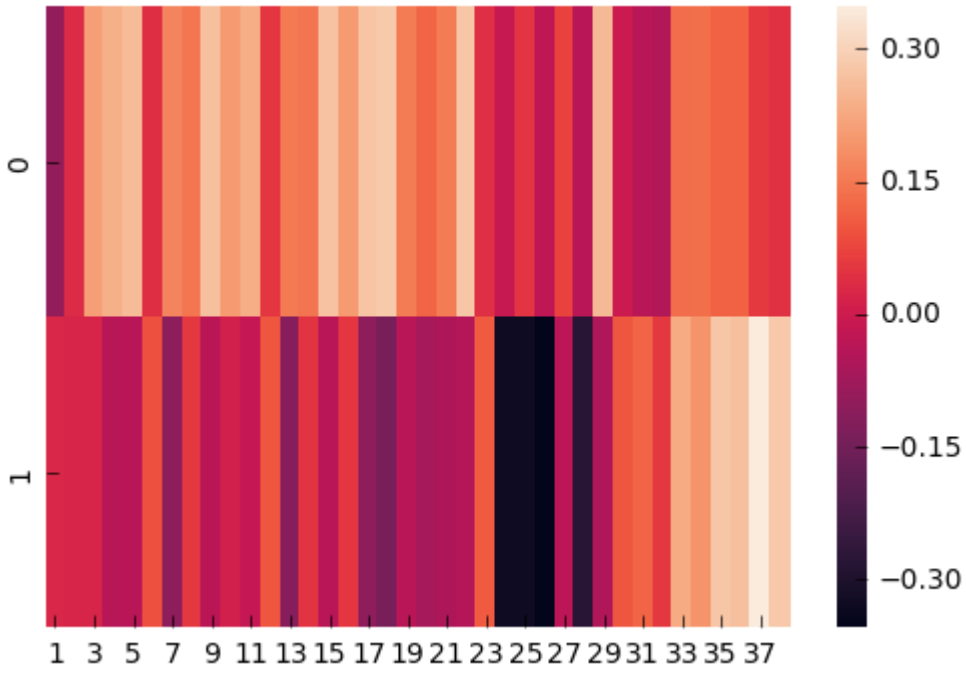
```
(527, 2)
```

In [335]:
```
# Show the PCA components
print(pca.components_)
```

```
[[-0.09607625  0.03404058  0.20829207  0.24007351  0.25809293  0.040229
   0.16702766  0.14580396  0.26566311  0.20340852  0.23675963  0.05168039
   0.15127654  0.14275293  0.27136624  0.20246357  0.27376737  0.28617692
   0.15425277  0.11842546  0.153793    0.27793829  0.04171031 -0.00904403
   0.05043745 -0.02302684  0.07194353 -0.03469232  0.25436505 -0.00426902
  -0.03510447 -0.04680218  0.13162004  0.13405186  0.11654486  0.11693155
   0.05690212  0.04311658]
 [ 0.02592851  0.0226068   0.02370611 -0.03887028 -0.03565239  0.09137956
  -0.103618    0.05467435 -0.03317637  0.01134908 -0.0169161   0.10020294
  -0.11585381  0.0467968  -0.03506975  0.05368288 -0.10234408 -0.14042989
  -0.03563338 -0.06542278 -0.05517845 -0.04312377  0.10349285 -0.32903646
  -0.32960542 -0.35439283 -0.03006309 -0.28014293 -0.05164938  0.09859565
   0.1193992   0.05455503  0.23046802  0.19290569  0.27690464  0.2666469
   0.34737629  0.28368542]]
```

In [336]:
```
# Create a visual heat map of the prinmcipal components
df_pca_comps = pd.DataFrame(data = pca.components_ , columns =
np.arange(1,39))
sns.heatmap(data = df_pca_comps)
plt.figure(figsize=(40,10))
plt.show()
```

```
<matplotlib.figure.Figure at 0xc050090>
```




```
<matplotlib.figure.Figure at 0xc0cab30>
```

In [337]:
```
np.arange(1,39)
```

Out[337]:
```
array([ 1,  2,  3,  4,  5,  6,  7,  8,  9, 10, 11, 12, 13, 14, 15, 16, 17,
       18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34,
       35, 36, 37, 38])
```

In [339]:
```
print(df_pca_comps)
```

```
          1         2         3         4         5         6         7   \
0 -0.096076  0.034041  0.208292  0.240074  0.258093  0.040229  0.167028
1  0.025929  0.022607  0.023706 -0.038870 -0.035652  0.091380 -0.103618

          8         9         10    ...           29        30        31  \
0  0.145804  0.265663  0.203409    ...     0.254365 -0.004269 -0.035104
1  0.054674 -0.033176  0.011349    ...    -0.051649  0.098596  0.119399

          32        33        34        35        36        37        38
0 -0.046802  0.131620  0.134052  0.116545  0.116932  0.056902  0.043117
1  0.054555  0.230468  0.192906  0.276905  0.266647  0.347376  0.283685

[2 rows x 38 columns]
```

In [340]:
```
from scipy.cluster.hierarchy import dendrogram, linkage
```

In [341]:
```
Water_matrix = np.array(Water_cleaned_scaled)
```

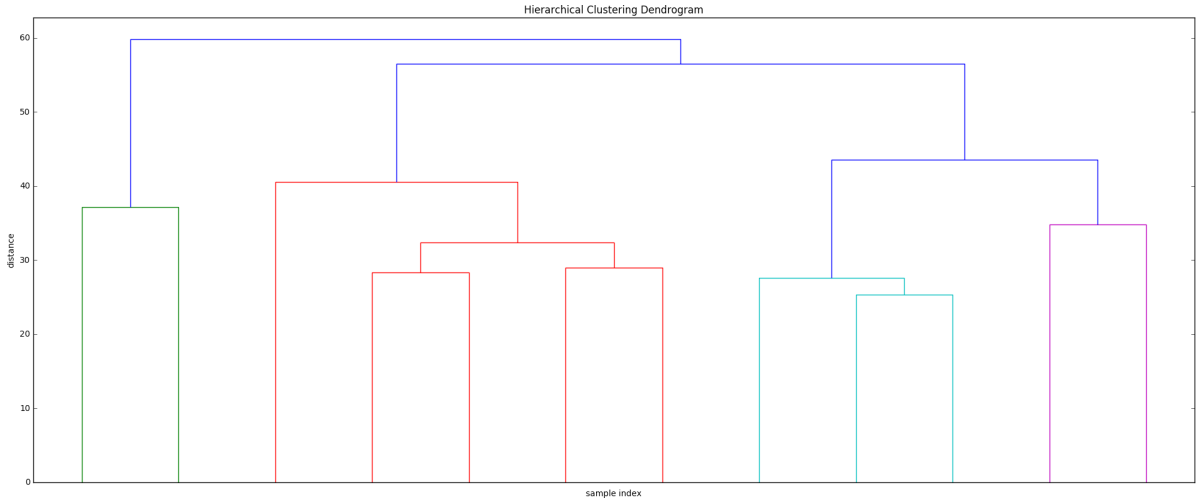
In [342]:
```
# generate the linkage matrix
Z = linkage(Water_matrix, 'ward')
```

In [343]:
```
from scipy.cluster.hierarchy import cophenet
from scipy.spatial.distance import pdist
```

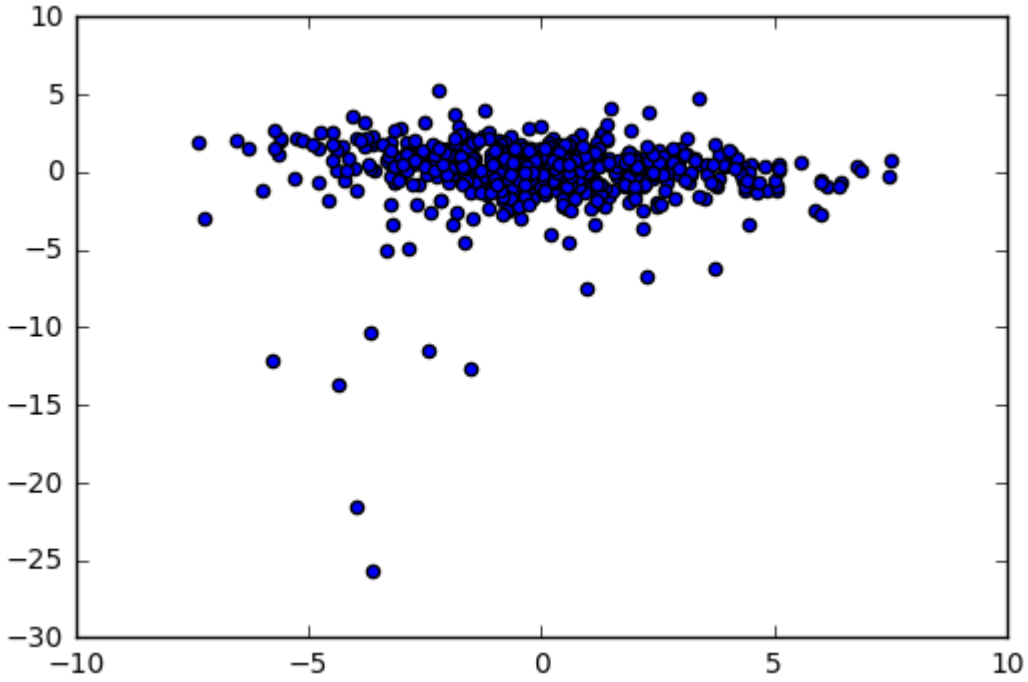
In [344]:
```
c, coph_dists = cophenet(Z, pdist(Water_matrix))
print(c)
```

```
0.298966756872
```

In [345]:
```
# calculate full dendrogram
plt.figure(figsize=(25, 10))
plt.title('Hierarchical Clustering Dendrogram')
plt.xlabel('sample index')
plt.ylabel('distance')
dendrogram(
    Z,
    truncate_mode='lastp',  # show only the last p merged clusters
    p=12,  # show only the last p merged clusters
    show_leaf_counts=False,  # otherwise numbers in brackets are counts
    leaf_rotation=90.,  # rotates the x axis labels
    leaf_font_size=8.,  # font size for the x axis labels
)
plt.show()
```



In [346]:
```
plt.scatter(pca_comps.transpose()[0], pca_comps.transpose()[1])
plt.show()
```



In [348]:
```
print(Water_matrix.shape)
```

```
(527, 38)
```

In [349]:
```
# Import Kmeans
from sklearn.cluster import KMeans
```

In [350]:
```
# Create an instance of a Kmeans model with 4 clusters
kmeans = KMeans(n_clusters = 4)
```

In [351]:
```
# fit the data
kmeans.fit(Water_cleaned_scaled)
```

Out[351]:
```
KMeans(copy_x=True, init='k-means++', max_iter=300, n_clusters=4, n_init=10,
    n_jobs=1, precompute_distances='auto', random_state=None, tol=0.0001,
    verbose=0)
```

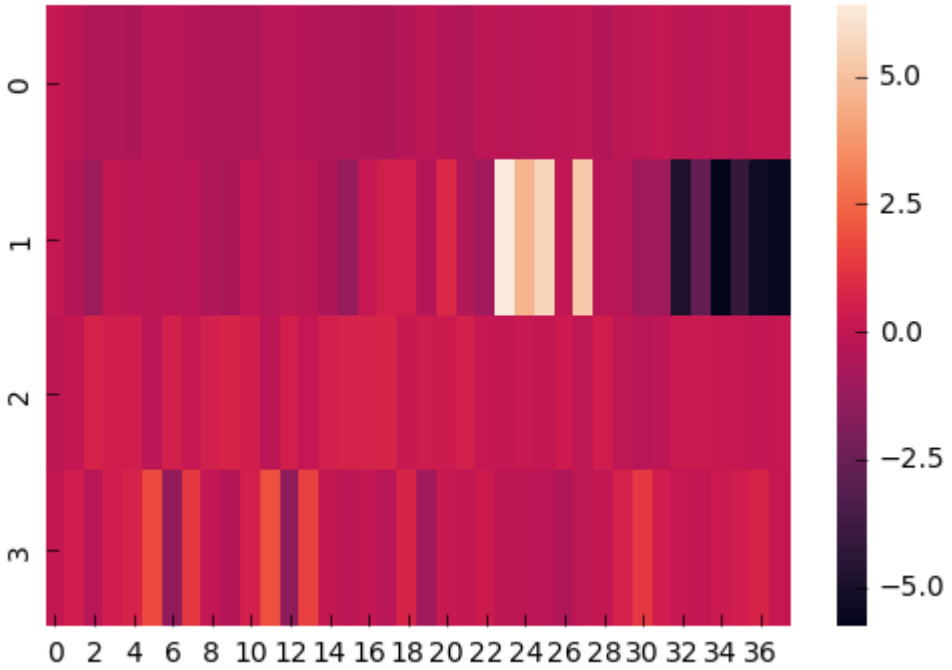
In [354]:
```
print(kmeans.cluster_centers_.shape)
```

```
(4, 38)
```

