Name: Md Toufique Hasan

Student Number: 151129267

Exercise date: 17/11/2022

Data Mining Exercise: 2

Answer 1

```
%% Read File
file = readtable('D:\TUNI\Courses\Period-2\DATA.ML.340 [Data Mining]\Weekly exercises 1\inco13p
```

file = 529×16 table

...

	NO	DIAGNOSI	UVA	US	PVR	PMU	CYM	PTR
1	2	0	0	8	1	0.0500	0	68
2	3	0	0	4	30	0.2000	0	72
3	4	0	0	4	40	0.1000	0	72
4	5	0	0	11	60	0.1500	0	71
5	6	0	0	8	5	NaN	0	NaN
6	8	0	0	NaN	30	0.9500	0	57
7	10	0	0	5	5	NaN	NaN	NaN
8	11	0	0	5	1	NaN	0	NaN
9	13	0	0	6	80	NaN	0	NaN
10	14	0	0	8	20	NaN	0	27
11	15	0	0	6	NaN	0.9500	NaN	NaN
12	16	0	0	3	1	0.4000	0	NaN
13	17	0	0	6	10	0.8000	0	102
14	19	0	0	6	NaN	NaN	NaN	NaN
15	21	0	0	6	1	0.3000	0	71
16	24	0	0	5	1	0.6000	0	27
17	25	0	0	6	1	NaN	0	82
18	26	0	0	6	2	NaN	0	48
19	27	0	0	5	100	NaN	0	86
20	29	0	0	6	100	NaN	0	62
21	33	0	0	4	20	NaN	0	80
22	34	0	0	NaN	1	NaN	0	80

	NO	DIAGNOSI	UVA	US	PVR	PMU	CYM	PTR
23	37	0	0	14	160	0.9500	0	62
24	38	0	0	6	20	0.8000	0	62
25	46	0	0	6	1	NaN	0	64
26	48	0	0	4	20	0.1000	0	NaN
27	49	0	0	5	20	0.8000	0	81
28	50	0	0	4	1	NaN	0	52
29	51	0	0	11	10	0.6000	0	21
30	53	0	0	4	10	NaN	0	98
31	55	0	0	11	1	0.9500	0	21
32	57	0	0	7	NaN	NaN	NaN	NaN
33	59	0	0	7	1	NaN	0	NaN
34	60	0	0	NaN	NaN	NaN	NaN	NaN
35	62	0	0	5	20	0.3000	0	63
36	66	0	0	5	10	0.5000	0	79
37	72	0	0	NaN	NaN	NaN	NaN	NaN
38	73	0	0	11	NaN	0.5000	NaN	NaN
39	77	0	0	5	NaN	NaN	NaN	NaN
40	78	0	0	3	NaN	0.1000	NaN	NaN
41	79	0	0	3	NaN	NaN	NaN	NaN
42	80	0	0	3	NaN	0.7000	NaN	NaN
43	81	0	0	7	1	0.2000	0	50
44	84	0	0	10	75	NaN	0	94
45	87	0	0	9	20	0.2000	0	89
46	88	0	0	7	5	NaN	0	56
47	91	0	0	6	5	NaN	0	NaN
48	92	0	0	7	60	NaN	0	57
49	95	0	0	18	40	NaN	0	82
50	96	0	0	7	1	0.1000	0	82
51	97	0	0	7	1	NaN	0	69
52	98	0	0	7	150	NaN	0	93
53	99	0	0	6	1	NaN	0	78
54	100	0	0	6	10	NaN	0	60
55	101	0	0	6	5	0.1000	0	78

