notebook

October 9, 2020

[1]:
```
from utils import *
from itertools import combinations
from math import fabs
from matplotlib import pyplot as plt
import pandas as pd
pd.options.display.max_rows = 20
pd.options.display.max_columns = 16
```

1 Read and normalize data

[2]:
```
merged = merge_files()
merged
```

[2]:

	0	1	2	3	4	5	6	7	8	9	…	35	36	37	38	39	40	41	\
0	1	59	52	70	67	73	66	72	61	58	…	66	56	62	56	72	62	74	
1	1	72	62	69	67	78	82	74	65	69	…	65	71	63	60	69	73	67	
2	1	71	62	70	64	67	64	79	65	70	…	73	70	66	65	64	55	61	
3	1	69	71	70	78	61	63	67	65	59	…	61	61	66	65	72	73	68	
4	1	70	66	61	66	61	58	69	69	72	…	67	69	70	66	70	64	60	
5	1	57	69	68	75	69	74	73	71	57	…	63	58	69	67	79	77	72	
6	1	69	66	62	75	67	71	72	76	69	…	69	70	72	72	69	68	70	
7	1	61	60	60	62	64	72	68	67	74	…	66	66	66	60	60	58	60	
8	1	65	62	67	68	65	67	71	71	64	…	67	63	74	63	77	79	68	
9	1	74	73	72	79	66	61	76	66	65	…	64	62	73	69	62	67	60	
..	..	..	..	..	..	..	..	..	..	..	..	..	..	..	..	..	..	..	
257	0	61	68	62	70	76	79	71	71	73	…	69	69	66	71	67	66	74	
258	0	73	80	78	78	75	73	78	75	70	…	65	70	68	69	64	64	63	
259	0	56	56	63	66	76	76	68	73	62	…	71	73	60	53	61	73	67	
260	0	73	74	65	66	69	69	67	81	65	…	66	69	72	70	74	76	75	
261	0	59	58	69	74	71	73	70	68	57	…	66	66	61	56	74	71	72	
262	0	74	69	75	70	70	74	77	77	65	…	66	67	63	61	71	68	66	
263	0	72	61	64	66	64	59	68	66	76	…	69	64	67	71	69	68	65	
264	0	75	73	72	77	68	67	76	73	67	…	70	67	72	71	79	75	77	
265	0	59	62	72	74	66	66	74	76	63	…	65	71	67	69	77	78	77	
266	0	64	66	68	71	62	64	74	73	63	…	70	69	68	65	75	72	62	

	42	43	44

```
0    74  64  67
1    71  56  58
2    41  51  46
3    68  59  63
4    55  49  41
5    70  61  65
6    73  63  59
7    67  49  52
8    70  59  56
9    56  53  46
..   ..  ..  ..
257  71  58  58
258  63  56  51
259  68  59  56
260  72  67  63
261  69  62  60
262  65  54  57
263  73  56  52
264  75  67  71
265  76  70  70
266  64  57  54

[267 rows x 45 columns]
```

[3]:
```
normalized = normalize(merged)
print('Normalized dataset:')
normalized
```

```
There are NO duplicated rows in the data.
Normalized dataset:
```

[3]:

	0	1	2	3	4	5	6	7	\
0	1.0	0.600000	0.516667	0.727273	0.66	0.852459	0.640625	0.796875	
1	1.0	0.816667	0.683333	0.704545	0.66	0.934426	0.890625	0.828125	
2	1.0	0.800000	0.683333	0.727273	0.60	0.754098	0.609375	0.906250	
3	1.0	0.766667	0.833333	0.727273	0.88	0.655738	0.593750	0.718750	
4	1.0	0.783333	0.750000	0.522727	0.64	0.655738	0.515625	0.750000	
5	1.0	0.566667	0.800000	0.681818	0.82	0.786885	0.765625	0.812500	
6	1.0	0.766667	0.750000	0.545455	0.82	0.754098	0.718750	0.796875	
7	1.0	0.633333	0.650000	0.500000	0.56	0.704918	0.734375	0.734375	
8	1.0	0.700000	0.683333	0.659091	0.68	0.721311	0.656250	0.781250	
9	1.0	0.850000	0.866667	0.772727	0.90	0.737705	0.562500	0.859375	
..	…	…	…	…	…	…	…	…	
257	0.0	0.633333	0.783333	0.545455	0.72	0.901639	0.843750	0.781250	
258	0.0	0.833333	0.983333	0.909091	0.88	0.885246	0.750000	0.890625	
259	0.0	0.550000	0.583333	0.568182	0.64	0.901639	0.796875	0.734375	
260	0.0	0.833333	0.883333	0.613636	0.64	0.786885	0.687500	0.718750	

```
261  0.0  0.600000  0.616667  0.704545  0.80  0.819672  0.750000  0.765625
262  0.0  0.850000  0.800000  0.840909  0.72  0.803279  0.765625  0.875000
263  0.0  0.816667  0.666667  0.590909  0.64  0.704918  0.531250  0.734375
264  0.0  0.866667  0.866667  0.772727  0.86  0.770492  0.656250  0.859375
265  0.0  0.600000  0.683333  0.772727  0.80  0.737705  0.640625  0.828125
266  0.0  0.683333  0.750000  0.681818  0.74  0.672131  0.609375  0.828125

            8         9   …        35        36        37        38  \
0    0.650794  0.707692  …  0.827586  0.671875  0.617021  0.622951
1    0.714286  0.876923  …  0.810345  0.906250  0.638298  0.688525
2    0.714286  0.892308  …  0.948276  0.890625  0.702128  0.770492
3    0.714286  0.723077  …  0.741379  0.750000  0.702128  0.770492
4    0.777778  0.923077  …  0.844828  0.875000  0.787234  0.786885
5    0.809524  0.692308  …  0.775862  0.703125  0.765957  0.803279
6    0.888889  0.876923  …  0.879310  0.890625  0.829787  0.885246
7    0.746032  0.953846  …  0.827586  0.828125  0.702128  0.688525
8    0.809524  0.800000  …  0.844828  0.781250  0.872340  0.737705
9    0.730159  0.815385  …  0.793103  0.765625  0.851064  0.836066
..        …       …  …       …         …         …         …
257  0.809524  0.938462  …  0.879310  0.875000  0.702128  0.868852
258  0.873016  0.892308  …  0.810345  0.890625  0.744681  0.836066
259  0.841270  0.769231  …  0.913793  0.937500  0.574468  0.573770
260  0.968254  0.815385  …  0.827586  0.875000  0.829787  0.852459
261  0.761905  0.692308  …  0.827586  0.828125  0.595745  0.622951
262  0.904762  0.815385  …  0.827586  0.843750  0.638298  0.704918
263  0.730159  0.984615  …  0.879310  0.796875  0.723404  0.868852
264  0.841270  0.846154  …  0.896552  0.843750  0.829787  0.868852
265  0.888889  0.784615  …  0.810345  0.906250  0.723404  0.836066
266  0.841270  0.784615  …  0.896552  0.875000  0.744681  0.770492

            39        40        41        42        43        44
0    0.847222  0.746667  0.896104  0.880000  0.746479  0.913043
1    0.805556  0.893333  0.805195  0.840000  0.633803  0.782609
2    0.736111  0.653333  0.727273  0.440000  0.563380  0.608696
3    0.847222  0.893333  0.818182  0.800000  0.676056  0.855072
4    0.819444  0.773333  0.714286  0.626667  0.535211  0.536232
5    0.944444  0.946667  0.870130  0.826667  0.704225  0.884058
6    0.805556  0.826667  0.844156  0.866667  0.732394  0.797101
7    0.680556  0.693333  0.714286  0.786667  0.535211  0.695652
8    0.916667  0.973333  0.818182  0.826667  0.676056  0.753623
9    0.708333  0.813333  0.714286  0.640000  0.591549  0.608696
..        …         …         …         …         …         …
257  0.777778  0.800000  0.896104  0.840000  0.661972  0.782609
258  0.736111  0.773333  0.753247  0.733333  0.633803  0.681159
259  0.694444  0.893333  0.805195  0.800000  0.676056  0.753623
260  0.875000  0.933333  0.909091  0.853333  0.788732  0.855072
261  0.875000  0.866667  0.870130  0.813333  0.718310  0.811594
```