In [355]:
```
print(kmeans.cluster_centers_)
```

```
[[  7.44631005e-02  -1.04508645e-01  -4.96284909e-01  -4.75560726e-01
   -5.53981098e-01  -2.34544662e-01  -2.00086601e-01  -3.55889313e-01
   -4.61135125e-01  -5.15113282e-01  -5.30001446e-01  -2.76004456e-01
   -1.72102716e-01  -3.96802827e-01  -4.69883380e-01  -5.01264436e-01
   -5.76768057e-01  -5.91260066e-01  -3.80515538e-01  -1.09953146e-01
   -3.31590081e-01  -4.87507083e-01  -1.21214736e-01  -1.82521561e-01
   -2.64345137e-01  -1.38915710e-01  -1.11561273e-01  -7.37445111e-02
   -4.31578861e-01  -8.16547210e-02  -3.26312569e-02   7.86622579e-02
   -1.13640279e-01  -1.48247419e-01  -4.05549479e-02  -1.05720851e-01
    2.34301794e-02   5.42345440e-02]
 [  6.95462438e-02  -4.31687940e-01  -1.02835160e+00   6.57169026e-02
   -1.49040335e-01  -7.75894413e-02  -1.63763121e-01  -2.09608344e-01
   -4.76286869e-01  -6.36350928e-01   1.08098429e-02  -3.09987192e-01
   -3.78493793e-02  -1.90675922e-01  -5.85427435e-01  -1.20622006e+00
    1.53992793e-01   5.34520413e-01   5.28617416e-01  -3.98734820e-01
    8.54242030e-01  -4.48645017e-01  -8.93101092e-01   6.38031986e+00
    4.64385846e+00   5.68521702e+00  -5.05936155e-02   5.28441679e+00
   -2.42371456e-01  -2.68845718e-01  -9.76339435e-01  -1.00801592e+00
   -4.86510250e+00  -2.67346962e+00  -5.77410168e+00  -4.09688152e+00
   -5.28457576e+00  -5.51832681e+00]
 [ -9.39950896e-02   3.30999512e-02   6.71593068e-01   4.47730285e-01
    4.85033927e-01  -1.53412193e-01   5.56897143e-01   9.53555425e-02
    5.40691397e-01   6.95311497e-01   4.84276675e-01  -1.34753236e-01
    5.32445641e-01   9.50608560e-02   5.50581577e-01   6.50880811e-01
    6.50482264e-01   7.05399471e-01   2.54305186e-01   3.57344789e-01
    3.11869777e-01   5.60096122e-01   9.93218200e-02   2.81574380e-02
    1.76114944e-01   2.09849510e-02   2.58505435e-01  -6.82652271e-02
    4.91709690e-01  -4.25595300e-02  -2.52365974e-01  -1.62027967e-01
    2.64338752e-01   2.56844568e-01   1.68835342e-01   1.46133532e-01
   -2.60844302e-02   8.46068821e-02]
 [  2.66986670e-02   4.21792067e-01  -2.91107695e-01   4.00480023e-01
    6.50022322e-01   1.78121442e+00  -1.35068433e+00   1.33946305e+00
    1.43666292e-02  -3.52790946e-01   5.16391874e-01   1.93360057e+00
   -1.39966395e+00   1.53484945e+00   2.96548726e-02  -1.56391681e-01
    2.51717541e-02  -1.86609826e-01   6.92116000e-01  -9.15802572e-01
    1.71734897e-01   5.60832949e-02   2.86600671e-01  -9.94263516e-02
   -9.25727447e-02  -1.85429056e-01  -5.40571371e-01  -7.06340433e-02
    4.61269798e-02   6.07184189e-01   1.34630716e+00   4.36781611e-01
    9.30757581e-02  -4.15334194e-03   2.64448924e-01   4.47139874e-01
    7.08094063e-01   1.27410603e-01]]
```

In [356]:
```
sns.heatmap(data = kmeans.cluster_centers_)
plt.figure(figsize=(40,10))
plt.show()
```




```
<matplotlib.figure.Figure at 0xbc11890>
```

In [357]:
```
C = kmeans.cluster_centers_.transpose()
```

In [359]:
```
print(C.shape)
```

```
(38, 4)
```

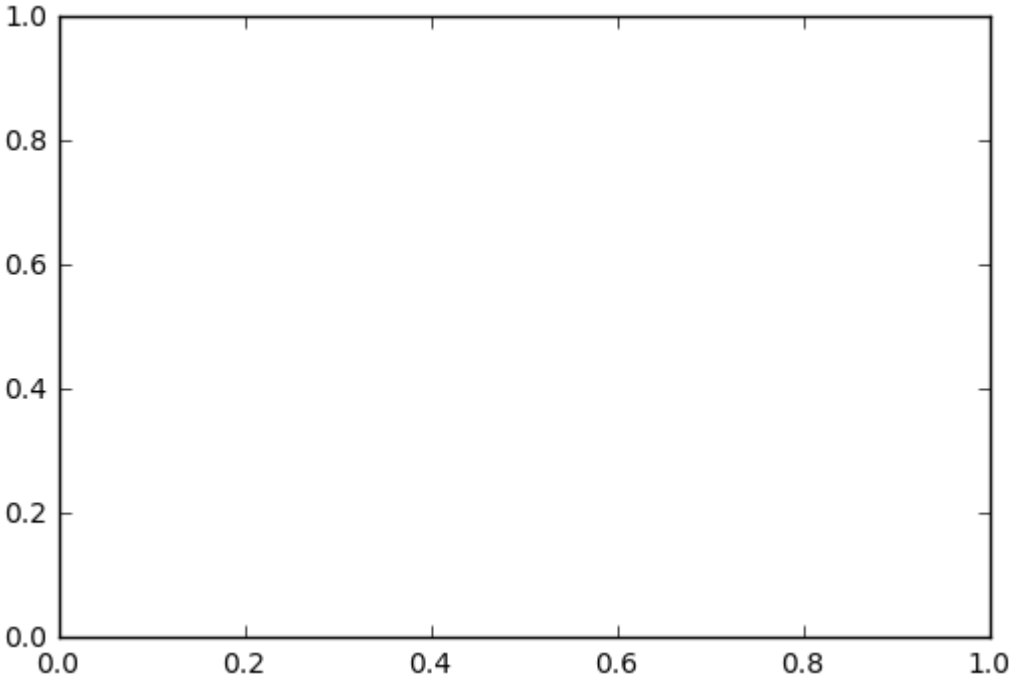
In [360]:
```
kmeans.labels_.shape
```

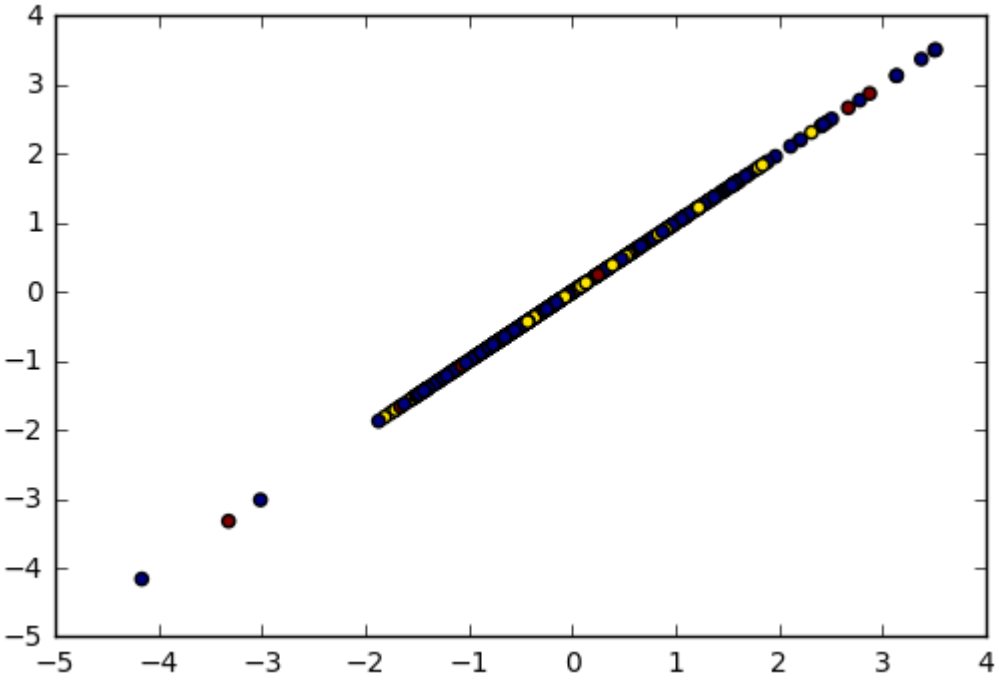
Out[360]:
```
(527,)
```

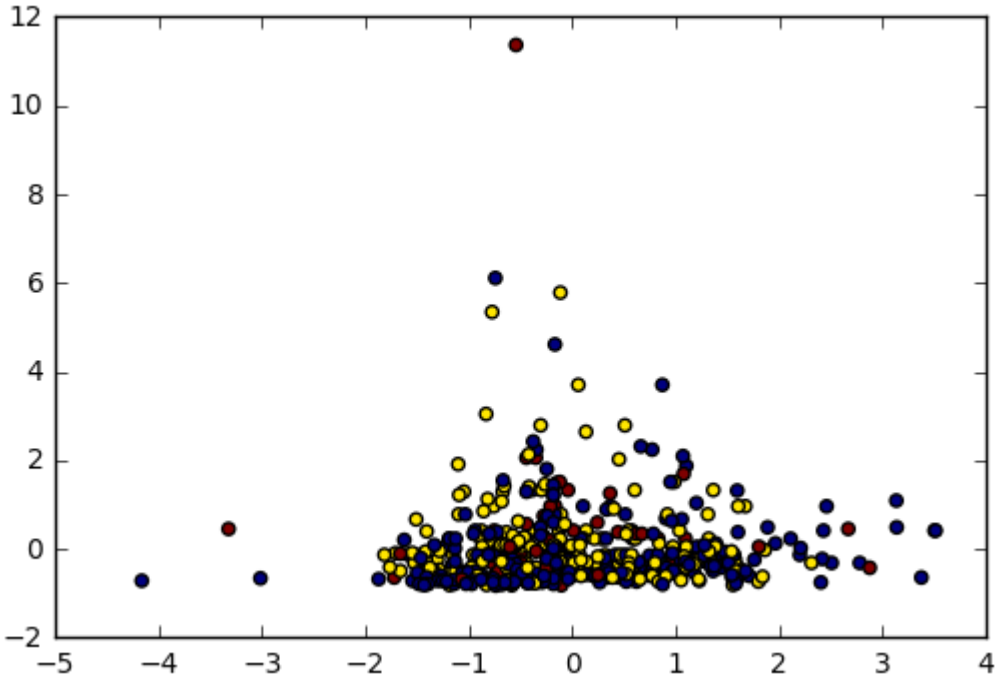
In [361]:
```
C[:,0].shape
```

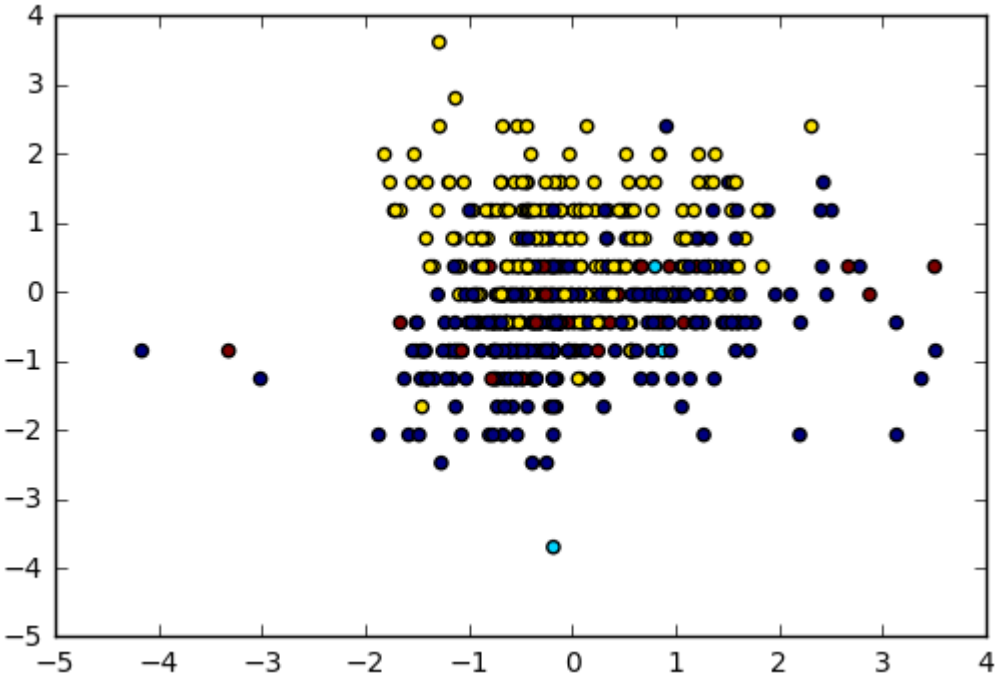
Out[361]:
```
(38,)
```

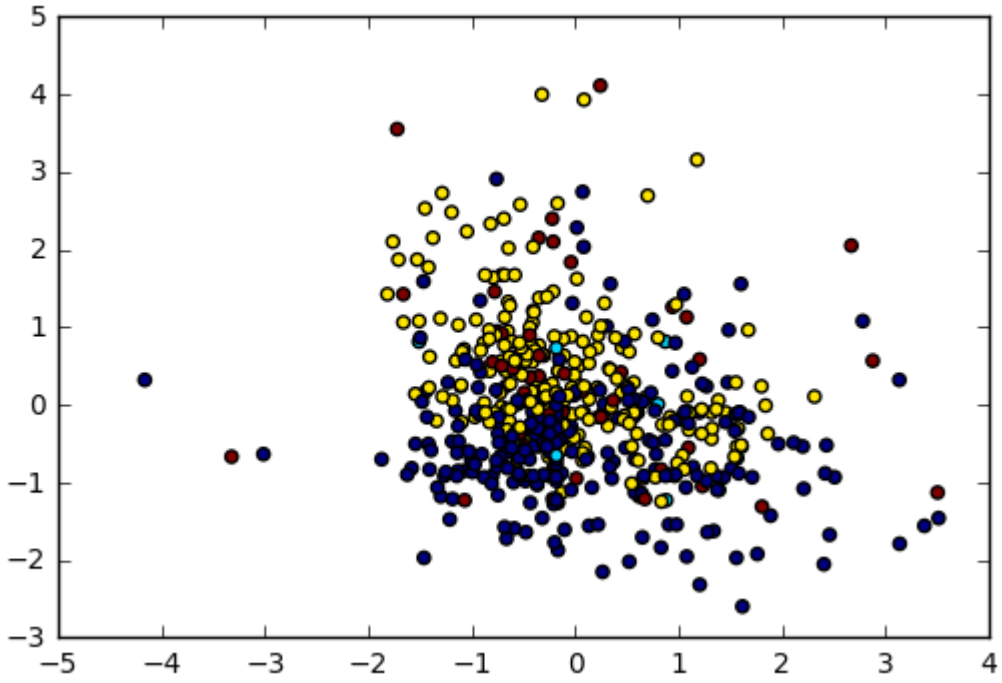
In [363]:
```
m,n = Water_matrix.shape
for i in range(n):
    #print(i)
    plt.figure()
    plt.scatter(Water_matrix[:,0],Water_matrix[:,i], c = kmeans.labels_)
    plt.show()
```