	NO	DIAGNOSI	UVA	US	PVR	PMU	CYM	PTR
56	102	0	0	8	1	NaN	0	70
57	103	0	0	2	1	NaN	0	119
58	104	0	0	5	100	0.8000	0	75
59	106	0	0	2	2	0.4000	0	52
60	108	0	0	NaN	60	NaN	0	91
61	109	0	0	4	10	0.4000	0	73
62	111	0	0	5	1	0.3000	0	57
63	113	0	0	5	10	NaN	0	64
64	114	0	0	8	50	0.2000	0	61
65	116	0	0	10	10	NaN	0	95
66	118	0	0	4	20	0.1000	0	88
67	123	0	0	6	50	NaN	0	83
68	124	0	0	9	1	0.3000	0	90
69	130	0	0	6	1	0.9500	0	84
70	132	0	0	3	30	NaN	0	NaN
71	133	0	0	3	1	NaN	0	72
72	135	0	0	NaN	20	NaN	0	NaN
73	139	0	0	NaN	NaN	NaN	NaN	NaN
74	140	0	0	3	1	NaN	0	40
75	142	0	0	NaN	NaN	NaN	NaN	NaN
76	143	0	0	7	40	NaN	0	85
77	144	0	0	3	NaN	0.3000	NaN	NaN
78	145	0	0	NaN	200	0.9500	0	80
79	146	0	0	9	1	NaN	0	95
80	149	0	0	6	45	NaN	0	60
81	151	0	0	3	1	NaN	0	76
82	153	0	0	5	10	NaN	0	75
83	154	0	0	NaN	NaN	NaN	NaN	NaN
84	155	0	0	4	30	0.2000	0	72
85	156	0	0	5	50	0.6000	0	80
86	157	0	0	3	1	0.2000	0	63
87	159	0	0	6	30	0.5000	0	109
88	160	0	0	NaN	NaN	0.2000	NaN	NaN

	NO	DIAGNOSI	UVA	US	PVR	PMU	CYM	PTR
89	161	0	0	9	1	0.7000	0	86
90	162	0	0	3	10	0.2000	0	67
91	163	0	0	10	10	NaN	0	68
92	164	0	0	6	1	0.2000	0	71
93	165	0	0	9	10	NaN	0	56
94	171	0	0	7	1	0.1000	0	87
95	173	0	0	NaN	NaN	NaN	NaN	NaN
96	175	0	0	5	5	0.4000	0	79
97	176	0	0	6	10	NaN	0	55
98	177	0	0	9	8	NaN	0	65
99	178	0	0	1	5	NaN	0	68
100	179	0	0	4	5	0.3000	0	65

⋮

This **inco13par.txt** file has 529 Rows and 16 Columns.

NO is an Index Variable and no statistic can be calculated.

DIAGNOSI is Nominal variable and MODE can be calculated.

UVA is Binary variable and MODE can be calculated.

US is Nominal Variable and MODE can be calculated.

PVR is Nominal Variable and MODE can be calculated.

PMU is Nominal Variable and MODE can be calculated.

CYM is Nominal Variable and MODE can be calculated.

PTR is Nominal Variable and MODE can be calculated.

MUC is Nominal Variable and MODE can be calculated.

SS is Binary Variable and MODE can be calculated.

UVJ is Nominal Variable and MODE can be calculated.

SSY is Nominal Variable and MODE can be calculated.

CLU is Binary Variable and MODE can be calculated.

DV is Binary Variable and MODE can be calculated.

USY is Binary Variable and MODE can be calculated.

Age is Continuous Variable and MODE/MEDIAN can be calculated.

```
%% Different Diagnoses
unique(file.DIAGNOSI)
```

```
ans = 5×1
     0
     1
     2
     3
     4
```

Here, 0 means no diseases and 1-4 means different types of diseases.

```
%% Mean Age
meanAge = mean(file.AGE, 'omitnan')
```

```
meanAge = 52.3276
```