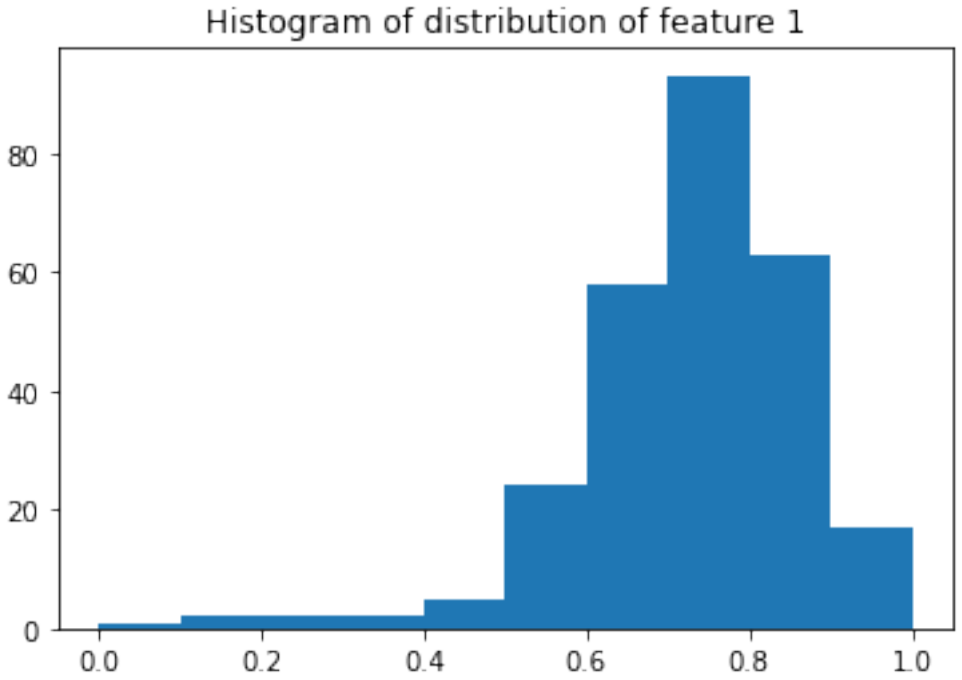
```
262  0.833333  0.826667  0.792208  0.760000  0.605634  0.768116
263  0.805556  0.826667  0.779221  0.866667  0.633803  0.695652
264  0.944444  0.920000  0.935065  0.893333  0.788732  0.971014
265  0.916667  0.960000  0.935065  0.906667  0.830986  0.956522
266  0.888889  0.880000  0.740260  0.746667  0.647887  0.724638

[267 rows x 45 columns]
```

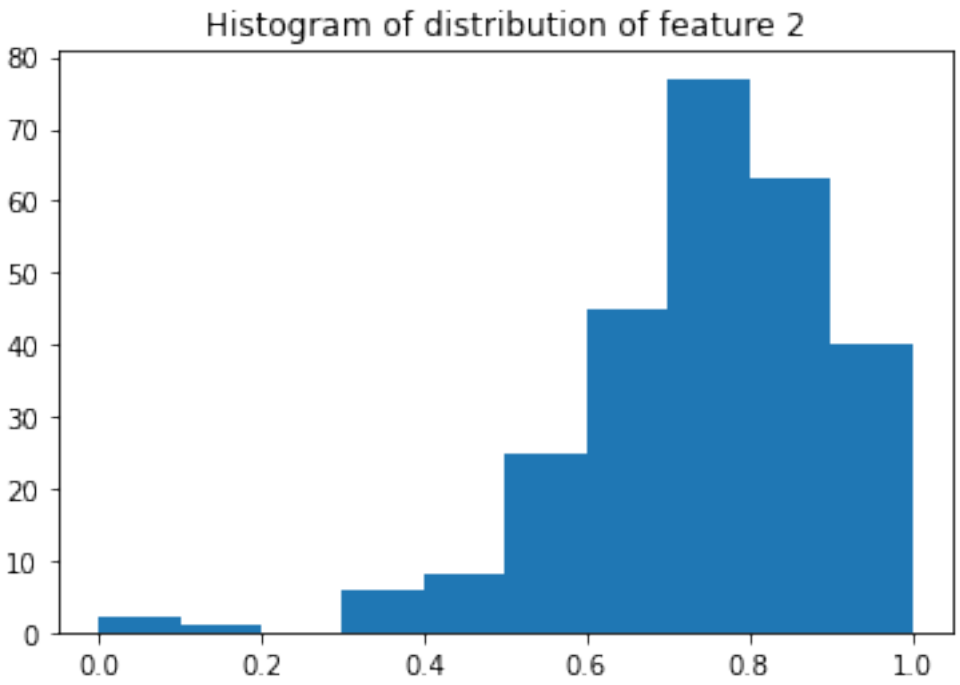
2 Check if any feature in the dataset is normally distributed