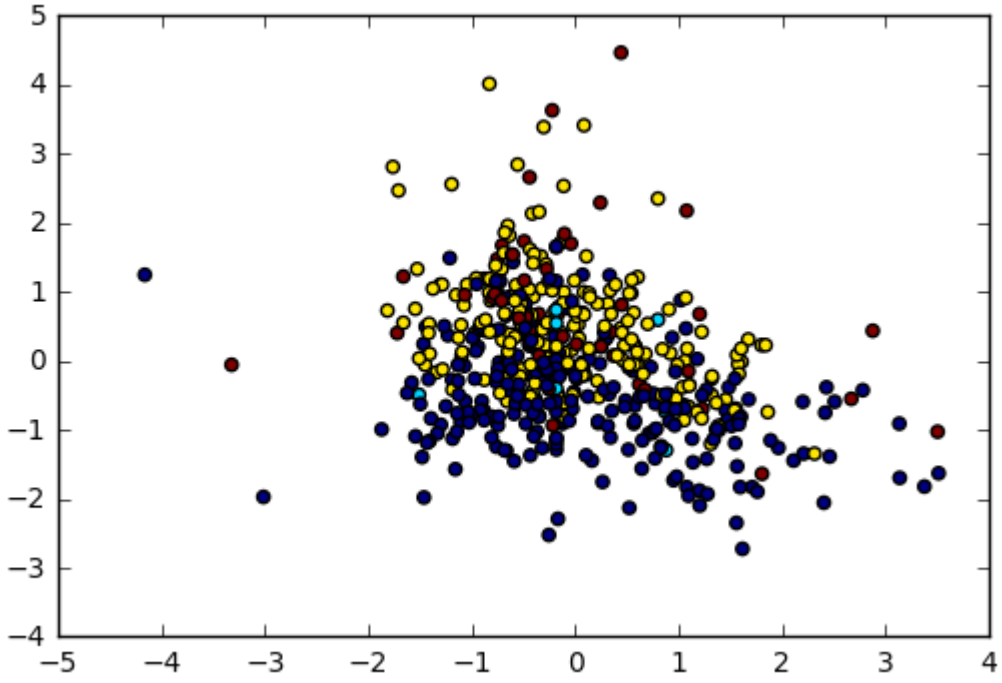

1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0

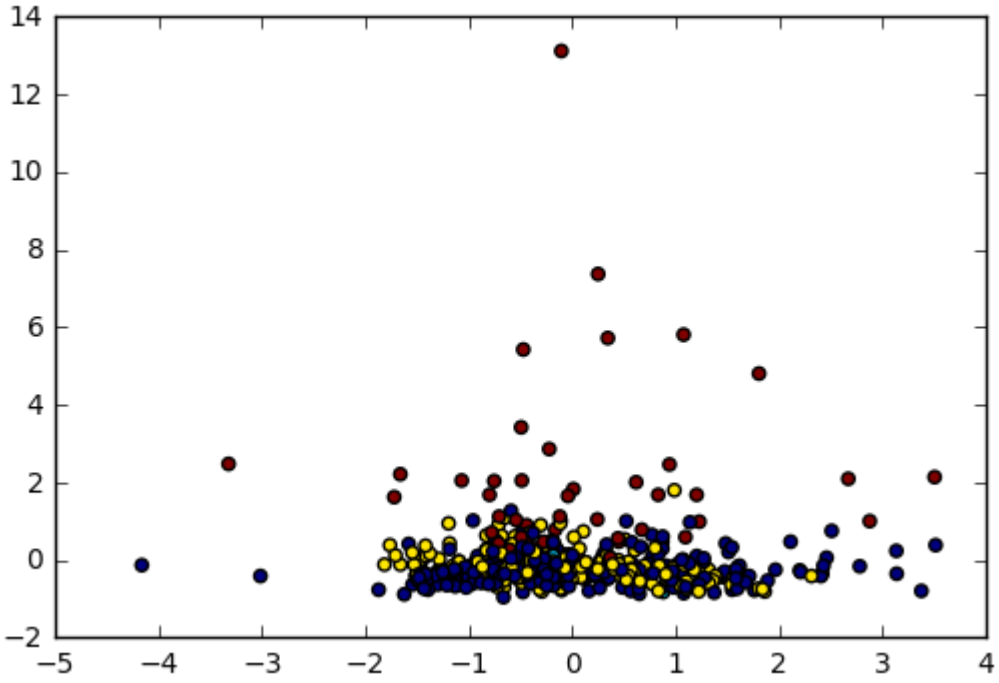

4
3
2
1
0
−1
−2
−3
−4
−5
−5
−4
−3
−2
−1
0
1
2
3
4

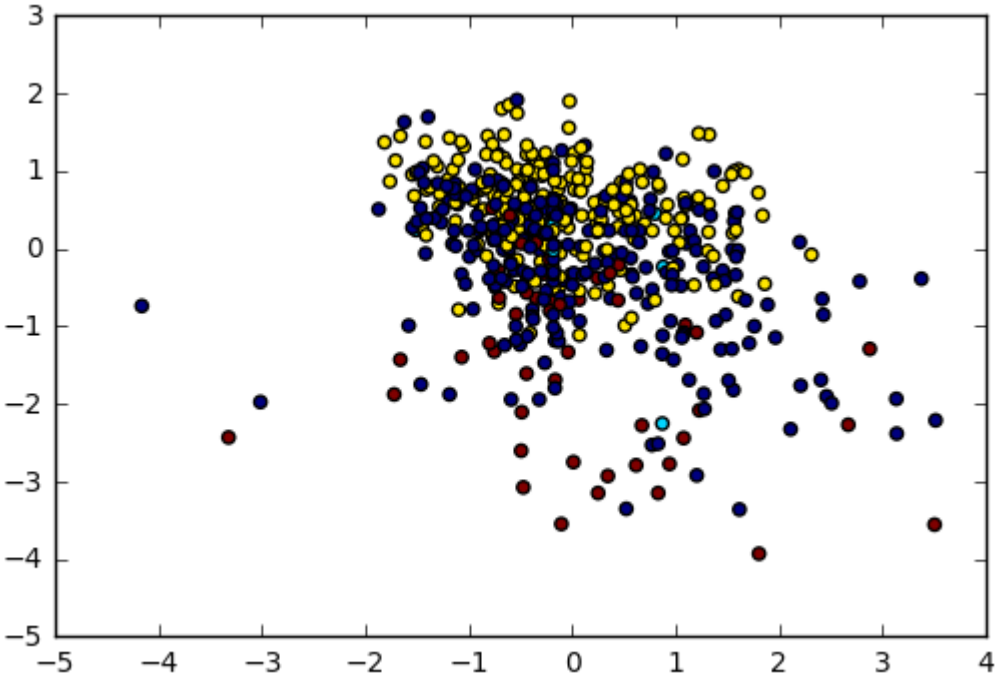

12
10
8
6
4
2
0
−2
−5
−4
−3
−2
−1
0
1
2
3
4

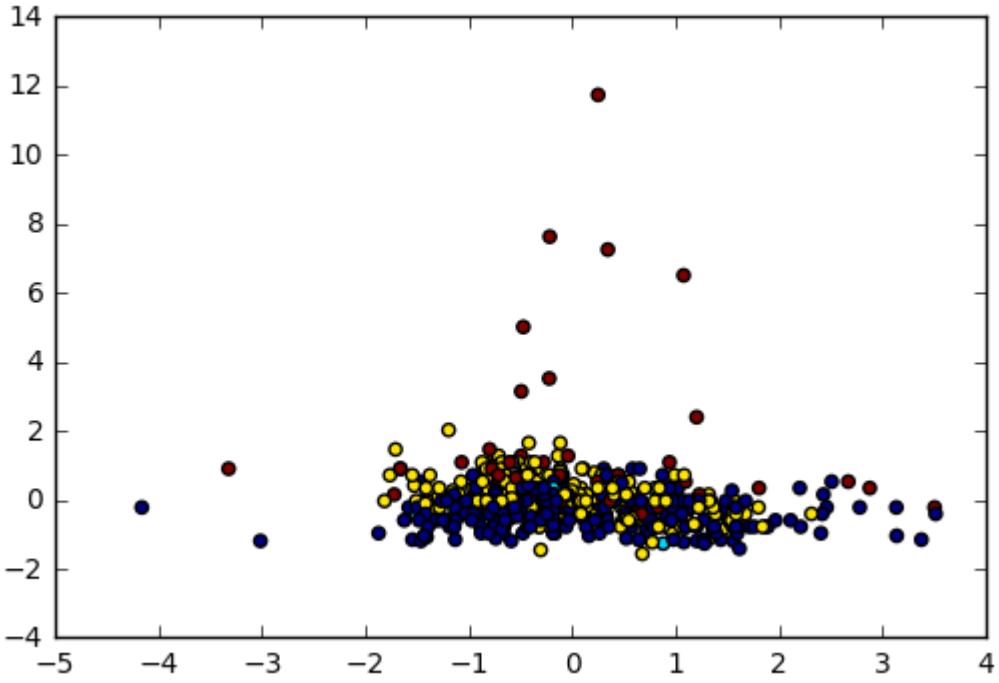

4
3
2
1
0
−1
−2
−3
−4
−5
−5
−4
−3
−2
−1
0
1
2
3
4

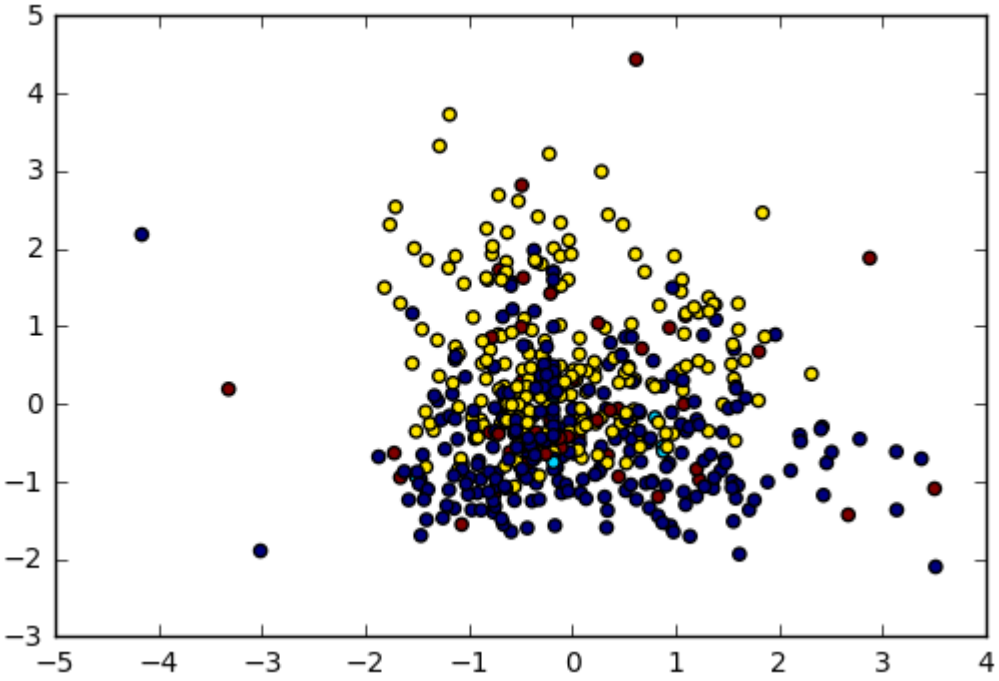

5
4
3
2
1
0
−1
−2
−3
−5
−4
−3
−2
−1
0
1
2
3
4

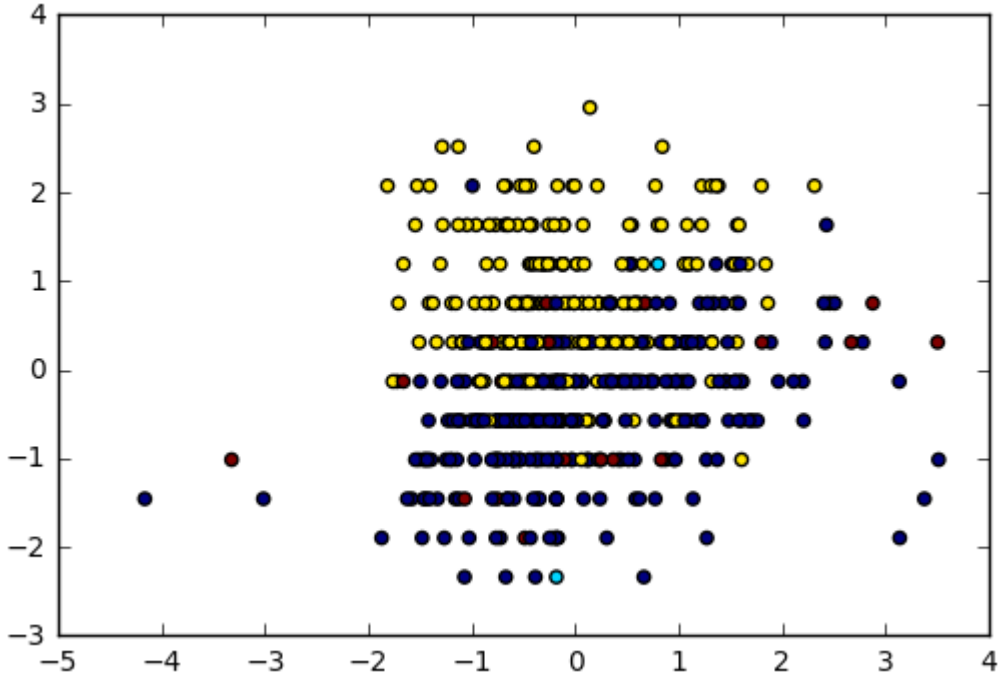

5
4
3
2
1
0
−1
−2
−3
−4
−5
−4
−3
−2
−1
0
1
2
3
4

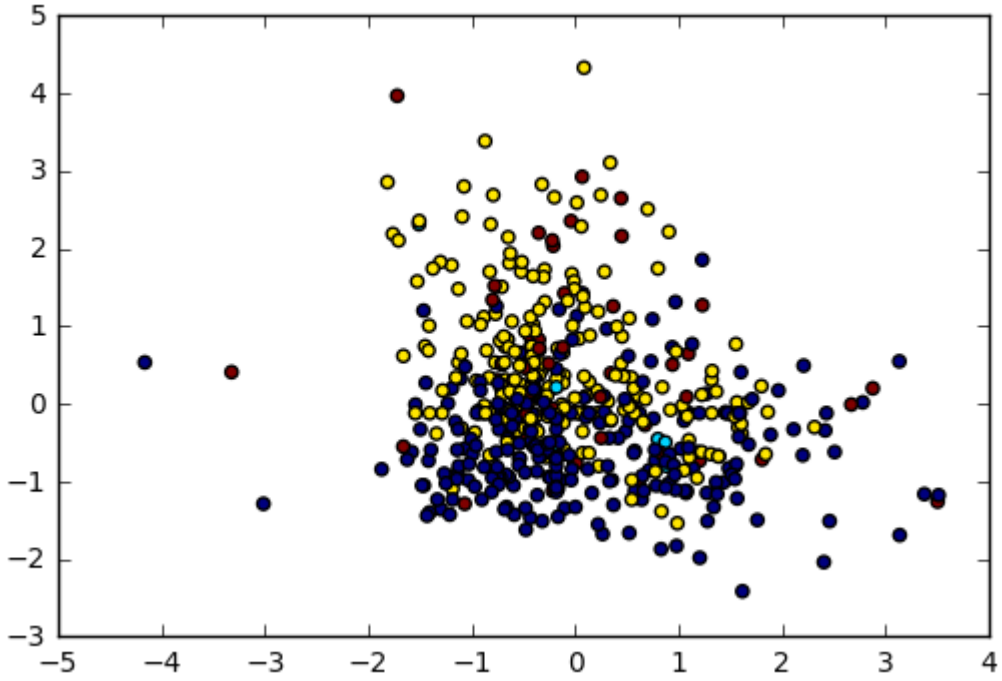