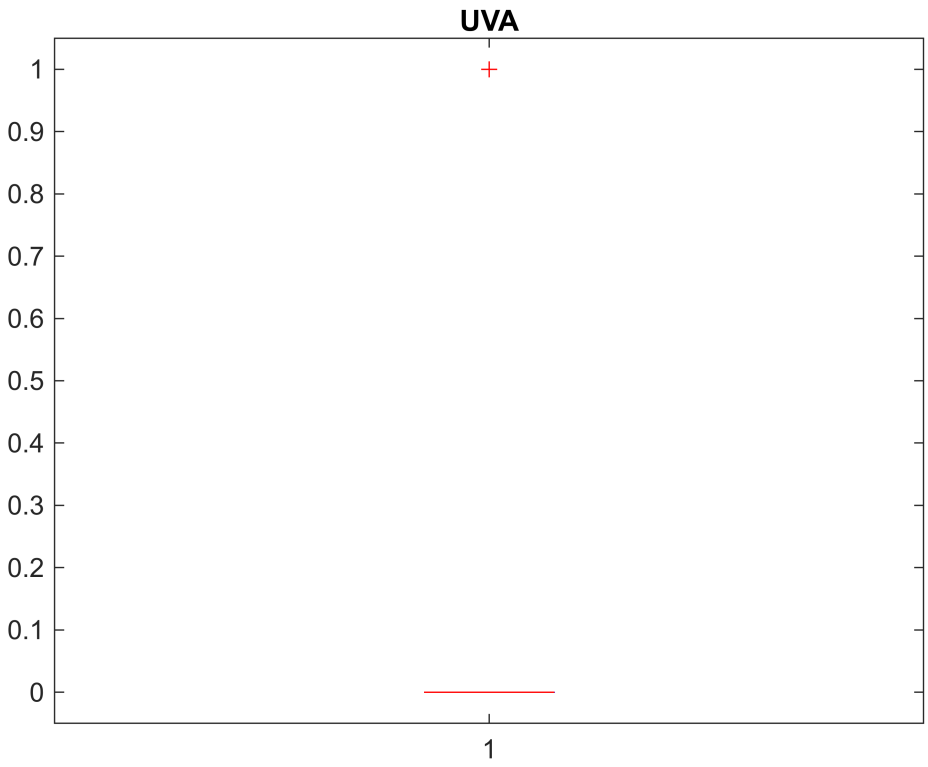
Answer 2

```
%% Replace all missing variable values (NaN) using respective mean

Diagnosis0 = file(file.DIAGNOSI == 0, :)
Diagnosis1 = file(file.DIAGNOSI == 1, :)
Diagnosis2 = file(file.DIAGNOSI == 2, :)
Diagnosis3 = file(file.DIAGNOSI == 3, :)
Diagnosis4 = file(file.DIAGNOSI == 4, :)

row_column = size(file)
for n = 2:row_column(2)
Diagnosis0{:,n}(isnan(Diagnosis0{:,n}))= mean(Diagnosis0{:,n}, 'omitnan')
Diagnosis1{:,n}(isnan(Diagnosis1{:,n}))= mean(Diagnosis1{:,n}, 'omitnan')
Diagnosis2{:,n}(isnan(Diagnosis2{:,n}))= mean(Diagnosis2{:,n}, 'omitnan')
Diagnosis3{:,n}(isnan(Diagnosis3{:,n}))= mean(Diagnosis3{:,n}, 'omitnan')
Diagnosis4{:,n}(isnan(Diagnosis4{:,n}))= mean(Diagnosis4{:,n}, 'omitnan')
end
```

