data_checks

January 16, 2023

[1]:

```
from matplotlib.pyplot import subplots
from pandas import read_csv
```

1 EEG Eye State

A worked out solution for this data set is provided.

[2]:

```
DATA_PATH = 'data/eeg_eye_state.csv'

df = read_csv(DATA_PATH)

print(df.dtypes)
print('')

columns = df.columns

print(columns)
print('')

df.head()
```

```
AF3             float64
F7              float64
F3              float64
FC5             float64
T7              float64
P7              float64
O1              float64
O2              float64
P8              float64
T8              float64
FC6             float64
F4              float64
F8              float64
AF4             float64
eyeDetection       bool
dtype: object
```

```
Index(['AF3', 'F7', 'F3', 'FC5', 'T7', 'P7', 'O1', 'O2', 'P8', 'T8', 'FC6',
       'F4', 'F8', 'AF4', 'eyeDetection'],
      dtype='object')
```

```
[2]:        AF3       F7       F3      FC5       T7       P7       O1       O2  \
     0  4329.23  4009.23  4289.23  4148.21  4350.26  4586.15  4096.92  4641.03
     1  4324.62  4004.62  4293.85  4148.72  4342.05  4586.67  4097.44  4638.97
     2  4327.69  4006.67  4295.38  4156.41  4336.92  4583.59  4096.92  4630.26
     3  4328.72  4011.79  4296.41  4155.90  4343.59  4582.56  4097.44  4630.77
     4  4326.15  4011.79  4292.31  4151.28  4347.69  4586.67  4095.90  4627.69

             P8       T8      FC6       F4       F8      AF4  eyeDetection
     0  4222.05  4238.46  4211.28  4280.51  4635.90  4393.85         False
     1  4210.77  4226.67  4207.69  4279.49  4632.82  4384.10         False
     2  4207.69  4222.05  4206.67  4282.05  4628.72  4389.23         False
     3  4217.44  4235.38  4210.77  4287.69  4632.31  4396.41         False
     4  4210.77  4244.10  4212.82  4288.21  4632.82  4398.46         False
```

2 Breast Cancer

Taken from UCI repository.

```
[3]: DATA_PATH = 'data/breast_cancer.csv'

     df = read_csv(DATA_PATH)

     print(df.dtypes)
     print('')

     columns = df.columns

     print(columns)
     print('')

     print('Rows: ', len(df))
     print('')

     df.head()
```

```
diagnosis                    int64
radius_mean                float64
texture_mean               float64
perimeter_mean             float64
area_mean                  float64
smoothness_mean            float64
compactness_mean           float64
concavity_mean             float64
```