14
12
10
8
6
4
2
0
−2
−5
−4
−3
−2
−1
0
1
2
3
4

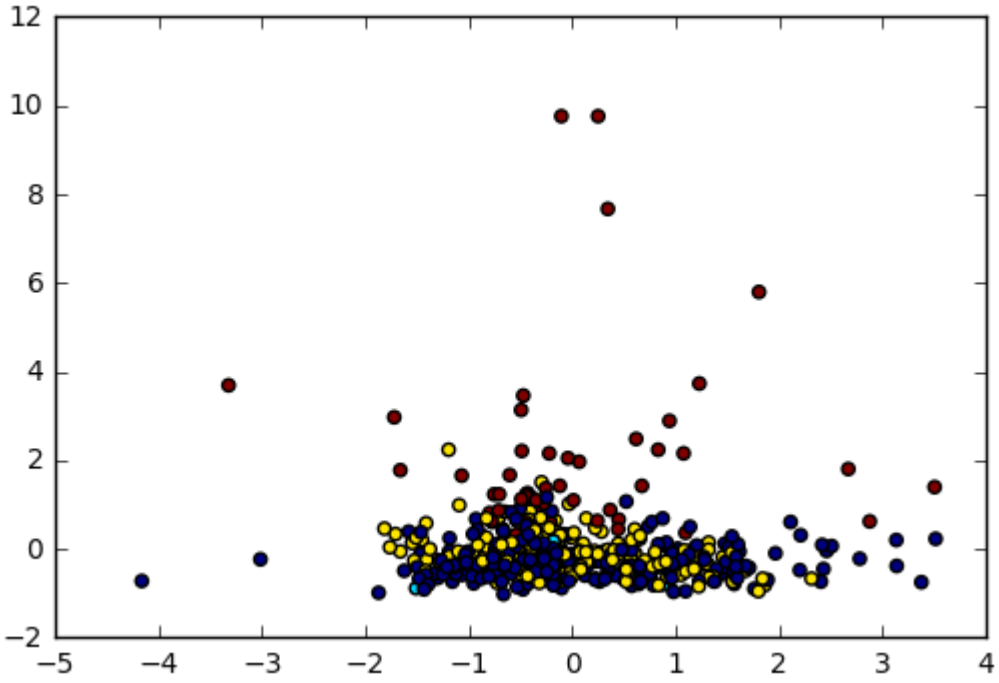

3
2
1
0
−1
−2
−3
−4
−5
−5
−4
−3
−2
−1
0
1
2
3
4

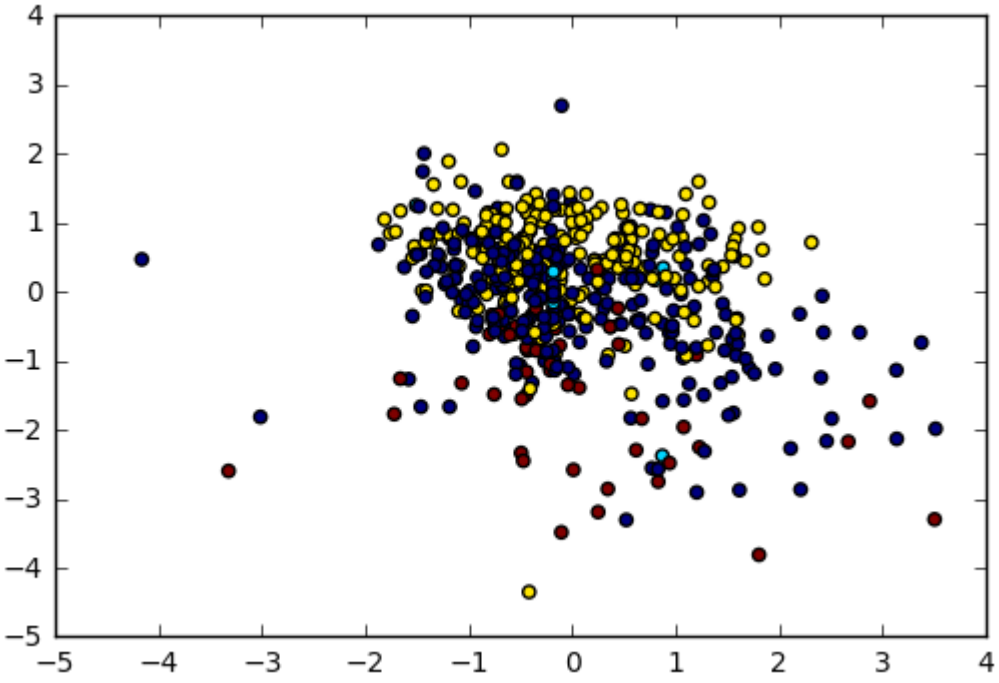

14
12
10
8
6
4
2
0
−2
−4
−5
−4
−3
−2
−1
0
1
2
3
4

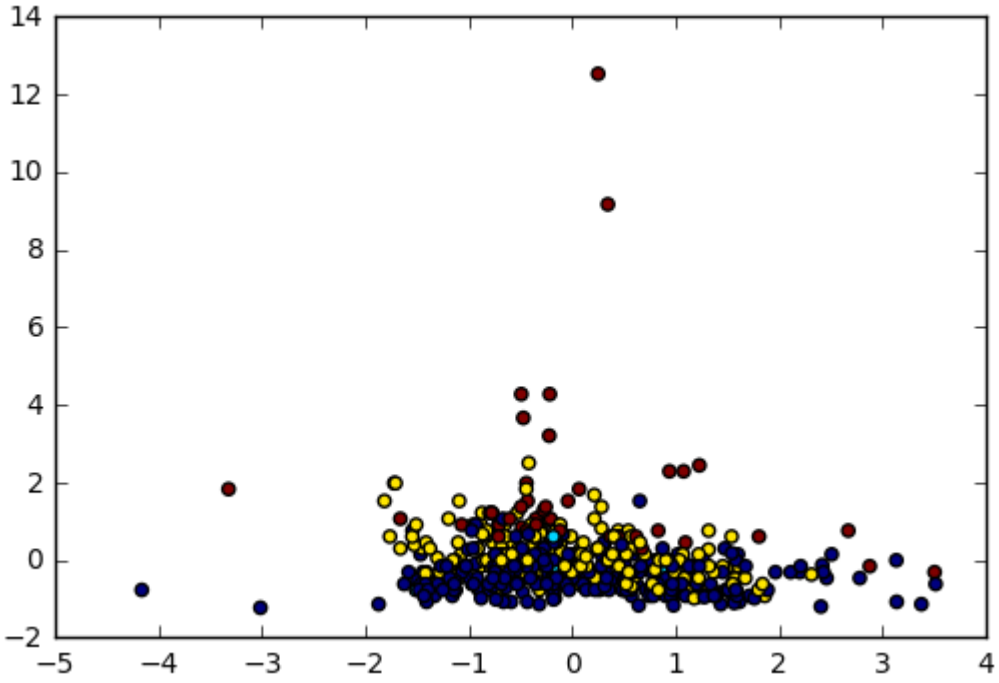

5
4
3
2
1
0
−1
−2
−3
−5
−4
−3
−2
−1
0
1
2
3
4

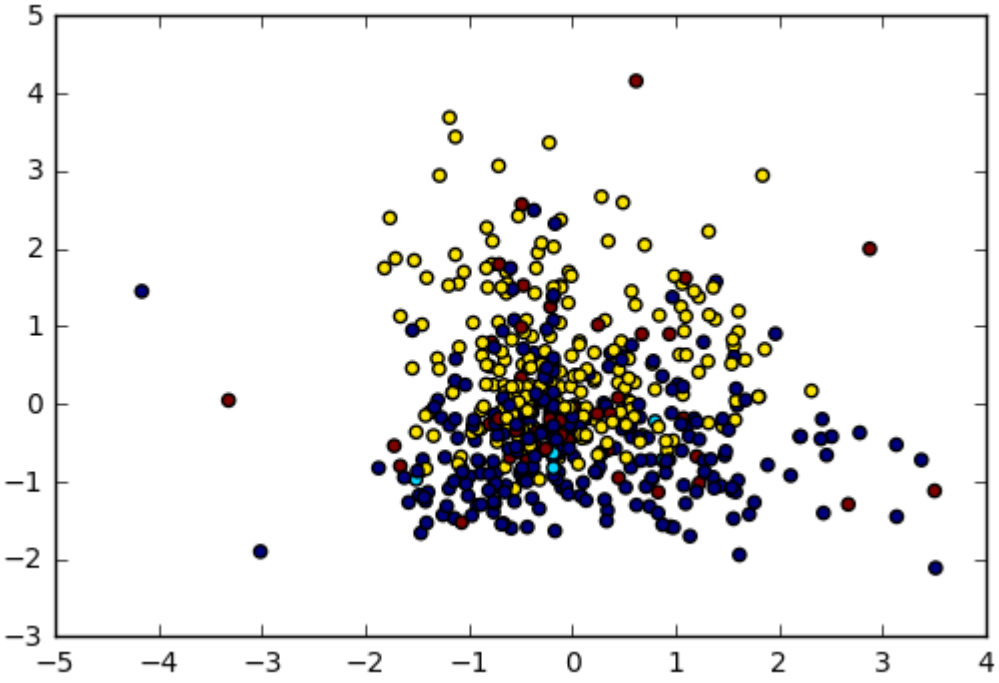

4
3
2
1
0
−1
−2
−3
−5
−4
−3
−2
−1
0
1
2
3
4

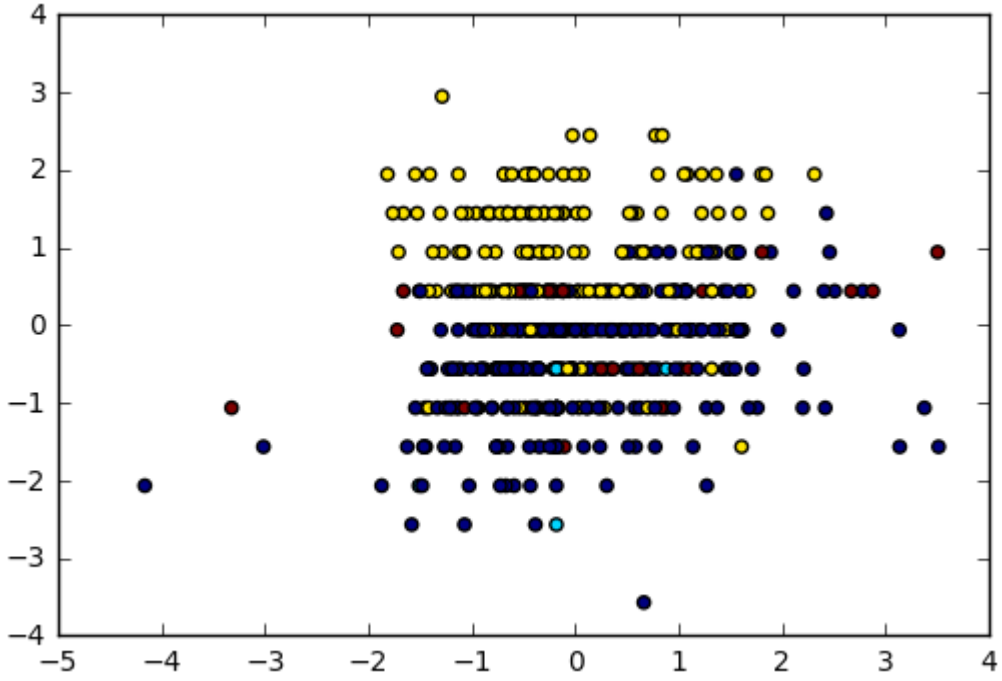

5
4
3
2
1
0
−1
−2
−3
−5
−4
−3
−2
−1
0
1
2
3
4

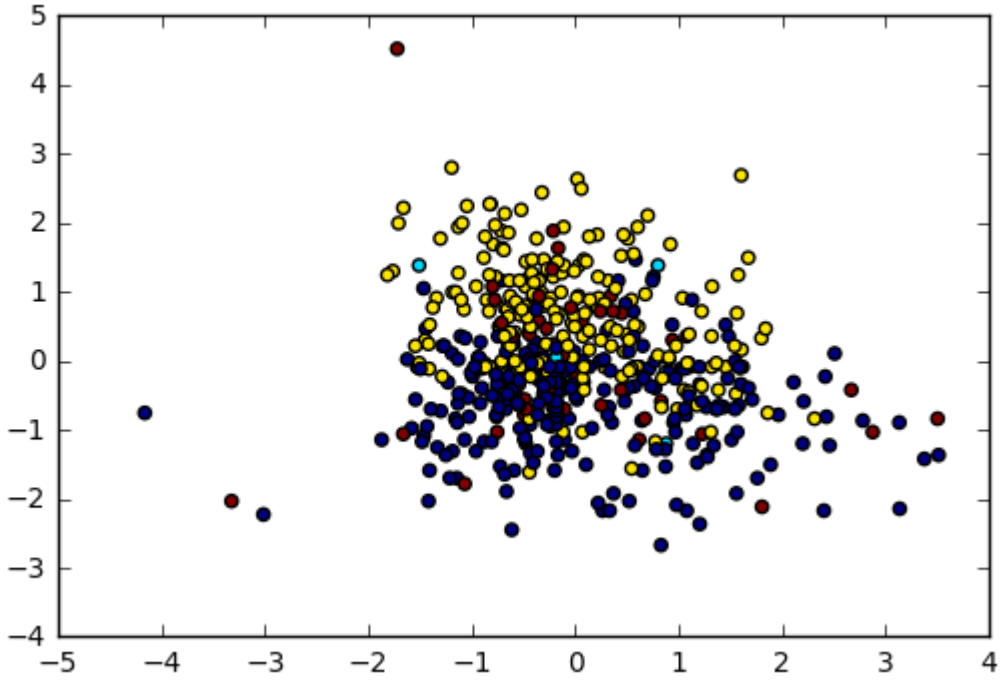

12
10
8
6
4
2
0
−2