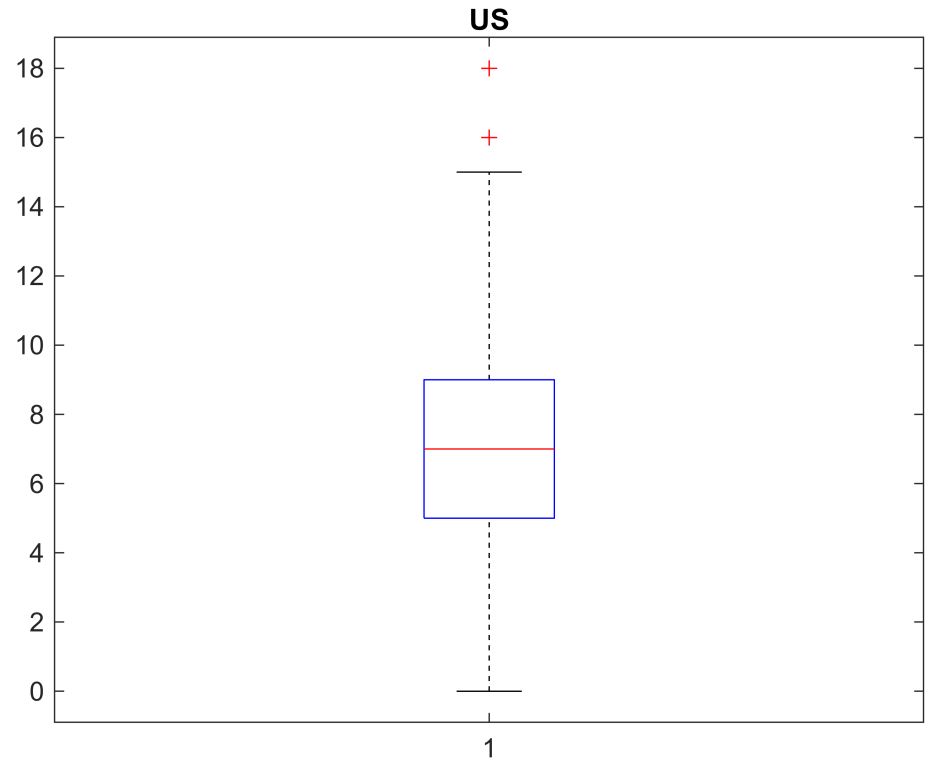
The output of Answer 2 is long, for this reason I do not show these here.