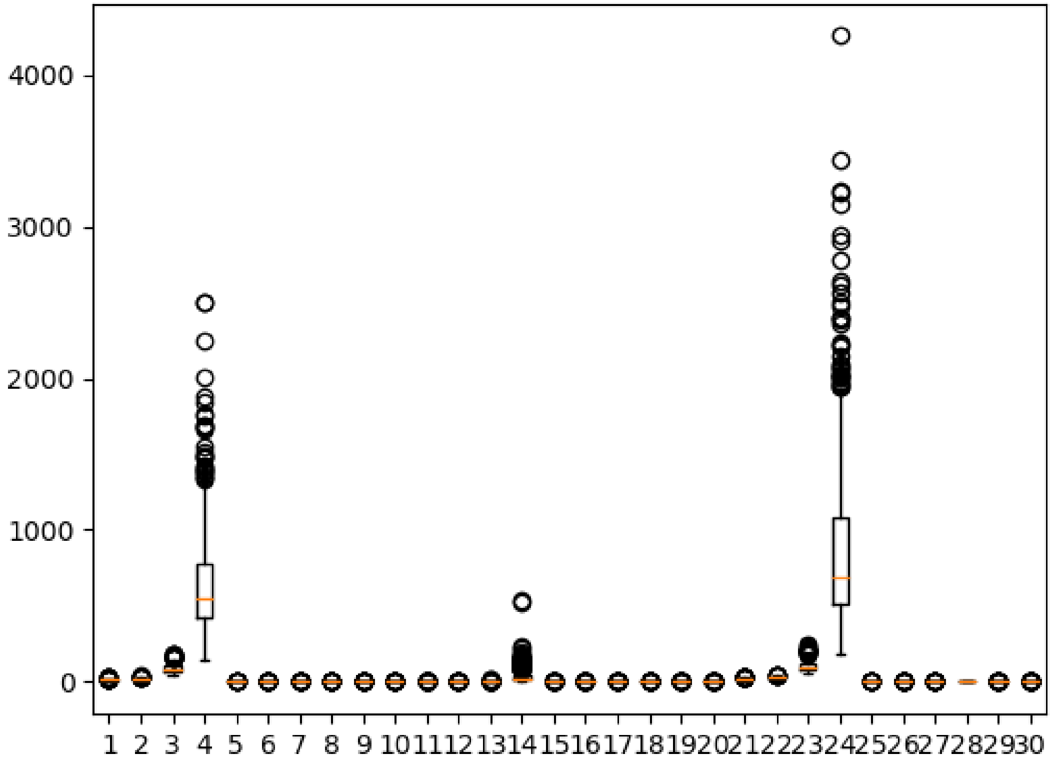
```
concave points_mean        float64
symmetry_mean              float64
fractal_dimension_mean     float64
radius_se                  float64
texture_se                 float64
perimeter_se               float64
area_se                    float64
smoothness_se              float64
compactness_se             float64
concavity_se               float64
concave points_se          float64
symmetry_se                float64
fractal_dimension_se       float64
radius_worst               float64
texture_worst              float64
perimeter_worst            float64
area_worst                 float64
smoothness_worst           float64
compactness_worst          float64
concavity_worst            float64
concave points_worst       float64
symmetry_worst             float64
fractal_dimension_worst    float64
dtype: object

Index(['diagnosis', 'radius_mean', 'texture_mean', 'perimeter_mean',
       'area_mean', 'smoothness_mean', 'compactness_mean', 'concavity_mean',
       'concave points_mean', 'symmetry_mean', 'fractal_dimension_mean',
       'radius_se', 'texture_se', 'perimeter_se', 'area_se', 'smoothness_se',
       'compactness_se', 'concavity_se', 'concave points_se', 'symmetry_se',
       'fractal_dimension_se', 'radius_worst', 'texture_worst',
       'perimeter_worst', 'area_worst', 'smoothness_worst',
       'compactness_worst', 'concavity_worst', 'concave points_worst',
       'symmetry_worst', 'fractal_dimension_worst'],
      dtype='object')

Rows:  569
```

```
[3]:    diagnosis  radius_mean  texture_mean  perimeter_mean  area_mean  \
     0          0        17.99         10.38          122.80     1001.0
     1          0        20.57         17.77          132.90     1326.0
     2          0        19.69         21.25          130.00     1203.0
     3          0        11.42         20.38           77.58      386.1
     4          0        20.29         14.34          135.10     1297.0

        smoothness_mean  compactness_mean  concavity_mean  concave points_mean  \
     0          0.11840           0.27760          0.3001              0.14710
     1          0.08474           0.07864          0.0869              0.07017
     2          0.10960           0.15990          0.1974              0.12790
     3          0.14250           0.28390          0.2414              0.10520
     4          0.10030           0.13280          0.1980              0.10430

        symmetry_mean  ...  radius_worst  texture_worst  perimeter_worst  \
     0         0.2419  ...         25.38          17.33           184.60
     1         0.1812  ...         24.99          23.41           158.80
     2         0.2069  ...         23.57          25.53           152.50
     3         0.2597  ...         14.91          26.50            98.87
     4         0.1809  ...         22.54          16.67           152.20

        area_worst  smoothness_worst  compactness_worst  concavity_worst  \
     0      2019.0            0.1622             0.6656           0.7119
     1      1956.0            0.1238             0.1866           0.2416
     2      1709.0            0.1444             0.4245           0.4504
     3       567.7            0.2098             0.8663           0.6869
     4      1575.0            0.1374             0.2050           0.4000

        concave points_worst  symmetry_worst  fractal_dimension_worst
     0                0.2654          0.4601                  0.11890
     1                0.1860          0.2750                  0.08902
     2                0.2430          0.3613                  0.08758
     3                0.2575          0.6638                  0.17300
     4                0.1625          0.2364                  0.07678

     [5 rows x 31 columns]
```

[4]:

```python
data = df.to_numpy()

data = data[:, 1:]

print('Data shape: ', data.shape)
print('')


fig, ax = subplots()

ax.boxplot(data);
```

```
Data shape:  (569, 30)
```



3 Cervical Cancer

Also from the UCI repository.