−5
−4
−3
−2
−1
0
1
2
3
4

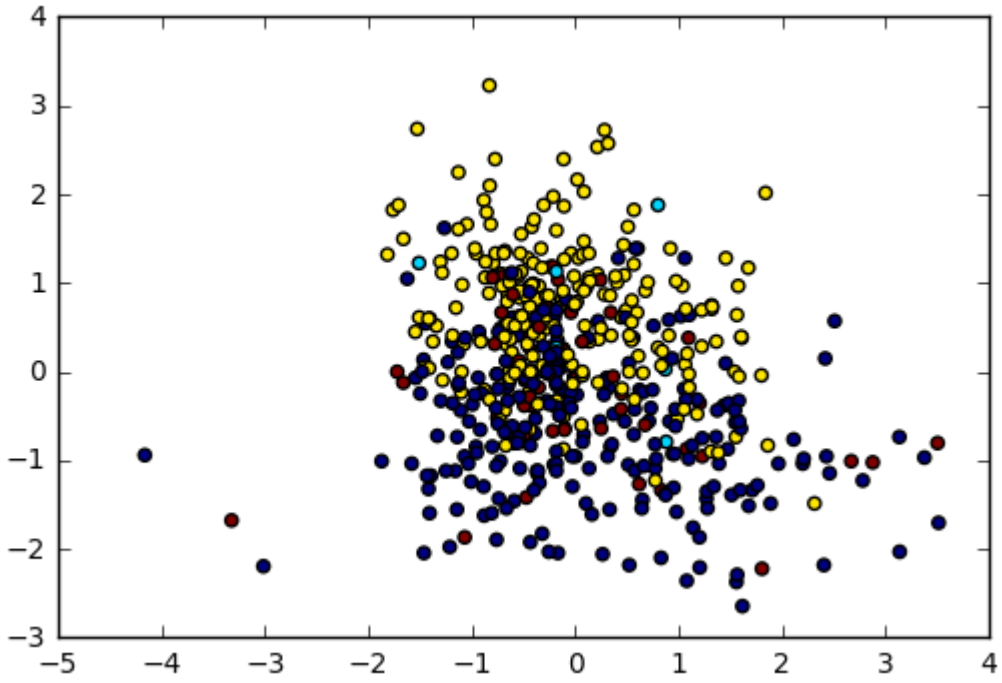

4
3
2
1
0
−1
−2
−3
−4
−5
−5
−4
−3
−2
−1
0
1
2
3
4

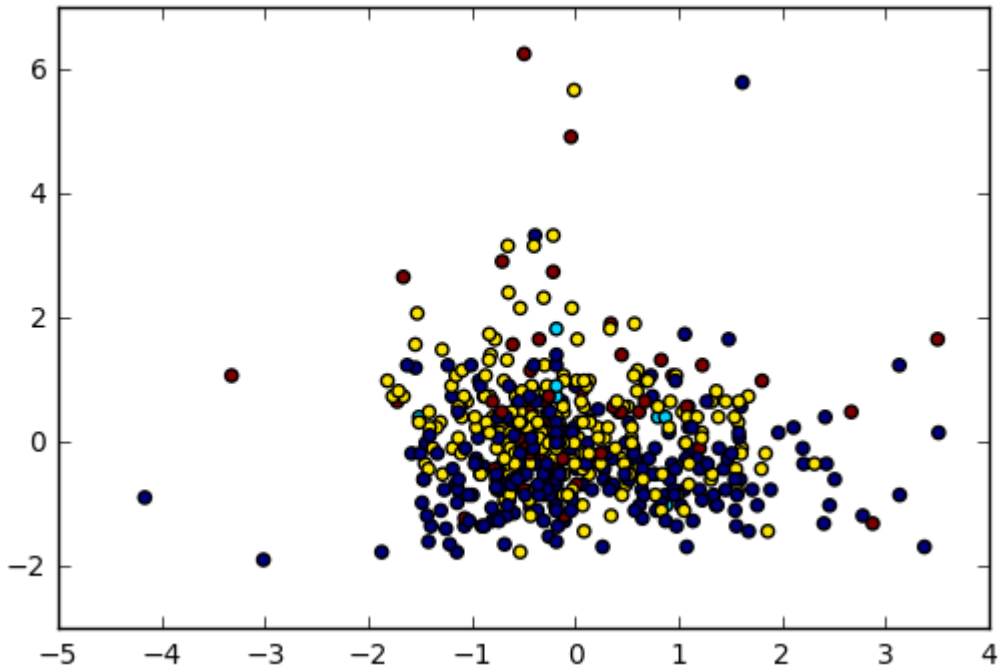

14
12
10
8
6
4
2
0
−2
−5
−4
−3
−2
−1
0
1
2
3
4

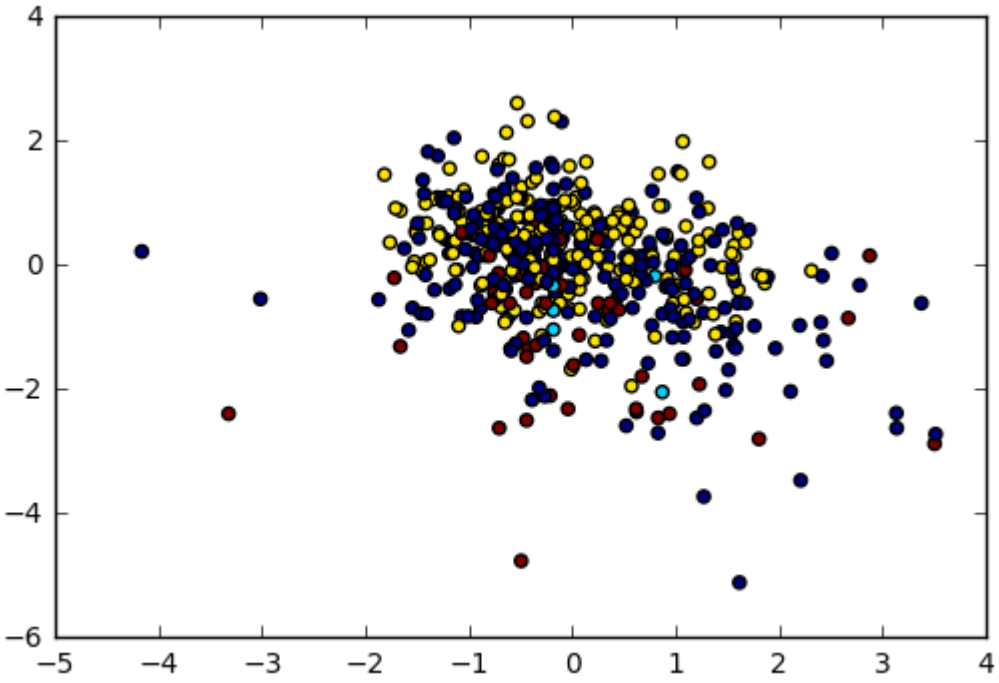

5
4
3
2
1
0
−1
−2
−3
−5
−4
−3
−2
−1
0
1
2
3
4

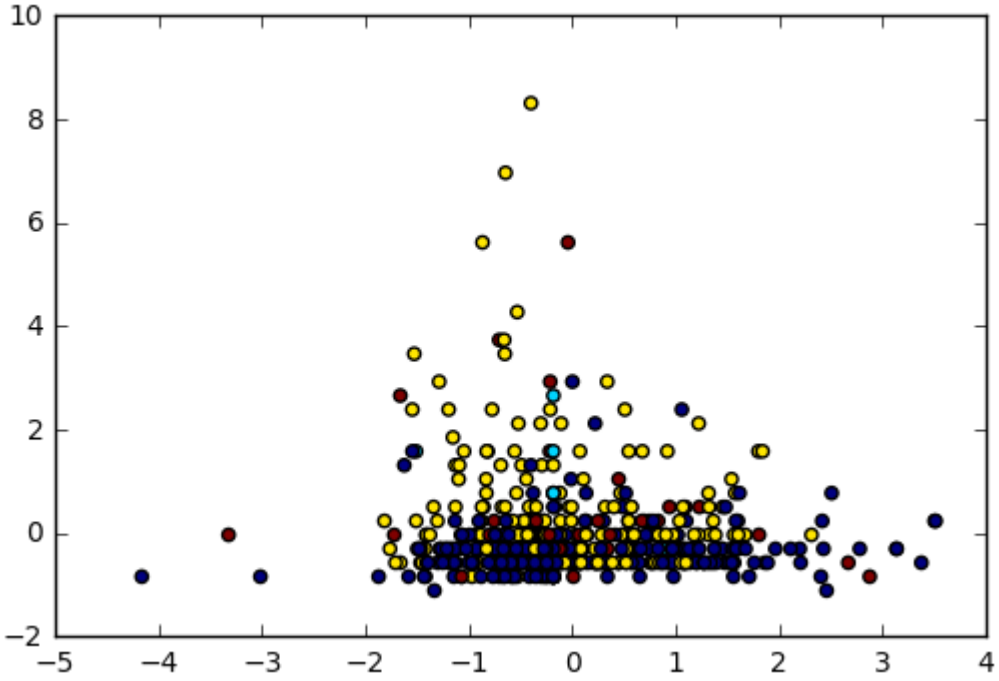

4
3
2
1
0
−1
−2
−3
−4
−5
−4
−3
−2
−1
0
1
2
3
4

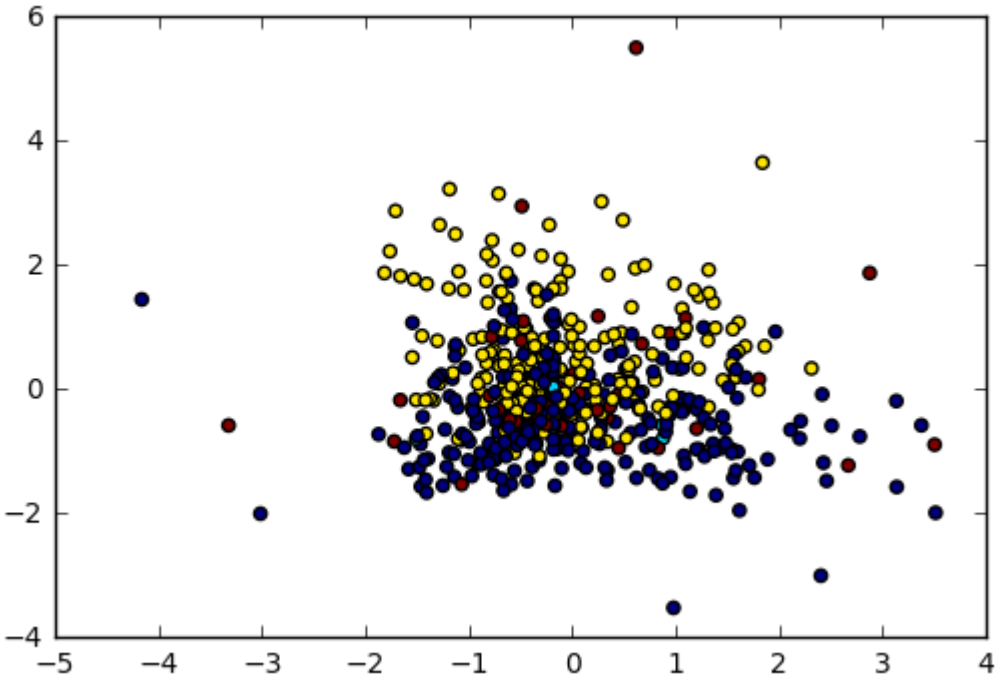

5
4
3
2
1
0
−1
−2
−3
−4
−5
−4
−3
−2
−1
0
1
2
3
4

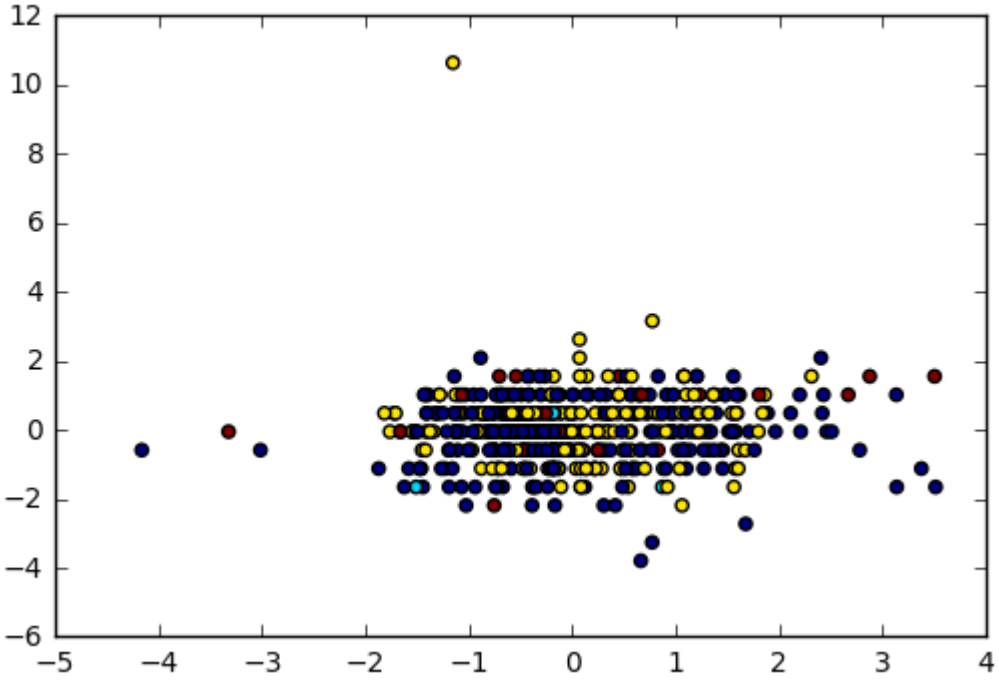

4
3
2
1
0
−1
−2
−3
−5
−4
−3
−2
−1
0
1
2
3
4