Answer 3

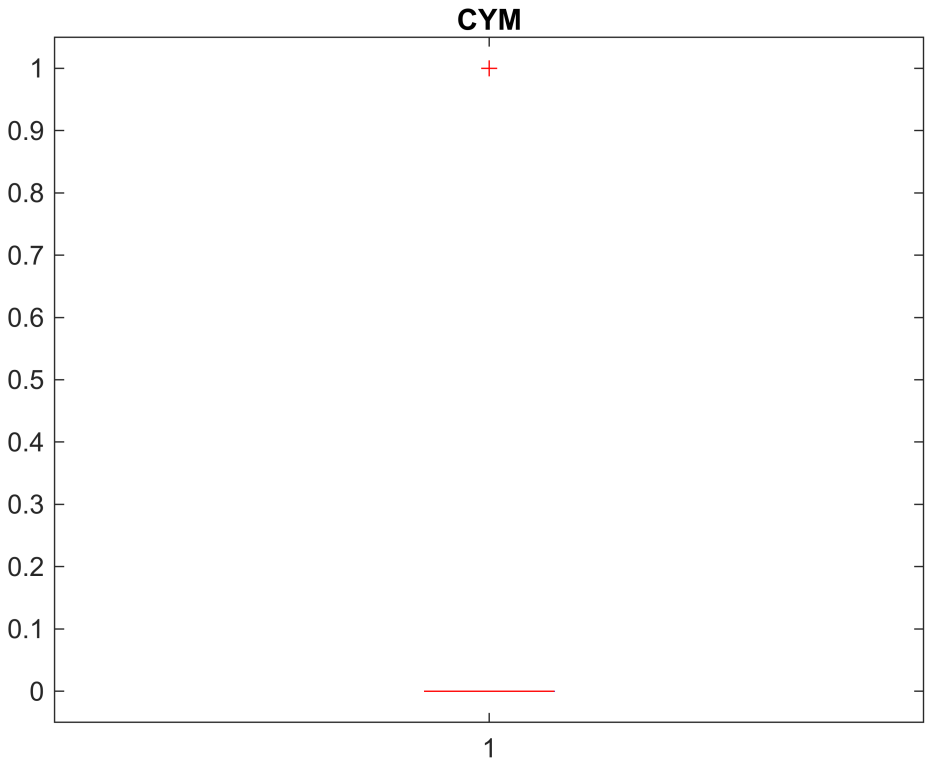
```
boxplot(file.UVA)
title('UVA')
```



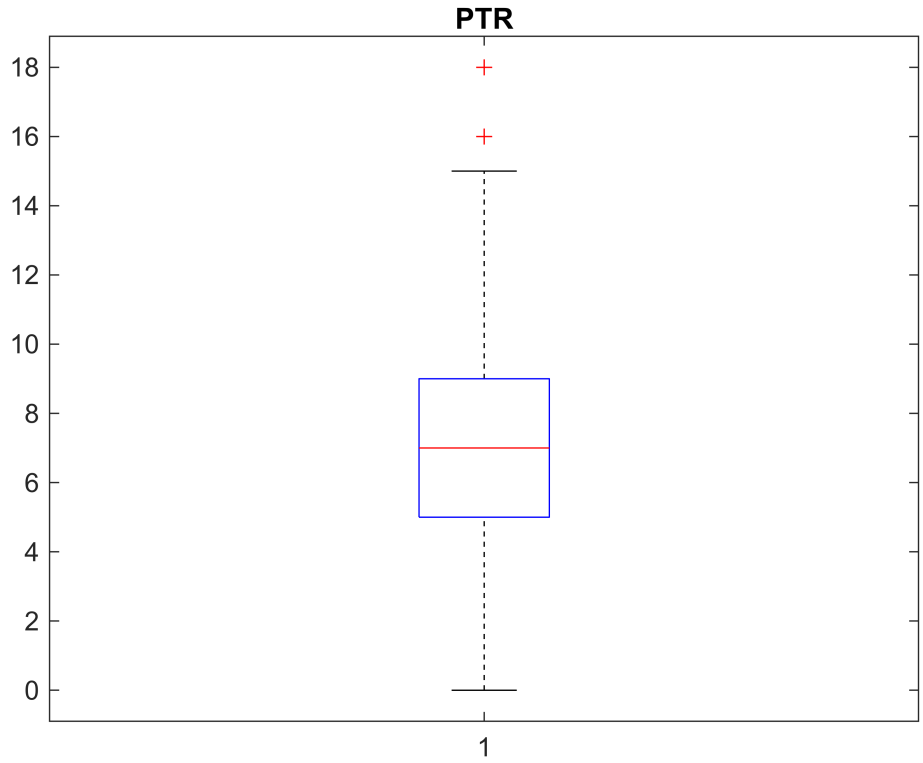

```
boxplot(file.US)
title('US')
```



```
boxplot(file.UVA)
title('CYM')
```




```
boxplot(file.US)
title('PTR')
```



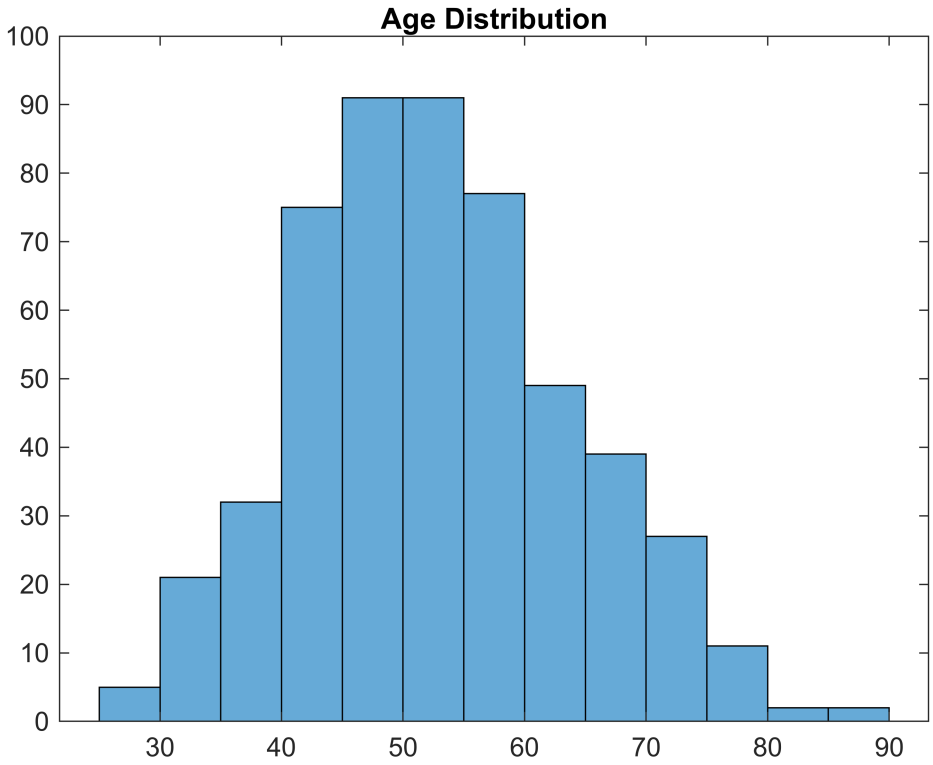
UVA is binary variable.