[4]:
```
plot_normality(normalized)
```

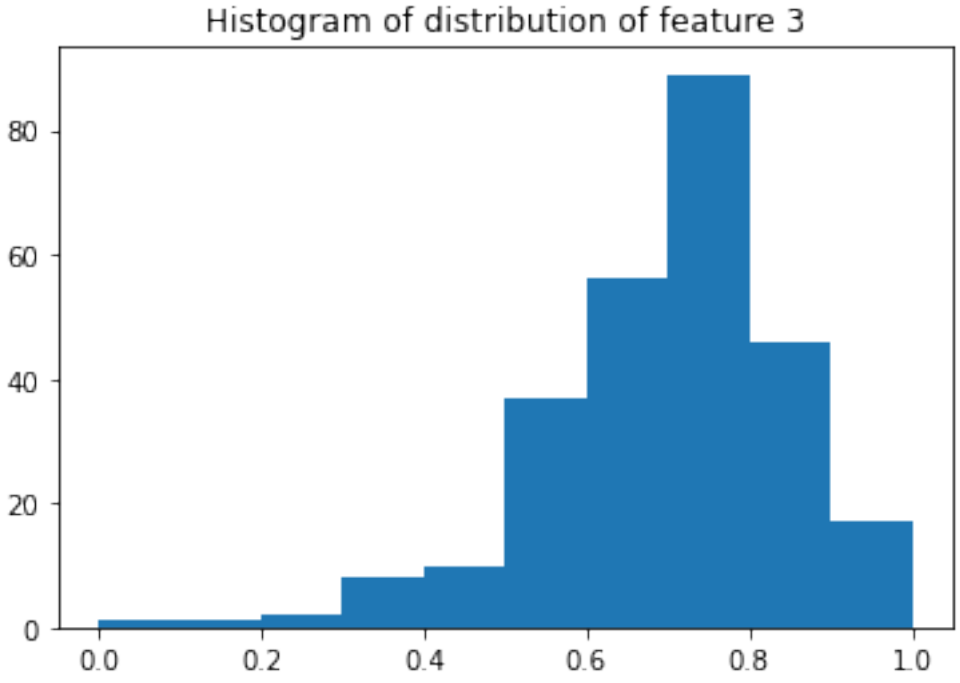
```
Normality test for feature 1:
        P-value: 9.388539447294658e-21  Samples do not come from a normal
distribution.
```



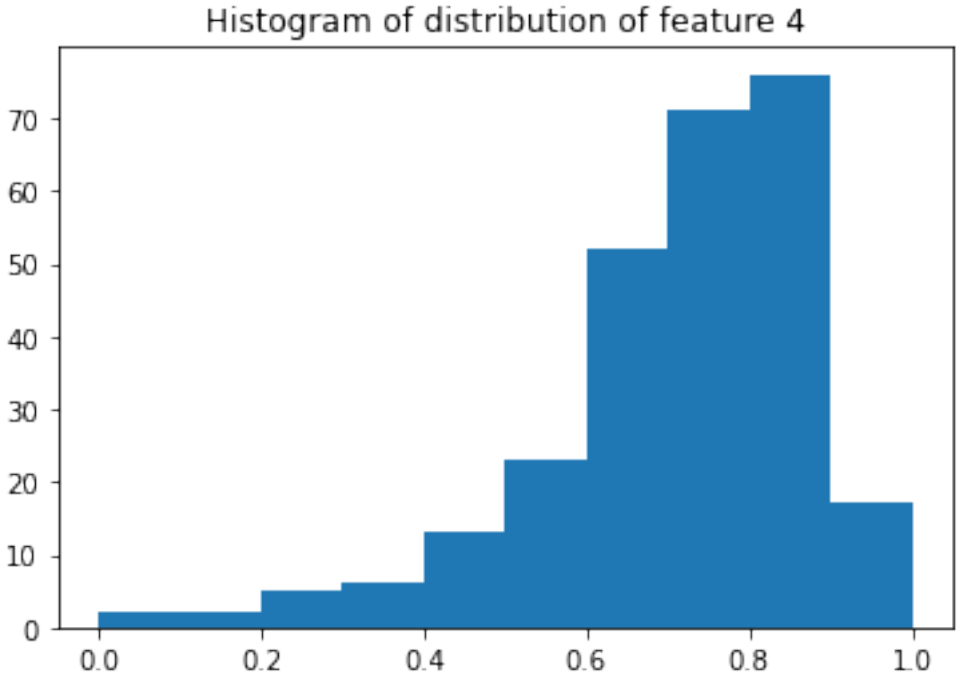

```
Normality test for feature 2:
        P-value: 2.4291153928630507e-15 Samples do not come from a normal
distribution.
```



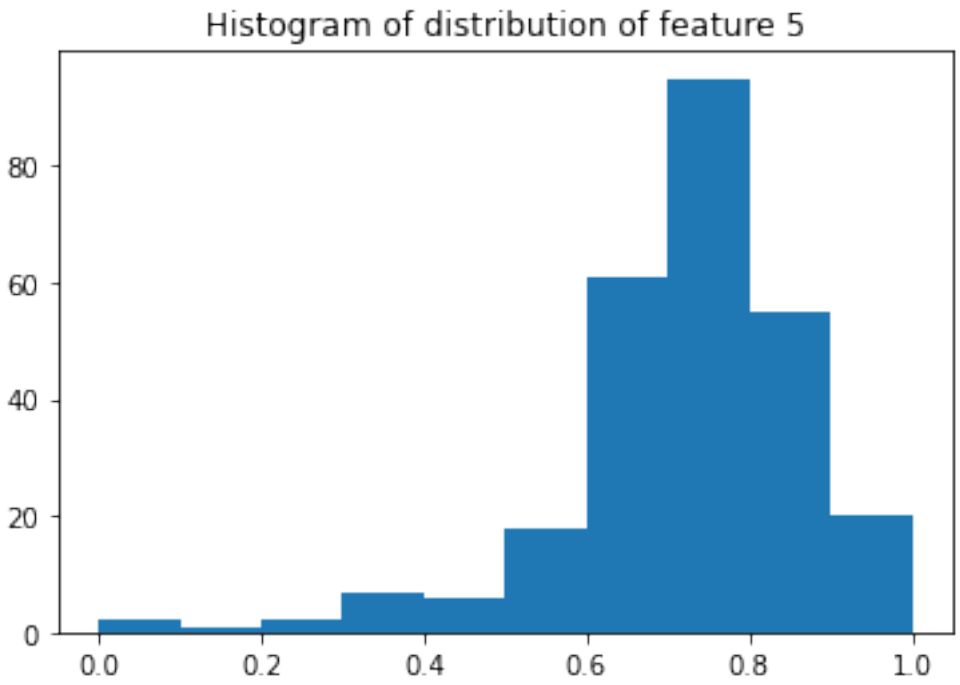

```
Normality test for feature 3:
        P-value: 2.285888250610929e-10  Samples do not come from a normal
distribution.
```