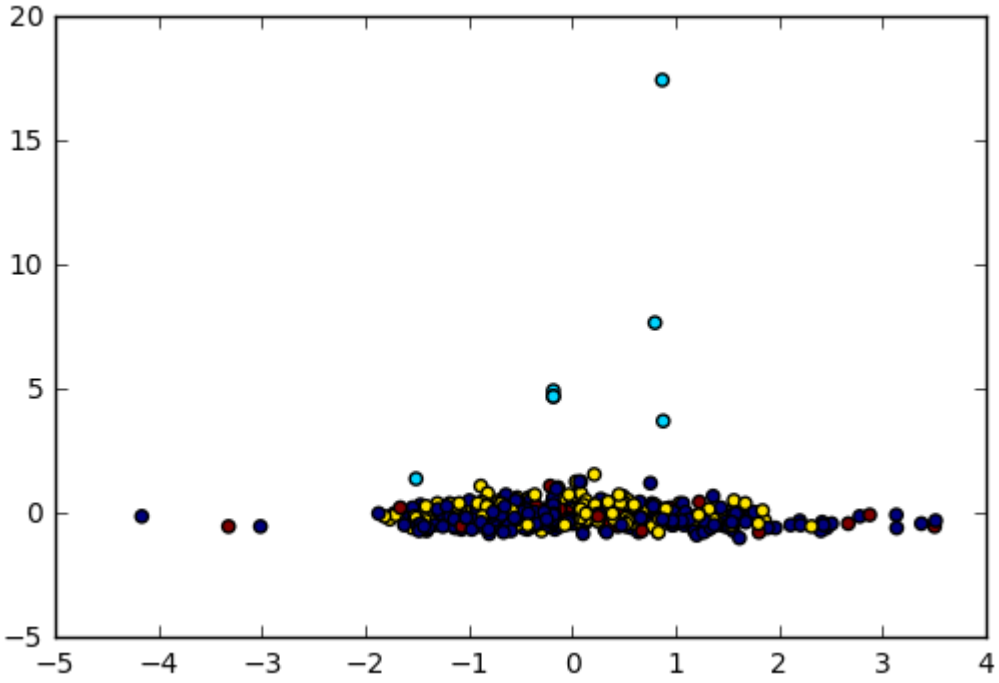

6
4
2
0
−2
−5
−4
−3
−2
−1
0
1
2
3
4

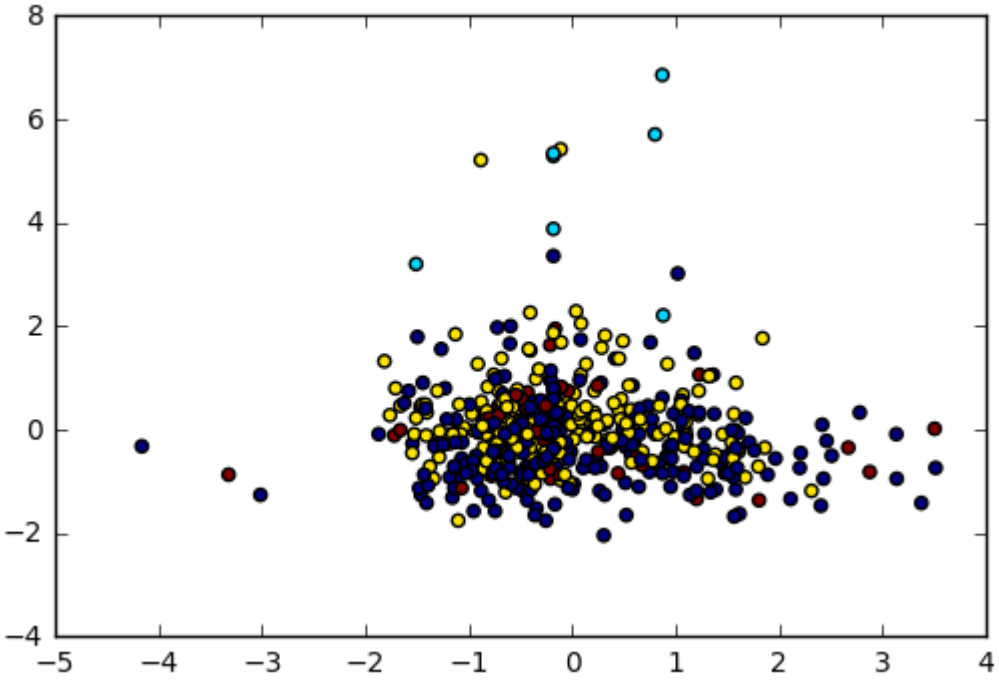

4
2
0
−2
−4
−6
−5
−4
−3
−2
−1
0
1
2
3
4

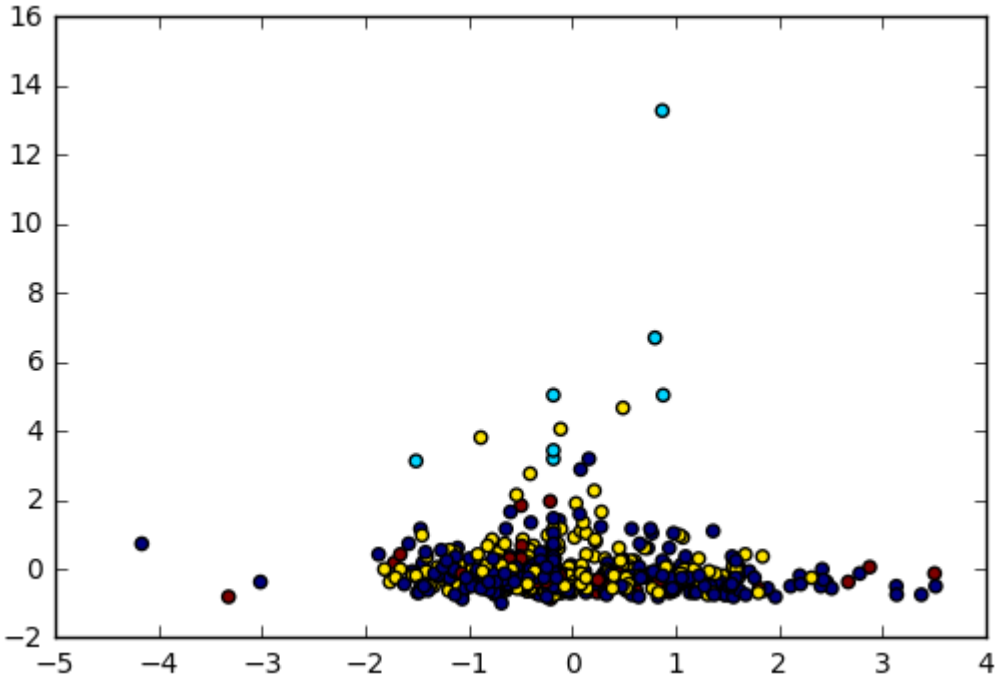

10
8
6
4
2
0
−2
−5
−4
−3
−2
−1
0
1
2
3
4

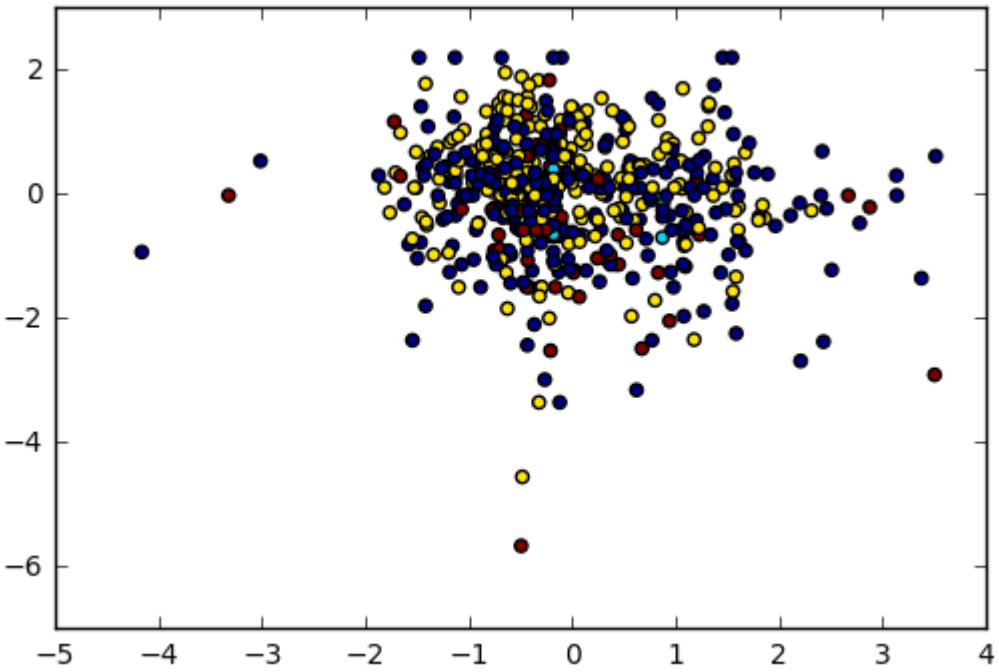

6
4
2
0
−2
−4
−5
−4
−3
−2
−1
0
1
2
3
4

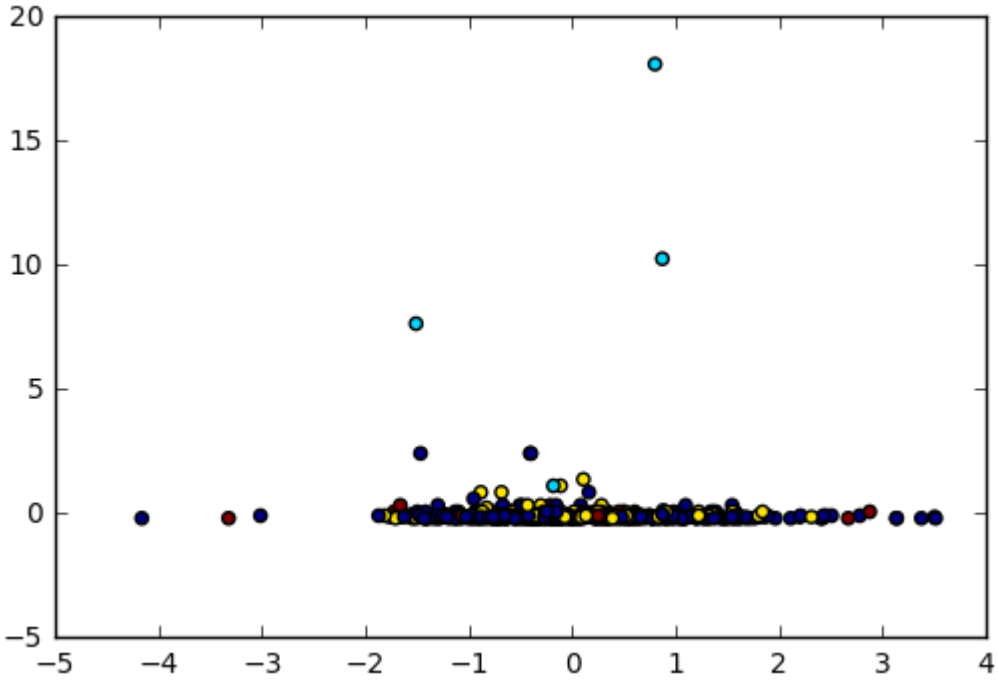

12
10
8
6
4
2
0
−2
−4
−6
−5
−4
−3
−2
−1
0
1
2
3
4

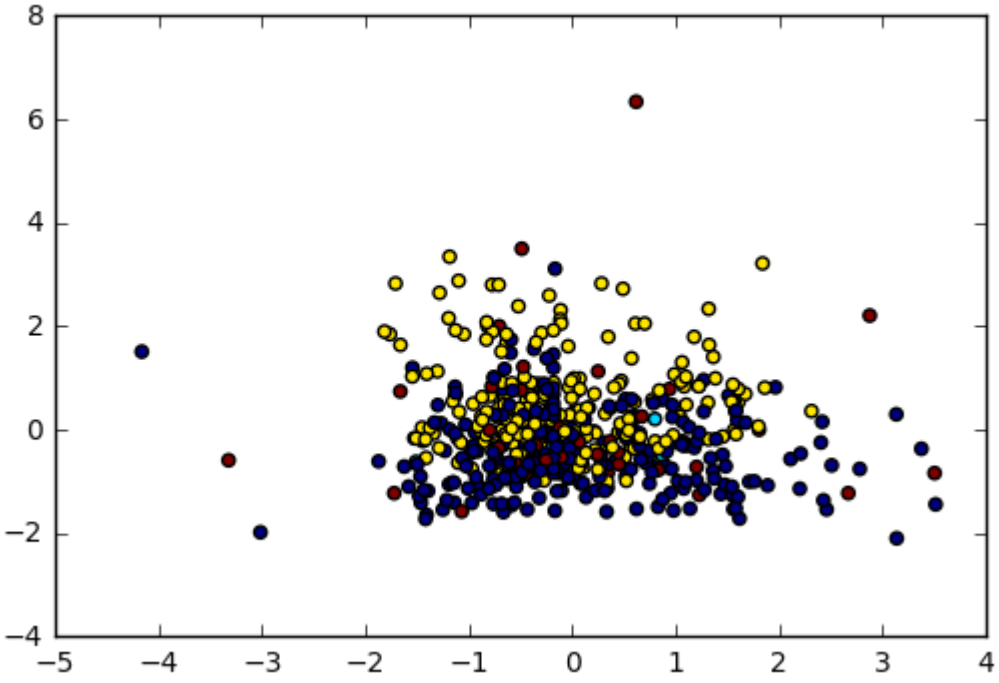

20
15
10
5
0
−5
−5
−4
−3
−2
−1
0
1
2
3
4

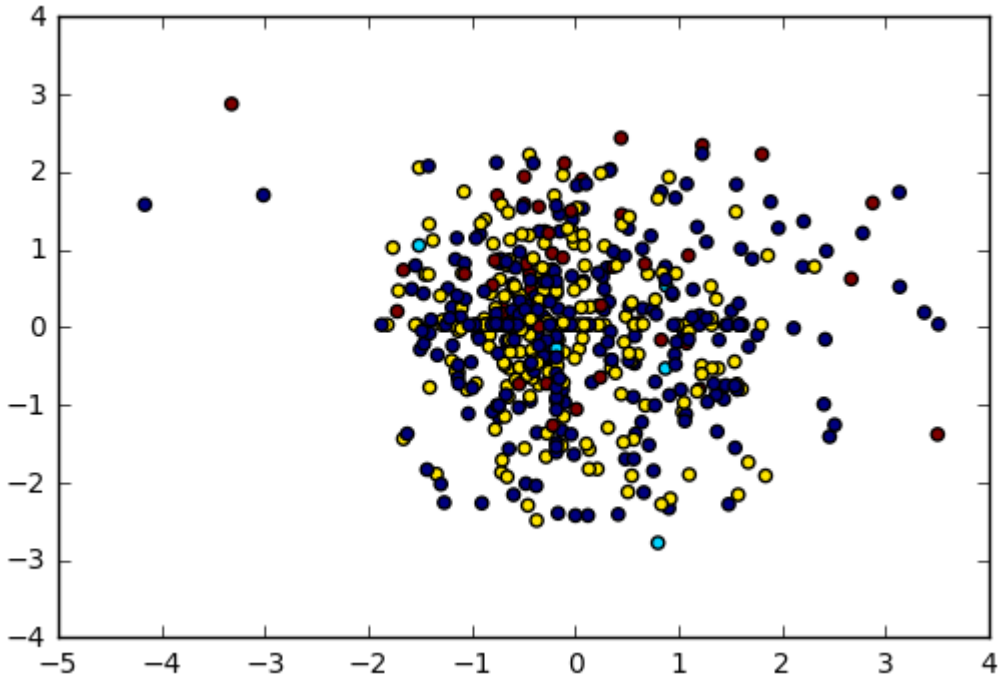

8
6
4
2
0
−2
−4
−5
−4
−3
−2
−1
0
1
2
3