US is Nominal variable.

CYM is binary variable.

PTR is Nominal variable.

```
histogram(file.AGE)
title('Age Distribution')
```



From the above Histogram of Age Distribution, the values are in uniform form and these are discrete values. So I can say that this is similar to the Normal Distribution of the data.

Answer 4

```
%% Euclidean distances
row_2 = inco13par(2,:)
```

row_2 = 1×16 table

...

	NO	DIAGNOSI	UVA	US	PVR	PMU	CYM	PTR
1	3	0	0	4	30	0.2000	0	72

```
row_269 = inco13par(269,:)
```

row_269 = 1×16 table

...

	NO	DIAGNOSI	UVA	US	PVR	PMU	CYM	PTR
1	458	0	0	NaN	1	0.4000	0	64

```
row_393 = inco13par(393,:)
```

row_393 = 1×16 table

...

	NO	DIAGNOSI	UVA	US	PVR	PMU	CYM	PTR
1	469	1	0	NaN	20	0.9500	0	68

```
%% Calculating Distance
distance_2_269 = sqrt(sum(table2array(row_2)-table2array(row_269)).^2)
```

```
distance_2_269 = NaN
```

```
distance_2_393 = sqrt(sum(table2array(row_2)-table2array(row_393)).^2)
```

```
distance_2_393 = NaN
```

```
distance_269_393 = sqrt(sum(table2array(row_269)-table2array(row_393)).^2)
```

```
distance_269_393 = NaN
```

269 and 393 are closest to each other.

I am using Euclidean Distance Measure with Nominal variables. Fot this reason the output is not coming. It is possible if the variable is Binary.

There are plenty of others approaches. Such as: Hamming Distance, Manhattan Distance, Chebyshev Distance, Minkowski Distance etc.

Answer 5

```
%% Load File
fileCSV = TetuanCitypowerconsumptionNumeric
```

```
fileCSV = 52416×9
10^4 ×
    0.0001    0.0007    0.0074    0.0000    0.0000    0.0000    3.4056    1.6129 ···
    0.0001    0.0006    0.0075    0.0000    0.0000    0.0000    2.9815    1.9375
    0.0001    0.0006    0.0075    0.0000    0.0000    0.0000    2.9128    1.9007
    0.0001    0.0006    0.0075    0.0000    0.0000    0.0000    2.8229    1.8361
    0.0001    0.0006    0.0076    0.0000    0.0000    0.0000    2.7336    1.7872
    0.0001    0.0006    0.0077    0.0000    0.0000    0.0000    2.6625    1.7416
    0.0001    0.0006    0.0078    0.0000    0.0000    0.0000    2.5999    1.6993
    0.0001    0.0005    0.0078    0.0000    0.0000    0.0000    2.5446    1.6661
    0.0001    0.0006    0.0078    0.0000    0.0000    0.0000    2.4778    1.6227
    0.0001    0.0005    0.0077    0.0000    0.0000    0.0000    2.4279    1.5939
     ⋮
```

```
%% Remove the first row (variable names)
fileCSV(1,:) = []
```