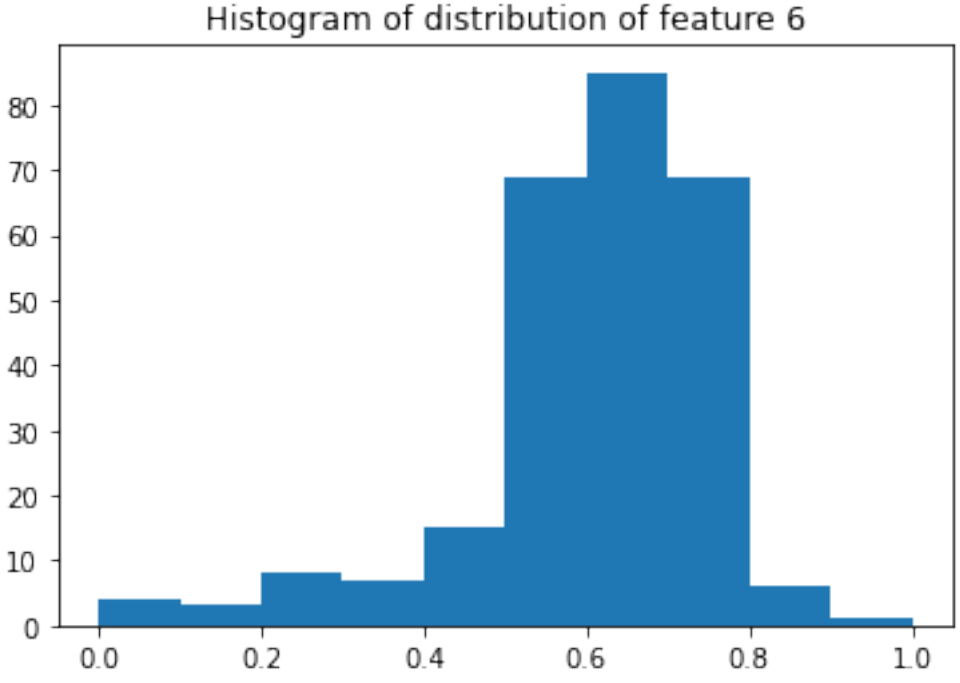
```
Normality test for feature 4:
        P-value: 2.414973546608377e-16  Samples do not come from a normal
distribution.
```



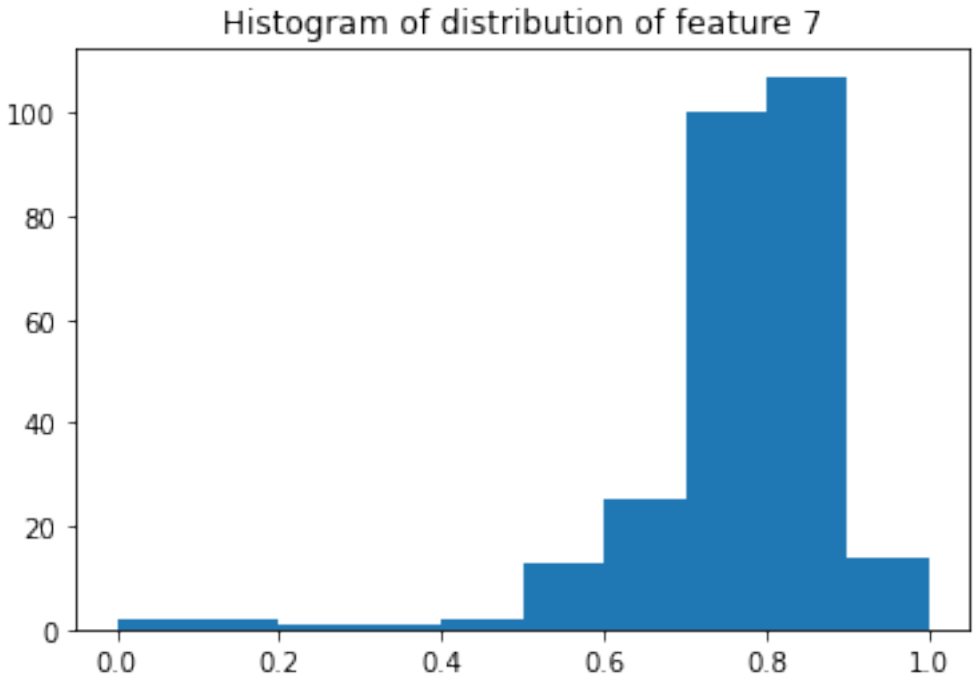

```
Normality test for feature 5:
        P-value: 5.249140735898784e-22  Samples do not come from a normal
distribution.
```



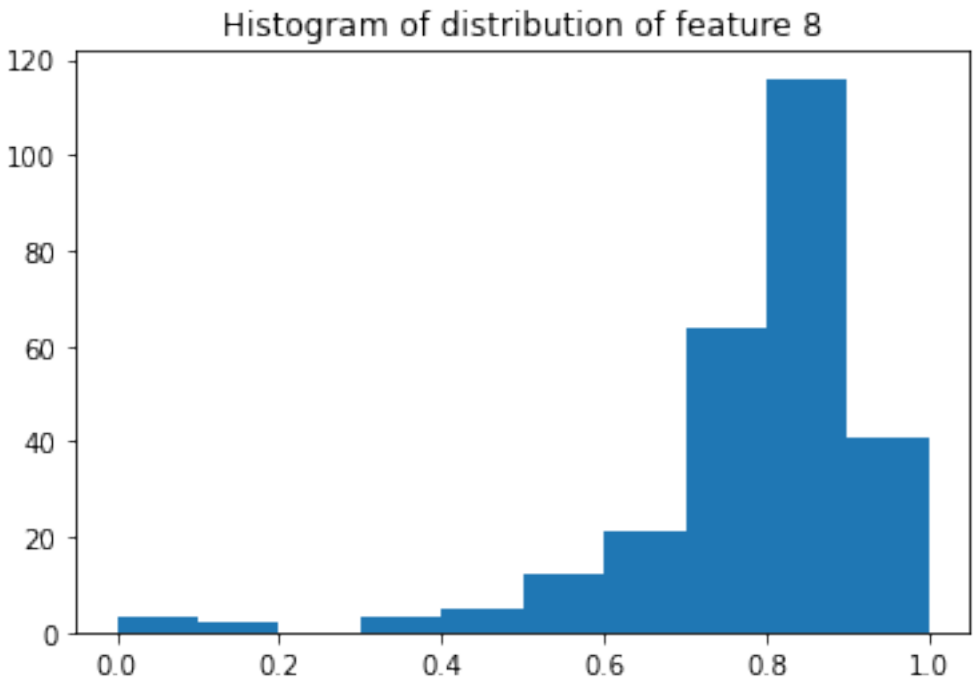

```
Normality test for feature 6:
        P-value: 2.891570227649463e-18  Samples do not come from a normal
distribution.
```