4

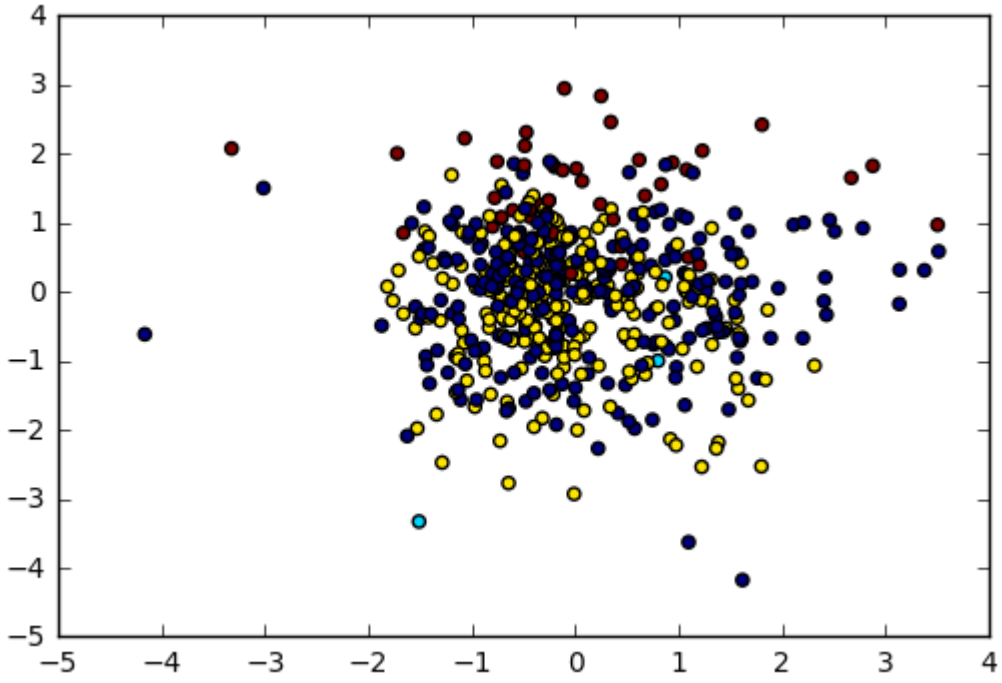

16
14
12
10
8
6
4
2
0
−2
−5
−4
−3
−2
−1
0
1
2
3
4

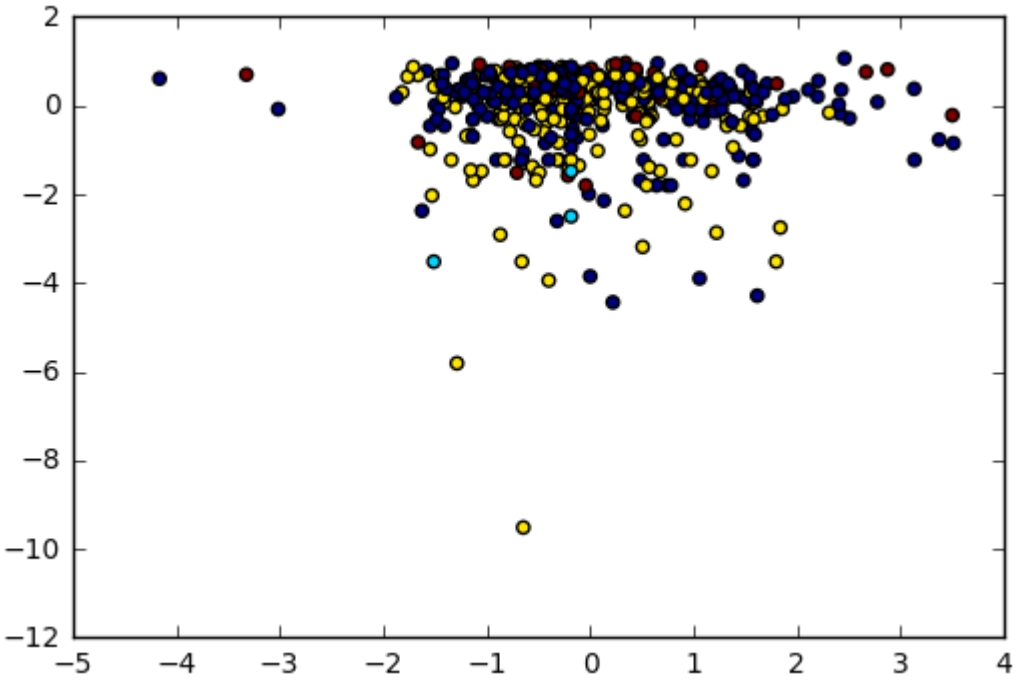

2
0
−2
−4
−6
−5
−4
−3
−2
−1
0
1
2
3
4

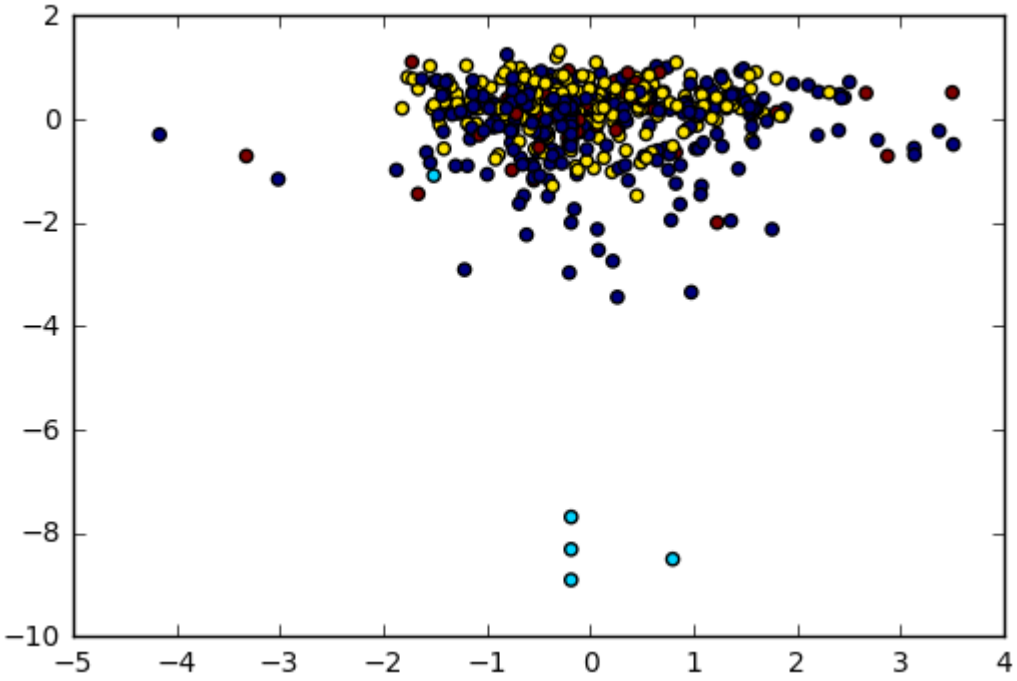

20
15
10
5
0
−5
−5
−4
−3
−2
−1
0
1
2
3
4

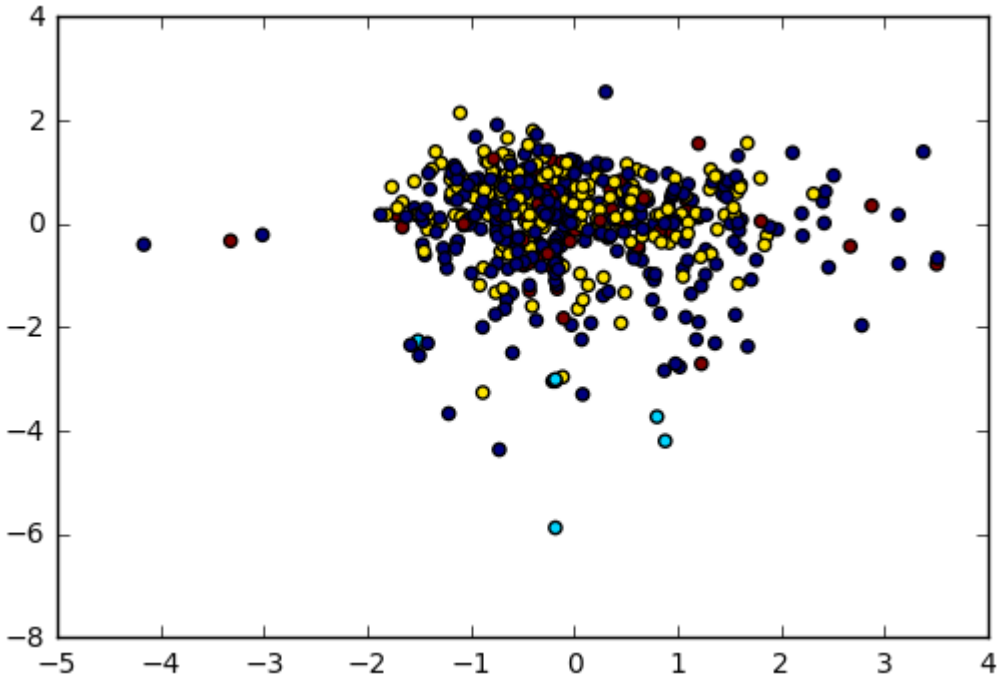

8
6
4
2
0
−2
−4
−5
−4
−3
−2
−1
0
1
2
3
4

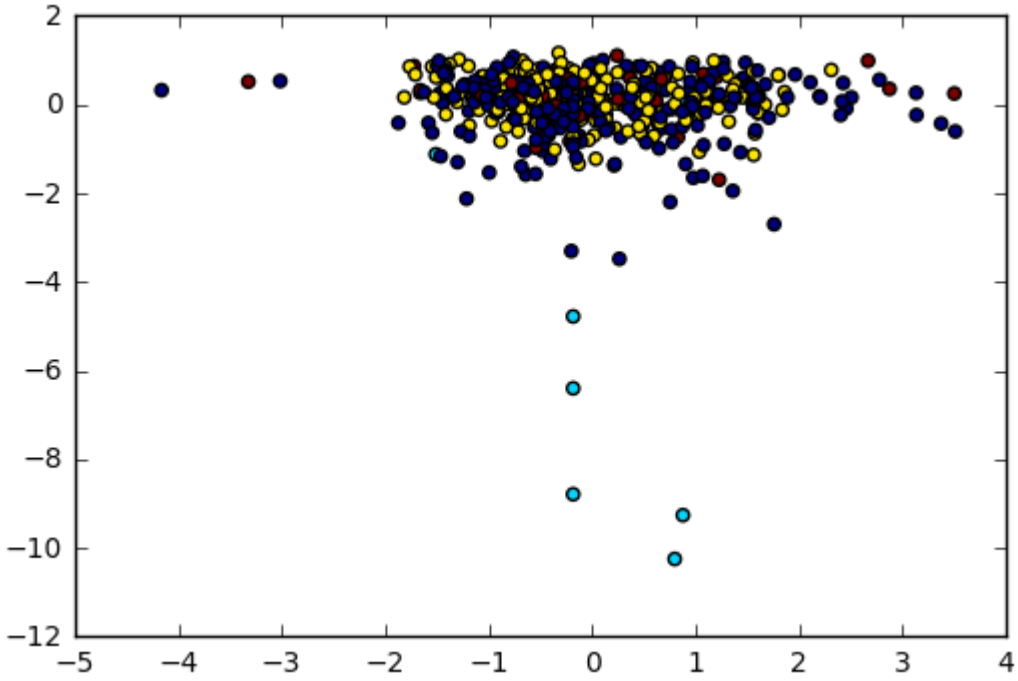

4
3
2
1
0
−1
−2
−3
−4
−5
−4
−3
−2
−1
0
1
2
3
4

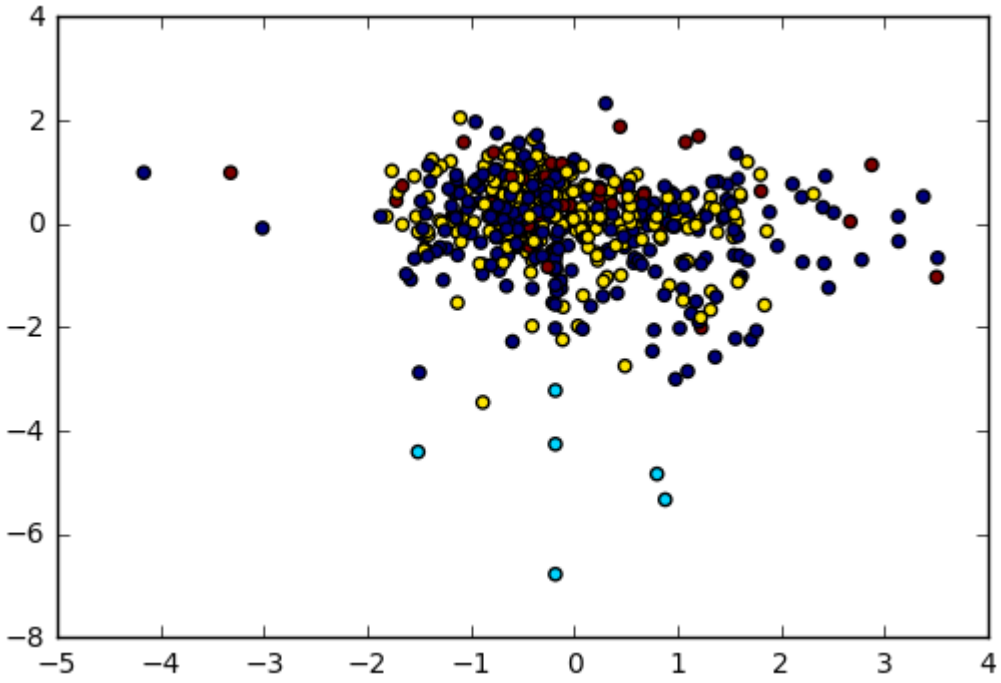

4
3
2
1
0
−1
−2
−3
−4
−5
−5
−4
−3
−2
−1
0
1
2
3
4

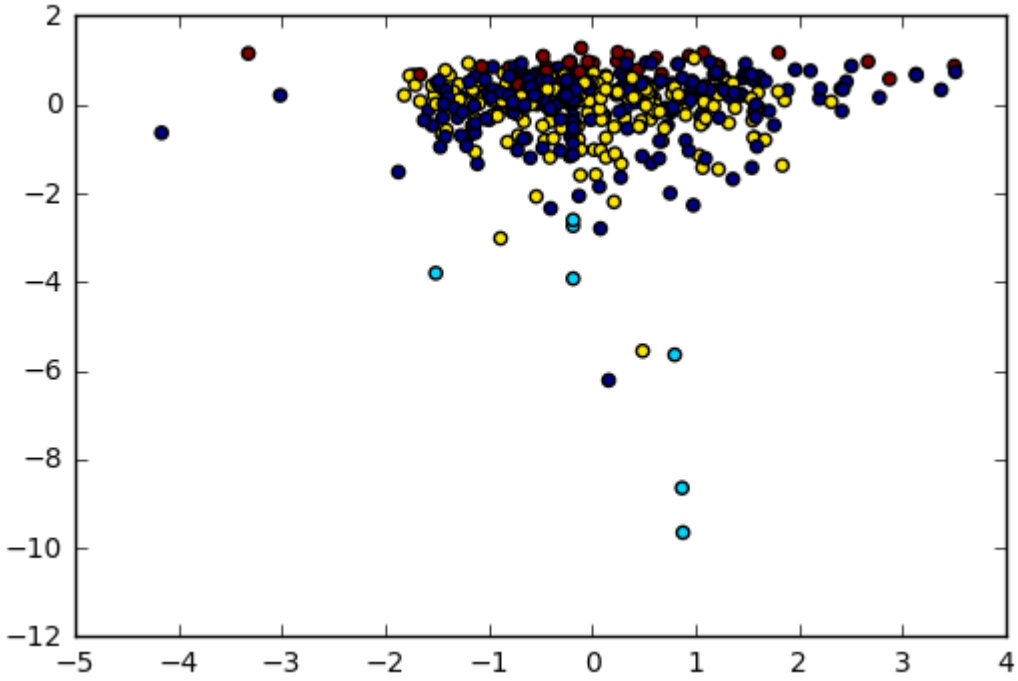

2
0
−2