[5]:

```
DATA_PATH = 'data/cervical_cancer.csv'

df = read_csv(DATA_PATH)

print(df.dtypes)
print('')

columns = df.columns

print(columns)
print('')

print('Rows: ', len(df))
print('')

df.head()
```

```
Age                                      int64
Number of sexual partners              float64
First sexual intercourse               float64
Num of pregnancies                     float64
Smokes                                 float64
Smokes (years)                         float64
Smokes (packs/year)                    float64
Hormonal Contraceptives                float64
Hormonal Contraceptives (years)        float64
IUD                                    float64
IUD (years)                            float64
STDs                                   float64
STDs (number)                          float64
STDs:condylomatosis                    float64
STDs:cervical condylomatosis           float64
STDs:vaginal condylomatosis            float64
STDs:vulvo-perineal condylomatosis     float64
STDs:syphilis                          float64
STDs:pelvic inflammatory disease       float64
STDs:genital herpes                    float64
STDs:molluscum contagiosum             float64
STDs:AIDS                              float64
STDs:HIV                               float64
STDs:Hepatitis B                       float64
STDs:HPV                               float64
STDs: Number of diagnosis                int64
Dx:Cancer                                int64
Dx:CIN                                   int64
Dx:HPV                                   int64
Dx                                       int64
Hinselmann                               int64
Schiller                                 int64
Citology                                 int64
Biopsy                                   int64
dtype: object

Index(['Age', 'Number of sexual partners', 'First sexual intercourse',
       'Num of pregnancies', 'Smokes', 'Smokes (years)', 'Smokes (packs/year)',
       'Hormonal Contraceptives', 'Hormonal Contraceptives (years)', 'IUD',
       'IUD (years)', 'STDs', 'STDs (number)', 'STDs:condylomatosis',
       'STDs:cervical condylomatosis', 'STDs:vaginal condylomatosis',
       'STDs:vulvo-perineal condylomatosis', 'STDs:syphilis',
       'STDs:pelvic inflammatory disease', 'STDs:genital herpes',
       'STDs:molluscum contagiosum', 'STDs:AIDS', 'STDs:HIV',
       'STDs:Hepatitis B', 'STDs:HPV', 'STDs: Number of diagnosis',
       'Dx:Cancer', 'Dx:CIN', 'Dx:HPV', 'Dx', 'Hinselmann', 'Schiller',
       'Citology', 'Biopsy'],
      dtype='object')
```

```
Rows:  668
```

[5]:

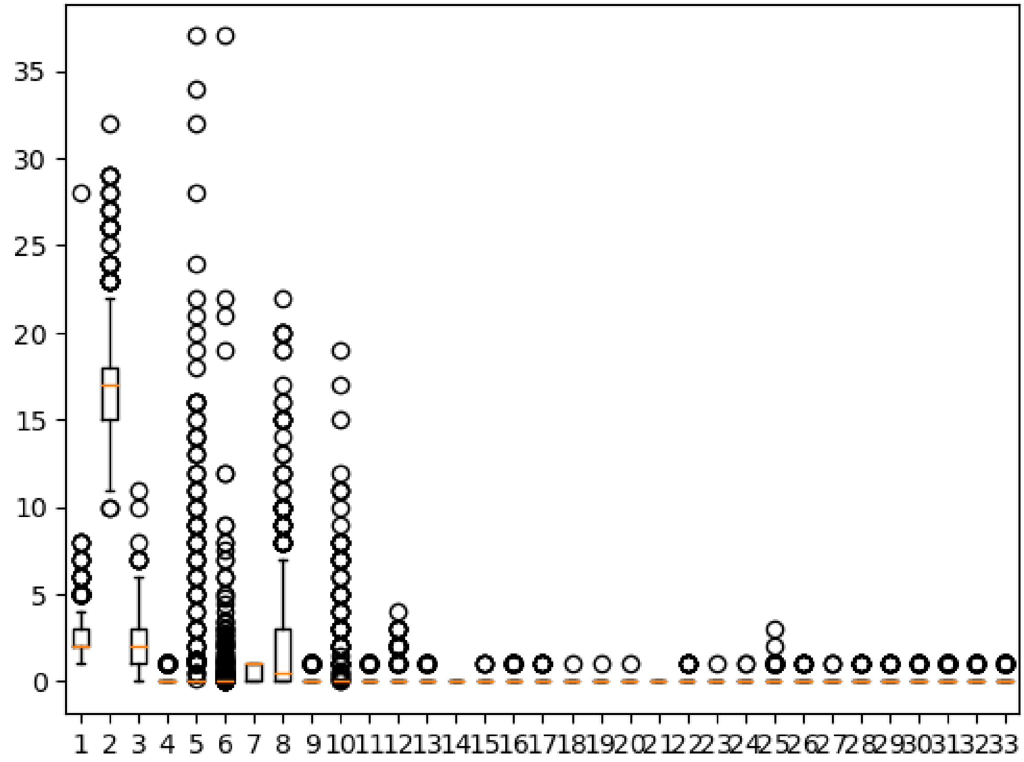
```
   Age  Number of sexual partners  First sexual intercourse  \
0   18                        4.0                      15.0   
1   15                        1.0                      14.0   
2   52                        5.0                      16.0   
3   46                        3.0                      21.0   
4   42                        3.0                      23.0   

   Num of pregnancies  Smokes  Smokes (years)  Smokes (packs/year)  \
0                 1.0     0.0             0.0                  0.0   
1                 1.0     0.0             0.0                  0.0   
2                 4.0     1.0            37.0                 37.0   
3                 4.0     0.0             0.0                  0.0   
4                 2.0     0.0             0.0                  0.0   

   Hormonal Contraceptives  Hormonal Contraceptives (years)  IUD  ...  \
0                      0.0                              0.0  0.0  ...   
1                      0.0                              0.0  0.0  ...   
2                      1.0                              3.0  0.0  ...   
3                      1.0                             15.0  0.0  ...   
4                      0.0                              0.0  0.0  ...   

   STDs:HPV  STDs: Number of diagnosis  Dx:Cancer  Dx:CIN  Dx:HPV  Dx  \
0       0.0                          0          0       0       0   0   
1       0.0                          0          0       0       0   0   
2       0.0                          0          1       0       1   0   
3       0.0                          0          0       0       0   0   
4       0.0                          0          0       0       0   0   

   Hinselmann  Schiller  Citology  Biopsy  
0           0         0         0       0  
1           0         0         0       0  
2           0         0         0       0  
3           0         0         0       0  
4           0         0         0       0  

[5 rows x 34 columns]
```

[6]:

```python
data = df.to_numpy()

data = data[:, 1:]

print('Data shape: ', data.shape)
print('')
```

```
fig, ax = subplots()

ax.boxplot(data);
```

```
Data shape:  (668, 33)
```



4 Diabetes

[7]:
```
from sklearn import datasets
from pandas import DataFrame

diabetes = datasets.load_diabetes()

data = diabetes.data

df = DataFrame(data)

print(df.dtypes)
print('')
```

```
columns = df.columns

print(columns)
print('')

print('Rows: ', len(df))
print('')

df.head()
```

```
0    float64
1    float64
2    float64
3    float64
4    float64
5    float64
6    float64
7    float64
8    float64
9    float64
dtype: object

RangeIndex(start=0, stop=10, step=1)

Rows:  442
```

[7]:

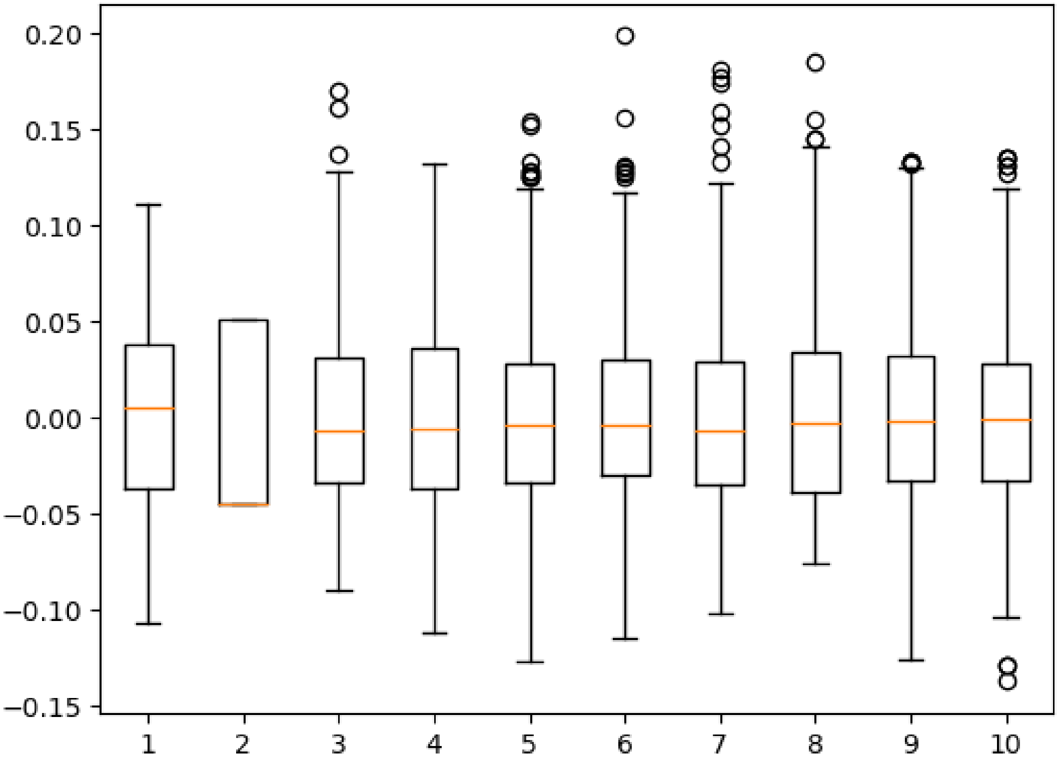
```
          0         1         2         3         4         5         6  \
0  0.038076  0.050680  0.061696  0.021872 -0.044223 -0.034821 -0.043401
1 -0.001882 -0.044642 -0.051474 -0.026328 -0.008449 -0.019163  0.074412
2  0.085299  0.050680  0.044451 -0.005671 -0.045599 -0.034194 -0.032356
3 -0.089063 -0.044642 -0.011595 -0.036656  0.012191  0.024991 -0.036038
4  0.005383 -0.044642 -0.036385  0.021872  0.003935  0.015596  0.008142

          7         8         9
0 -0.002592  0.019908 -0.017646
1 -0.039493 -0.068330 -0.092204
2 -0.002592  0.002864 -0.025930
3  0.034309  0.022692 -0.009362
4 -0.002592 -0.031991 -0.046641
```

[8]:

```
data = df.to_numpy()

data = data[:, :]

print('Data shape: ', data.shape)
print('')
```

```
fig, ax = subplots()

ax.boxplot(data);
```

```
Data shape:  (442, 10)
```



5 Echocardiogram

[9]:

```
DATA_PATH = 'data/echocardio.csv'

df = read_csv(DATA_PATH)

print(df.dtypes)
print('')

columns = df.columns

print(columns)
print('')

print('Rows: ', len(df))
print('')

df.head()
```

```
attack age                float64
pericardial effusion        int64
fractional shortening     float64
epss                      float64
lvdd                      float64
wall motion score         float64
wall motion index         float64
alive in a year              bool
survival                  float64
alive                       int64
dtype: object

Index(['attack age', 'pericardial effusion', 'fractional shortening', 'epss',
       'lvdd', 'wall motion score', 'wall motion index', 'alive in a year',
       'survival', 'alive'],
      dtype='object')

Rows:  61
```

[9]:

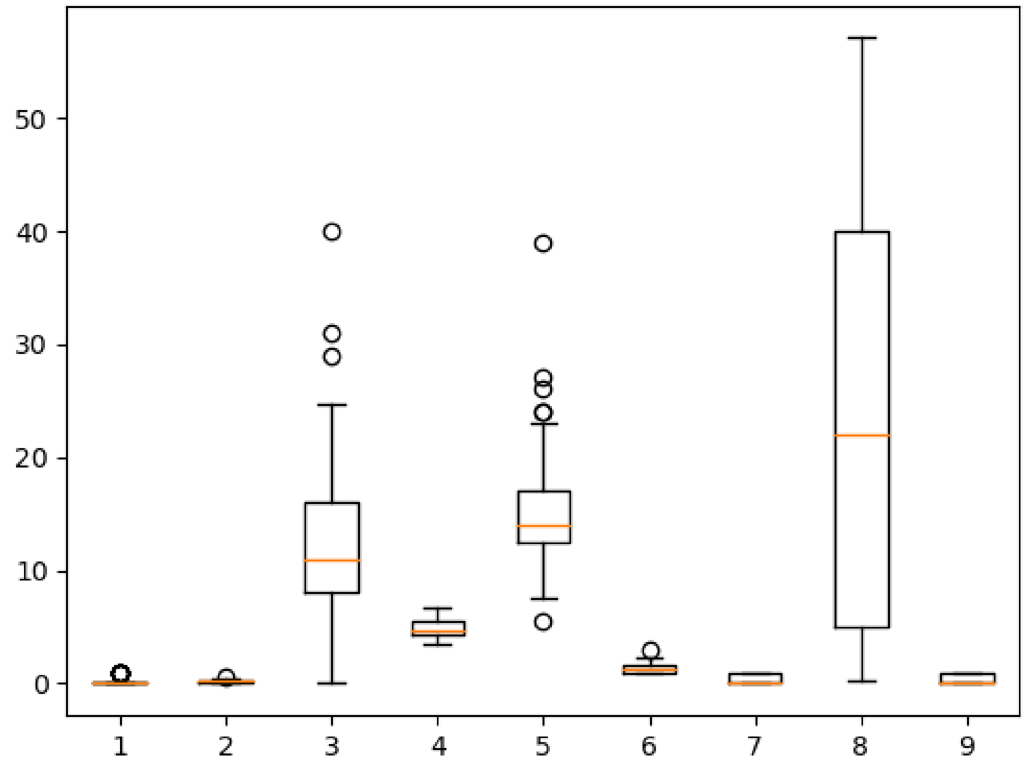
```
   attack age  pericardial effusion  fractional shortening    epss   lvdd  \
0        71.0                     0                  0.260   9.000  4.600
1        72.0                     0                  0.380   6.000  4.100
2        55.0                     0                  0.260   4.000  3.420
3        60.0                     0                  0.253  12.062  4.603
4        57.0                     0                  0.160  22.000  5.750

   wall motion score  wall motion index  alive in a year  survival  alive
0               14.0               1.00            False      11.0      0
1               14.0               1.70            False      19.0      0
2               14.0               1.00            False      16.0      0
3               16.0               1.45            False      57.0      0
4               18.0               2.25            False      19.0      1
```

[10]:

```
data = df.to_numpy()

data = data[:, 1:]

print('Data shape: ', data.shape)
print('')
```

```
fig, ax = subplots()

ax.boxplot(data);
```

```
Data shape:  (61, 9)
```



6 Gene Expression

From Gene Expression Omnibus

There is an accompanying paper called A Beginner's Guide to Analysis of RNA Sequencing Data