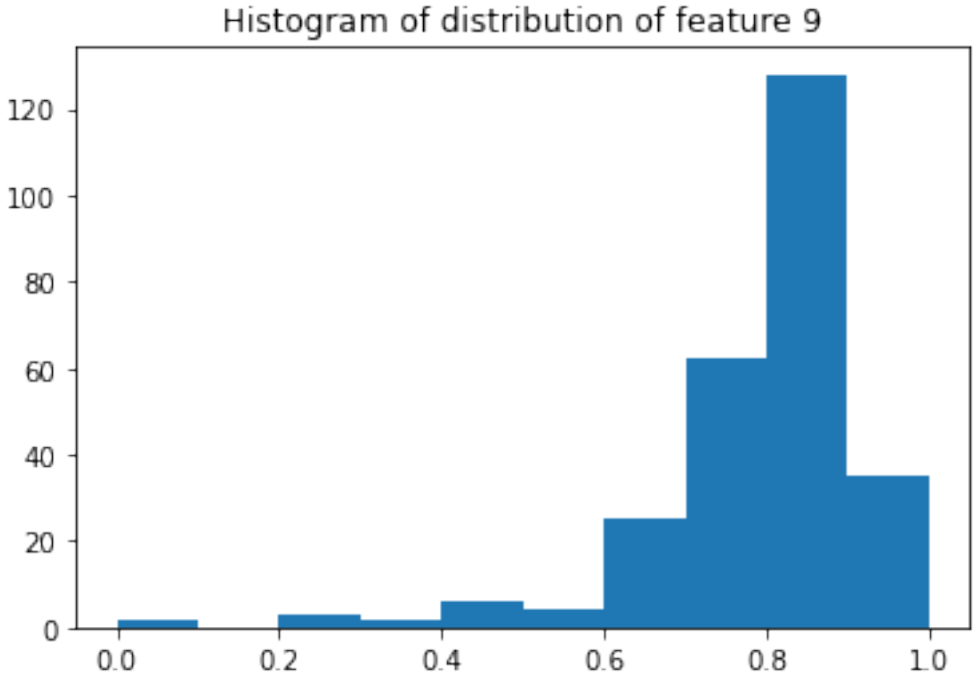
```
Normality test for feature 7:
        P-value: 9.225922046783433e-39  Samples do not come from a normal
distribution.
```



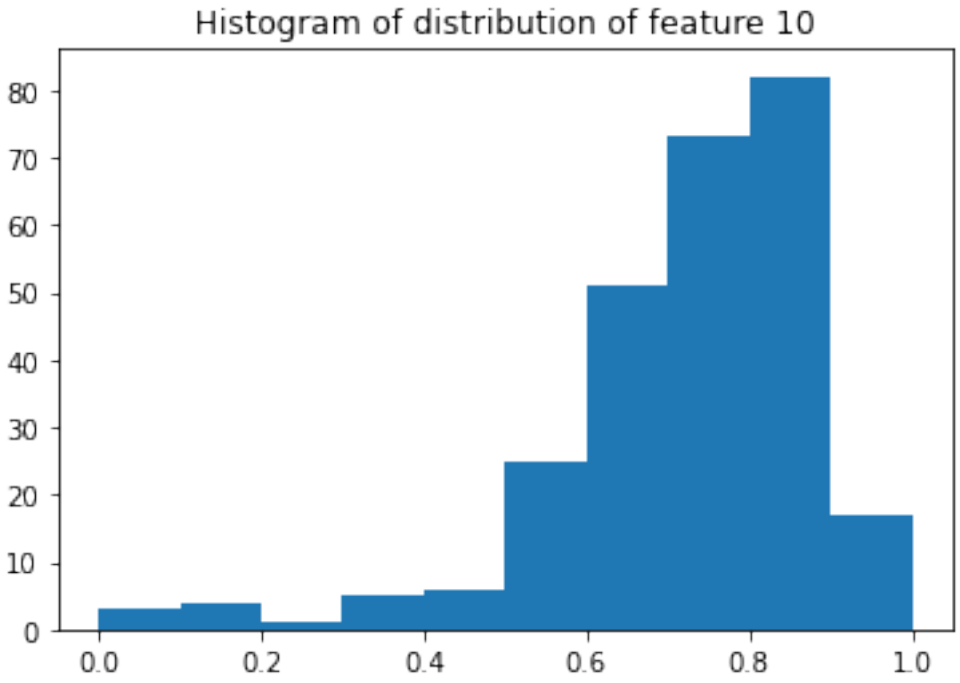

```
Normality test for feature 8:
        P-value: 4.805694765410013e-34  Samples do not come from a normal
distribution.
```