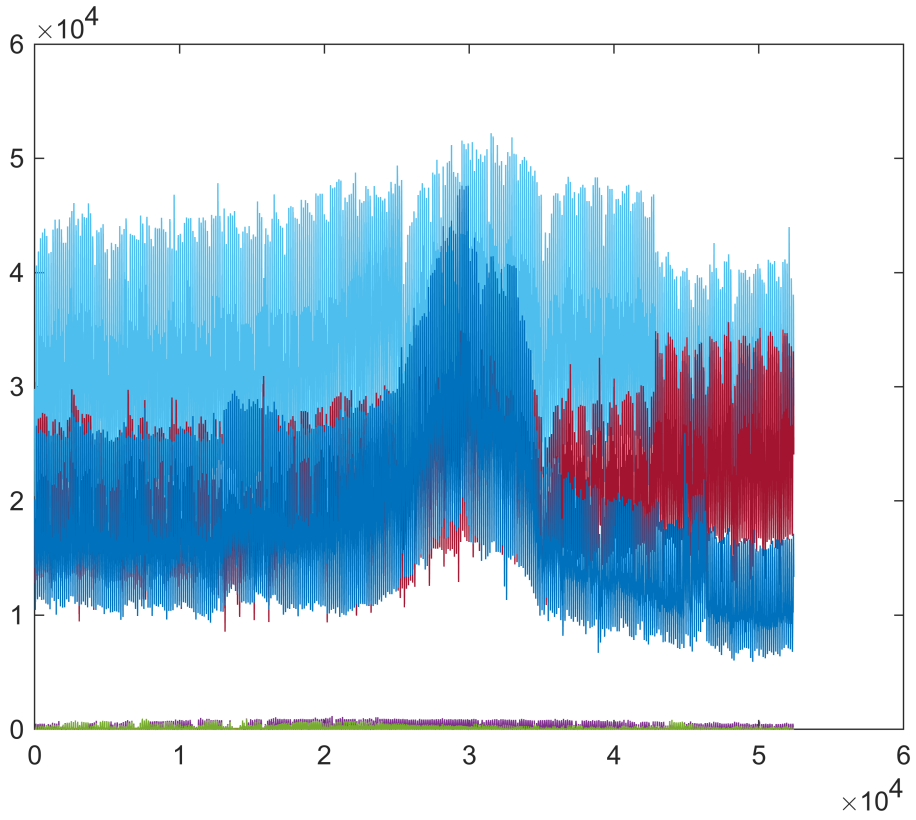
```
fileCSV = 52415×9
10^4 ×
    0.0001    0.0006    0.0075    0.0000    0.0000    0.0000    2.9815    1.9375 ···
    0.0001    0.0006    0.0075    0.0000    0.0000    0.0000    2.9128    1.9007
    0.0001    0.0006    0.0075    0.0000    0.0000    0.0000    2.8229    1.8361
    0.0001    0.0006    0.0076    0.0000    0.0000    0.0000    2.7336    1.7872
    0.0001    0.0006    0.0077    0.0000    0.0000    0.0000    2.6625    1.7416
    0.0001    0.0006    0.0078    0.0000    0.0000    0.0000    2.5999    1.6993
    0.0001    0.0005    0.0078    0.0000    0.0000    0.0000    2.5446    1.6661
```

```
    0.0001    0.0006    0.0078    0.0000    0.0000    0.0000    2.4778    1.6227
    0.0001    0.0005    0.0077    0.0000    0.0000    0.0000    2.4279    1.5939
    0.0001    0.0006    0.0077    0.0000    0.0000    0.0000    2.3897    1.5436
      ⋮
```

```
%% Remove the first column (date info)
fileCSV(:,1) = []
```

```
fileCSV = 52415×8
10^4 ×
    0.0006    0.0075    0.0000    0.0000    0.0000    2.9815    1.9375    2.0131
    0.0006    0.0075    0.0000    0.0000    0.0000    2.9128    1.9007    1.9668
    0.0006    0.0075    0.0000    0.0000    0.0000    2.8229    1.8361    1.8899
    0.0006    0.0076    0.0000    0.0000    0.0000    2.7336    1.7872    1.8442
    0.0006    0.0077    0.0000    0.0000    0.0000    2.6625    1.7416    1.8130
    0.0006    0.0078    0.0000    0.0000    0.0000    2.5999    1.6993    1.7945
    0.0005    0.0078    0.0000    0.0000    0.0000    2.5446    1.6661    1.7459
    0.0006    0.0078    0.0000    0.0000    0.0000    2.4778    1.6227    1.7026
    0.0005    0.0077    0.0000    0.0000    0.0000    2.4279    1.5939    1.6794
    0.0006    0.0077    0.0000    0.0000    0.0000    2.3897    1.5436    1.6638
      ⋮
```

```
%% Visually (plot) all variables
plot(fileCSV)
```