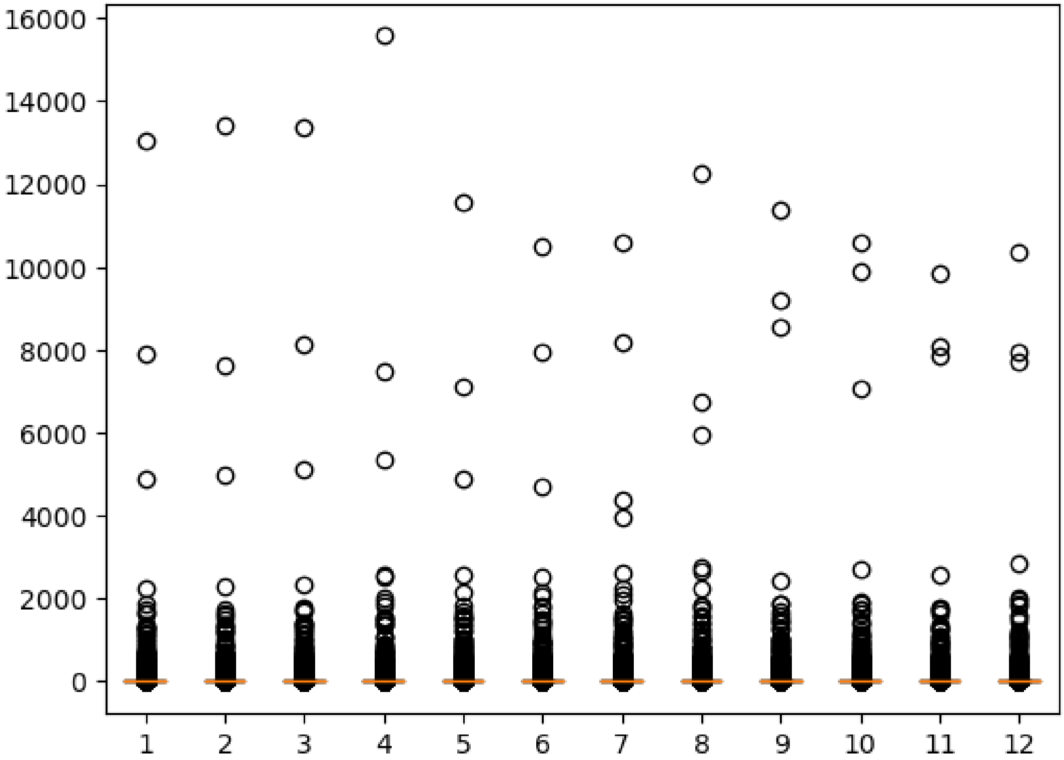
[11]:
```
data_file = "data/GSE116583_transplant.am.htseq.all.rpkm.txt"

df = read_csv(data_file, delimiter='\s+')

column_names = df.columns

print('Number of columns: ', len(column_names))
print('Number of rows: ', len(df))
print('')

df.head()
```

```
Number of columns:  13
Number of rows:  43430
```

```
[11]:               Symbol  N01_AM_Naive_01  N02_AM_Naive_02  N03_AM_Naive_03  \
      0  ENSMUSG00000000001        66.567260        72.604388        67.217287
      1  ENSMUSG00000000003         0.000000         0.000000         0.000000
      2  ENSMUSG00000000028         2.407605         2.163268         2.490079
      3  ENSMUSG00000000031         0.000000         0.000000         0.000000
      4  ENSMUSG00000000037         0.009852         0.005464         0.017368

         N04_AM_Naive_04  R01_AM_Allo_02H_01  R02_AM_Allo_02H_02  \
      0        70.263406           70.839475           64.616903
      1         0.000000            0.000000            0.000000
      2         1.593701            1.441733            1.002870
      3         0.000000            0.000000            0.000000
      4         0.014177            0.012641            0.000000

         R03_AM_Allo_02H_03  R04_AM_Allo_02H_04  R05_AM_Allo_24H_01  \
      0           69.642272           72.673890           68.579085
      1            0.000000            0.000000            0.000000
      2            1.700592            1.666637           11.861413
      3            0.000000            0.000000            0.000000
      4            0.010085            0.011964            0.028890

         R06_AM_Allo_24H_02  R07_AM_Allo_24H_03  R08_AM_Allo_24H_04
      0           69.754822           68.286563           68.031853
      1            0.000000            0.000000            0.000000
      2           13.541675           11.626514            9.789112
      3            0.000000            0.000000            0.000000
      4            0.015992            0.018193            0.018984
```

```
[12]: data = df.to_numpy()

      data = data[:, 1:]

      print('Data shape: ', data.shape)
      print('')


      fig, ax = subplots()

      ax.boxplot(data);
```

```
Data shape:  (43430, 12)
```



7 Imaging or Other Data

If you see something that looks interesting, please contact us with a public link to database or a suggestion for discussion of suitability. Do not use confidential data.

Good success!