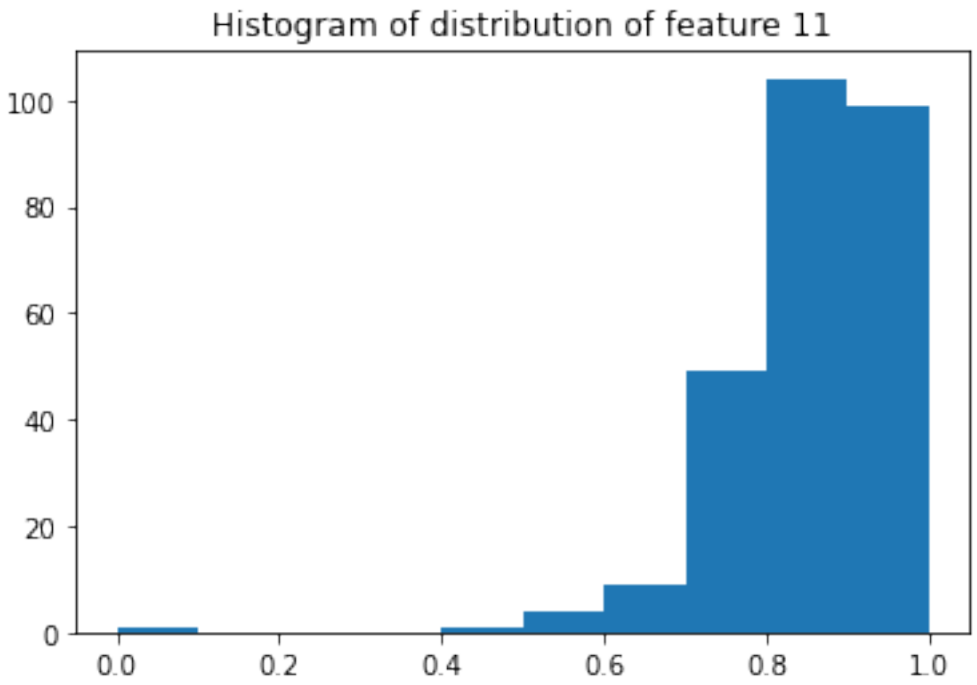

```
Normality test for feature 9:
        P-value: 9.388481895657816e-36  Samples do not come from a normal
distribution.
```



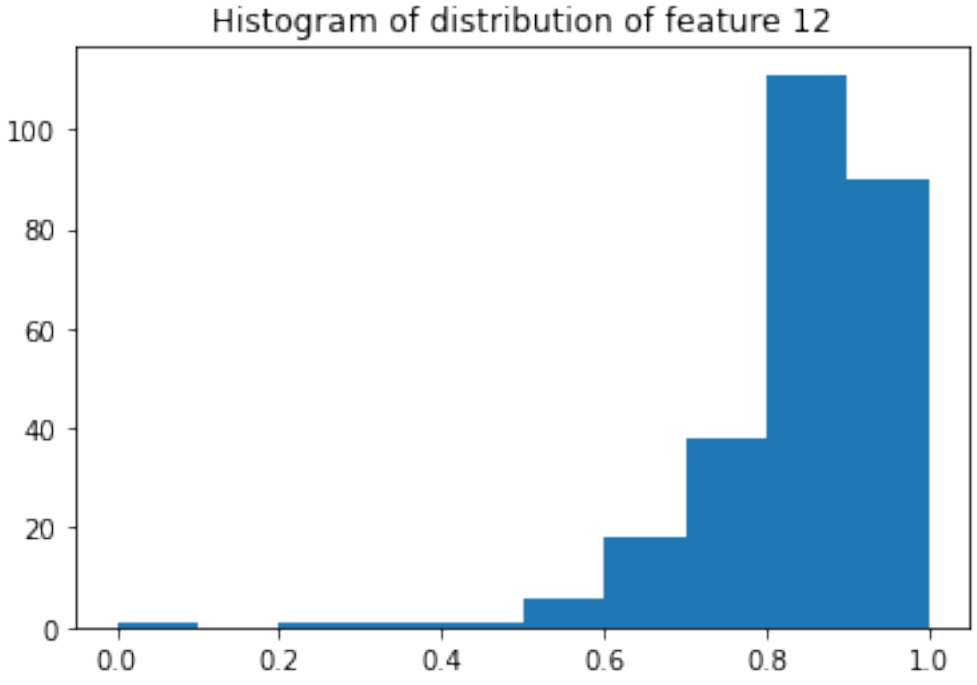
```
Normality test for feature 10:
        P-value: 6.059938944582491e-22  Samples do not come from a normal
distribution.
```



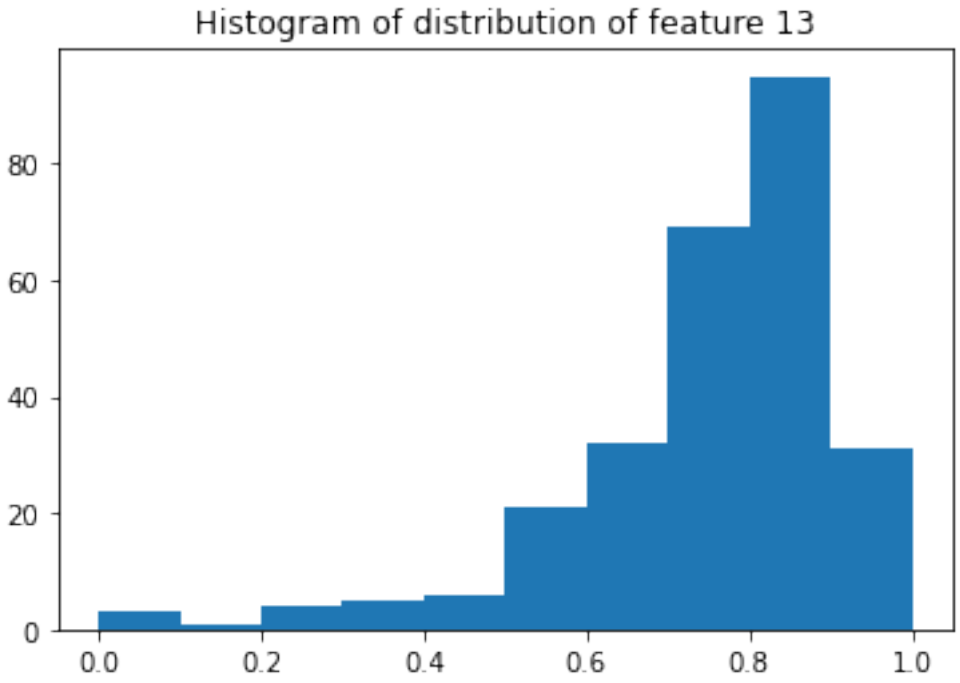

```
Normality test for feature 11:
        P-value: 7.559361280515924e-44  Samples do not come from a normal
distribution.
```