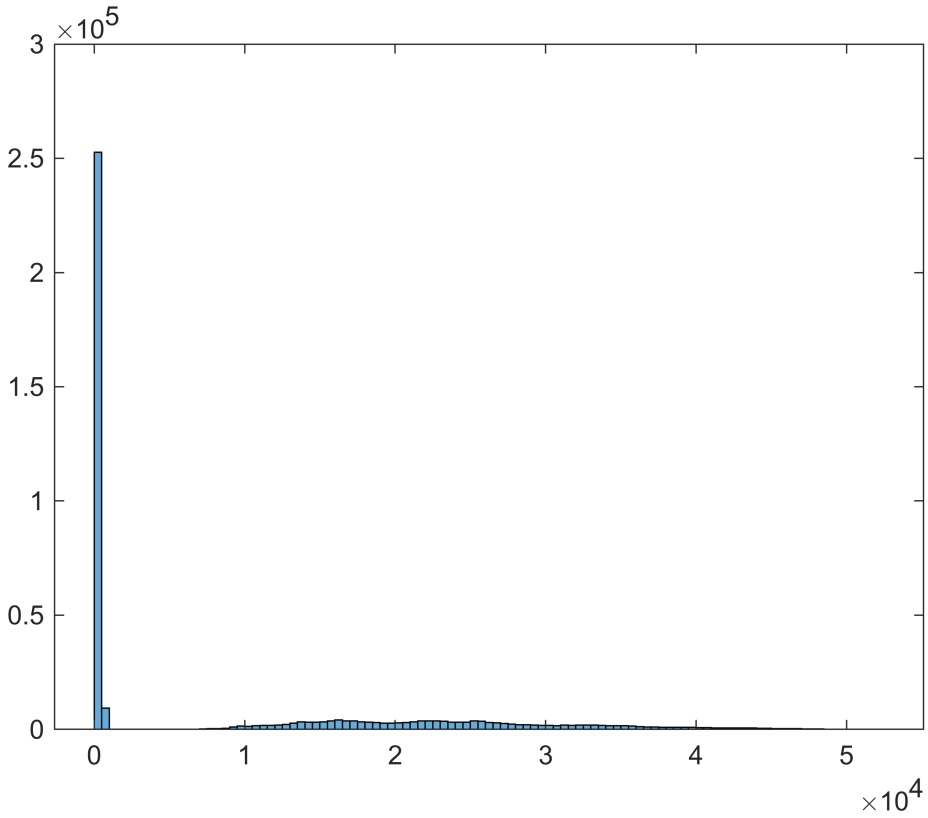


Answer 7

```
%% Select the last column
lastCol = fileCSV(:,8)
```

```
lastCol = 52415×1
10^4 ×
    2.0131
    1.9668
    1.8899
    1.8442
    1.8130
    1.7945
    1.7459
    1.7026
    1.6794
    1.6638
      ⋮
```

```
histogram(fileCSV)
```




```
%% k-means using the k value of 3
k_means_3 = kmeans(lastCol,3)
```

```
k_means_3 = 52415×1
     3
     3
     3
     3
     3
     3
     3
```

```
     3
     3
     3
     ⋮
```

```
%% %% k-means using the k value of 3
k_means_5 = kmeans(lastCol,5)
```

```
k_means_5 = 52415×1
     3
     3
     3
     3
     3
     3
     3
     4
     4
     4
     ⋮
```

```
gscatter(k_means_3, k_means_5)
```