−4
−6
−8
−10
−12
−5
−4
−3
−2
−1
0
1
2
3
4

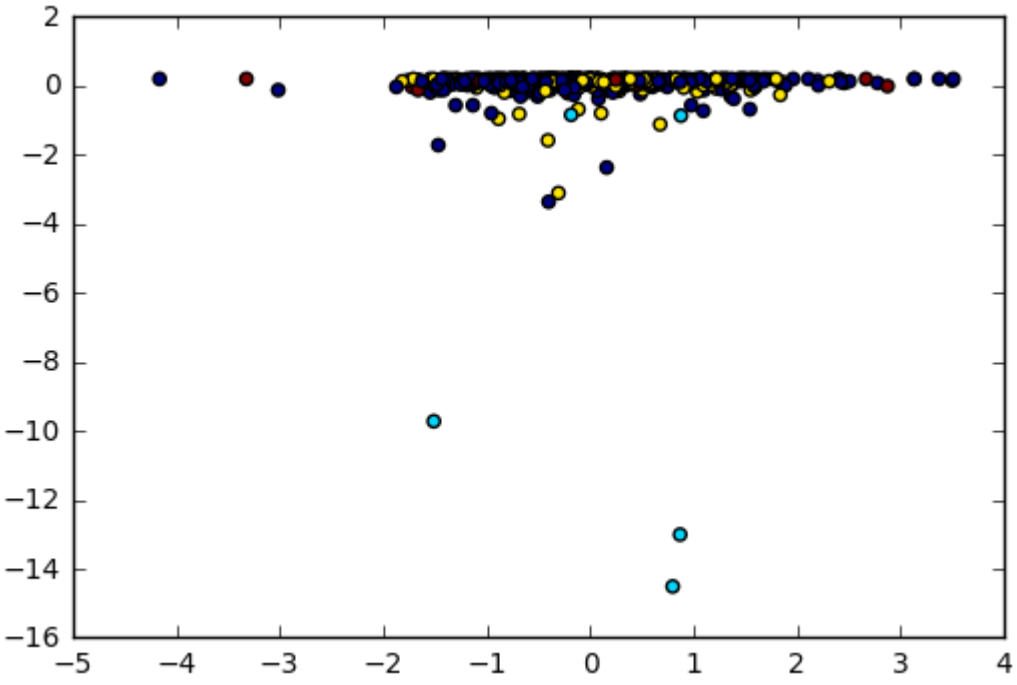

2
0
−2
−4
−6
−8
−10
−5
−4
−3
−2
−1
0
1
2
3
4


4
2
0
−2
−4
−6
−8
−5
−4
−3
−2
−1
0
1
2
3
4

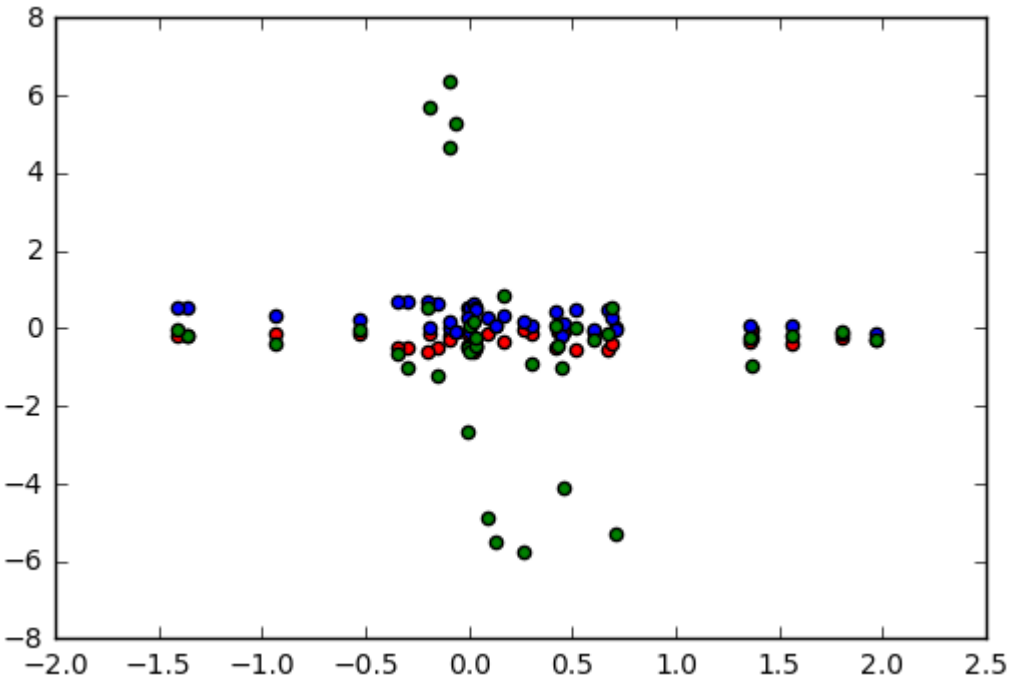

2
0
−2
−4
−6
−8
−10
−12
−5
−4
−3
−2
−1
0
1
2
3
4


4
2
0
−2
−4
−6
−8
−5
−4
−3
−2
−1
0
1
2
3
4


2
0
−2
−4
−6
−8
−10
−12
−5
−4
−3
−2
−1
0
1
2
3
4



In [309]:
```
plt.scatter(C[:,0], C[:,1], c ='r')
plt.scatter(C[:,0], C[:,2], c = 'b')
plt.scatter(C[:,0], C[:,3], c = 'g')
plt.show()
```



In [364]:
```
print('The End')
```

```
The End
```

In [ ]:
```

```