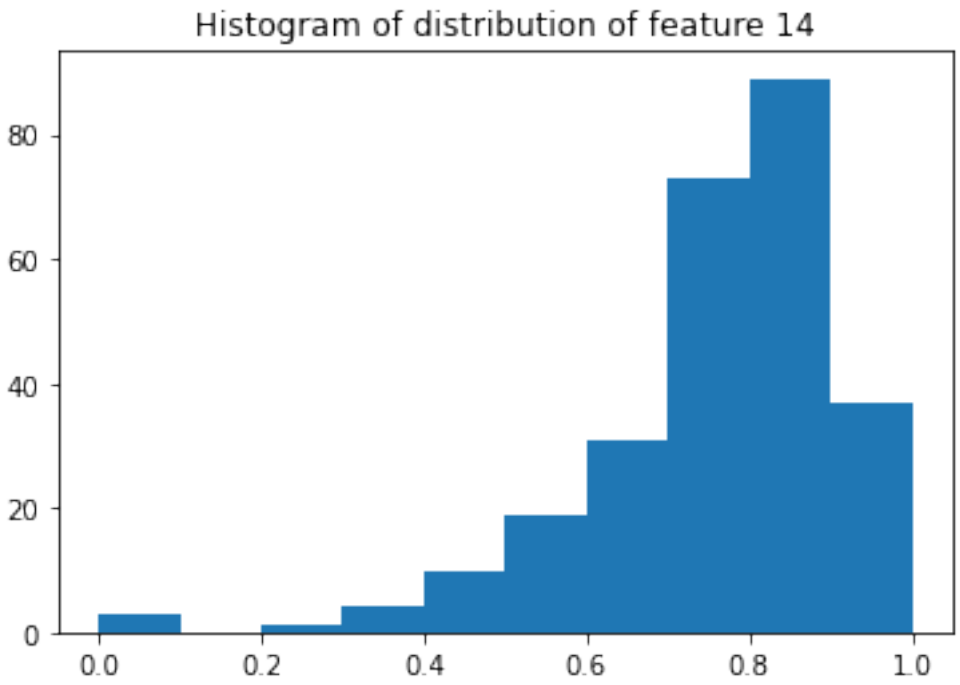

```
Normality test for feature 12:
        P-value: 1.5091914699506185e-36 Samples do not come from a normal
distribution.
```



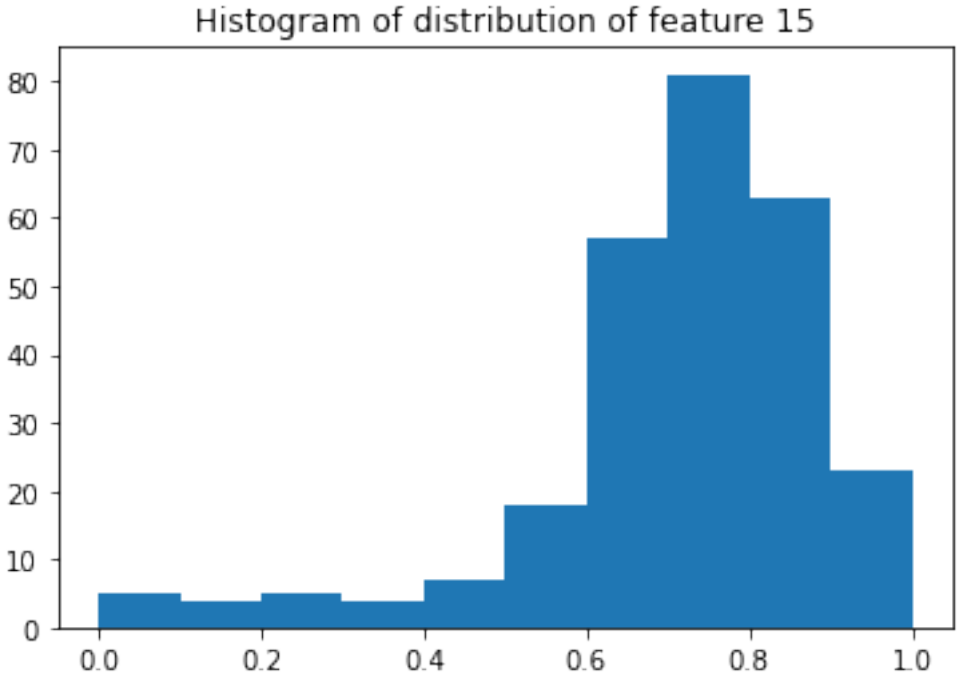
```
Normality test for feature 13:
        P-value: 4.859557613733641e-22  Samples do not come from a normal
distribution.
```



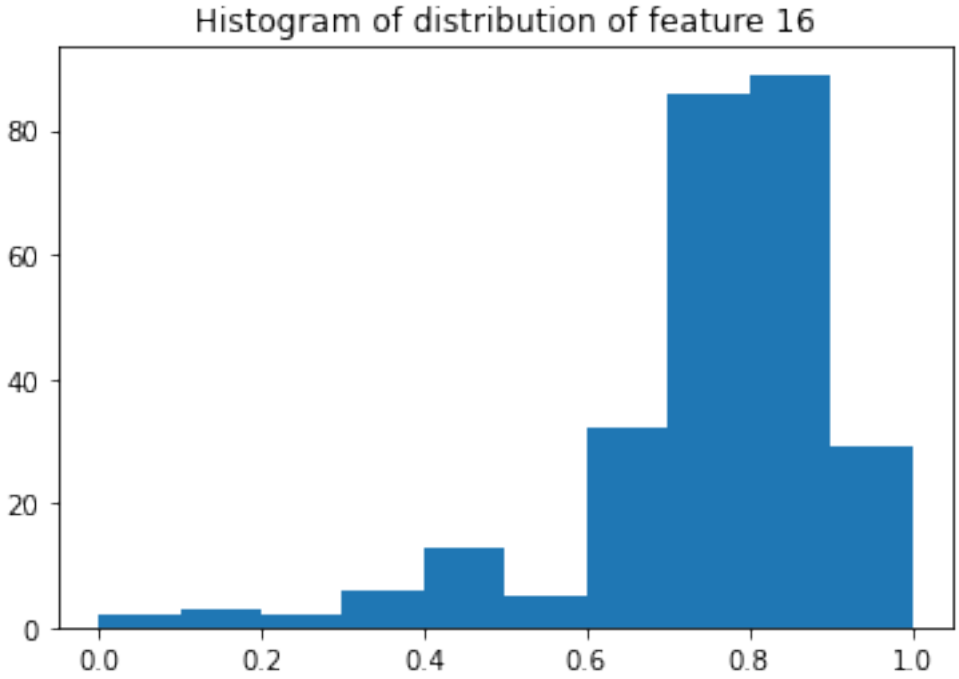

```
Normality test for feature 14:
        P-value: 3.566716933865817e-24  Samples do not come from a normal
distribution.
```



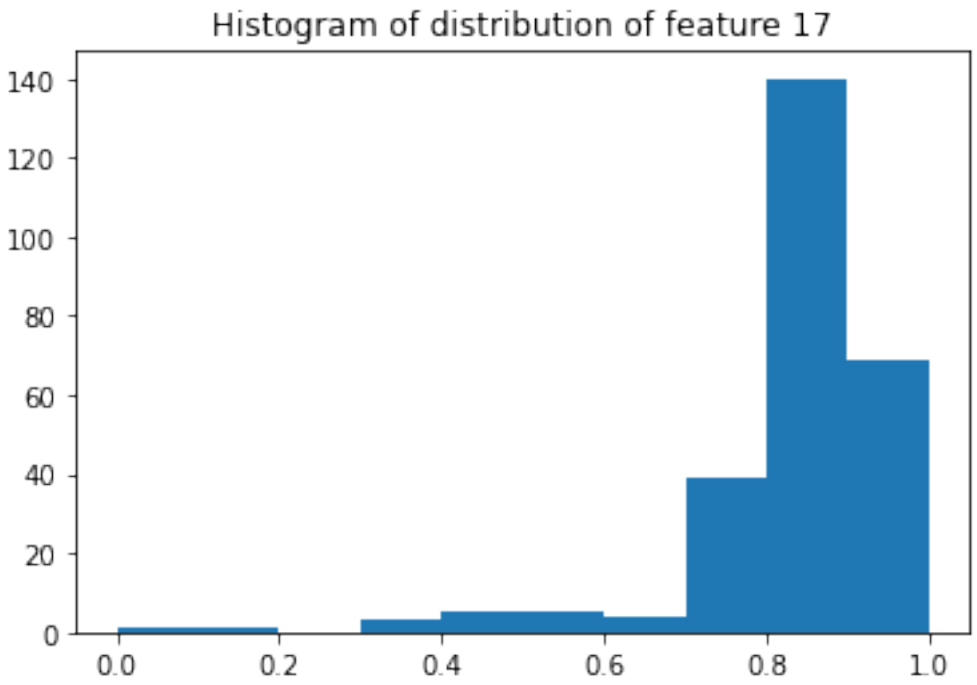

```
Normality test for feature 15:
        P-value: 1.480807054434905e-20  Samples do not come from a normal
distribution.
```



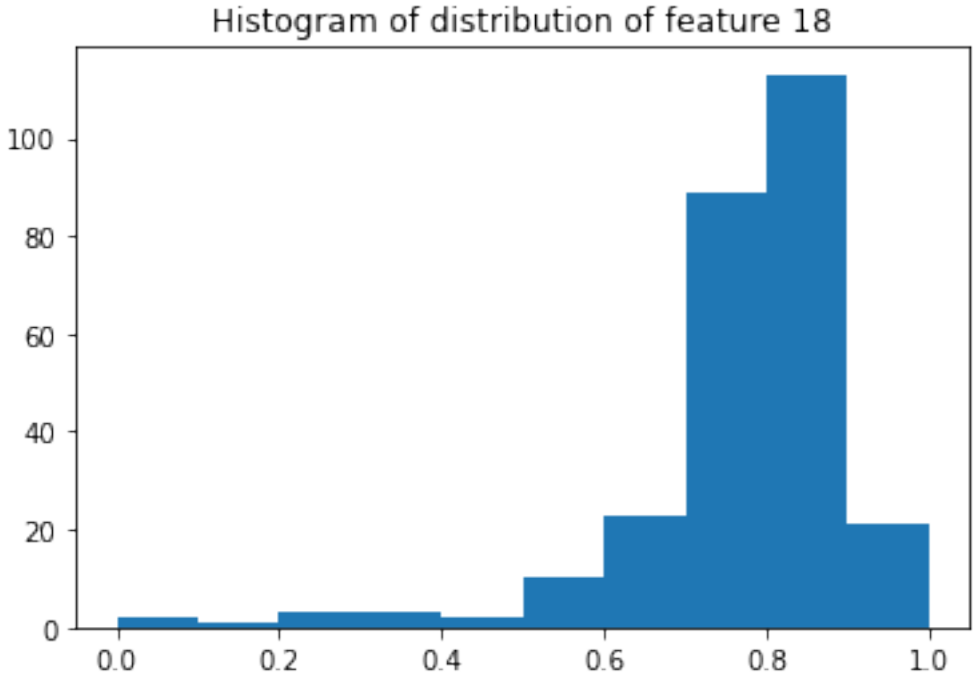
Normality test for feature 16:
P-value: 3.4537391274169423e-23 Samples do not come from a normal distribution.



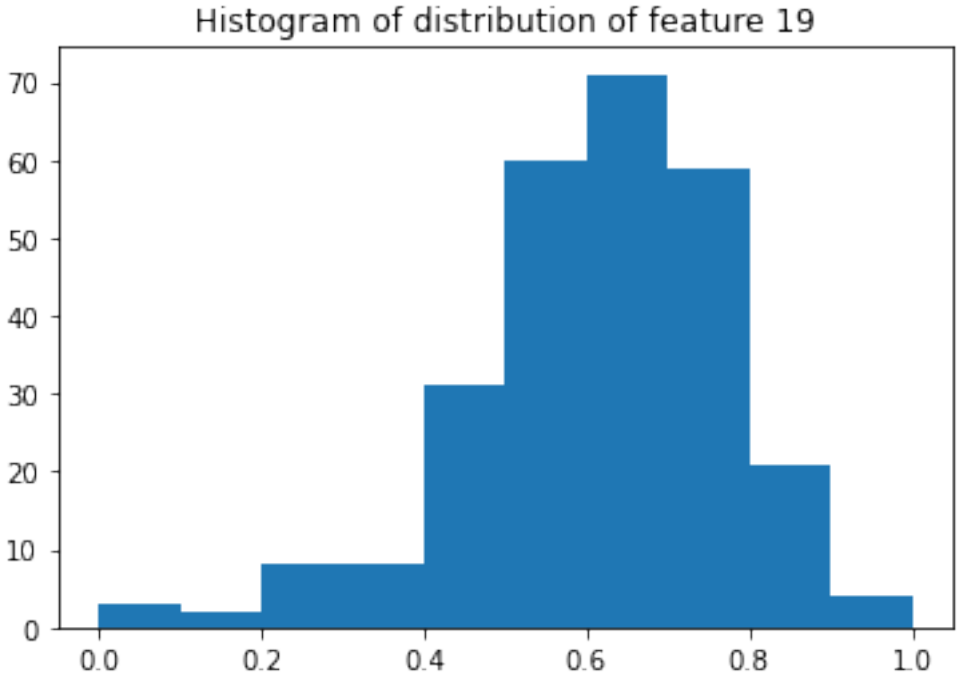
Normality test for feature 17:
P-value: 7.770661401049436e-44 Samples do not come from a normal distribution.



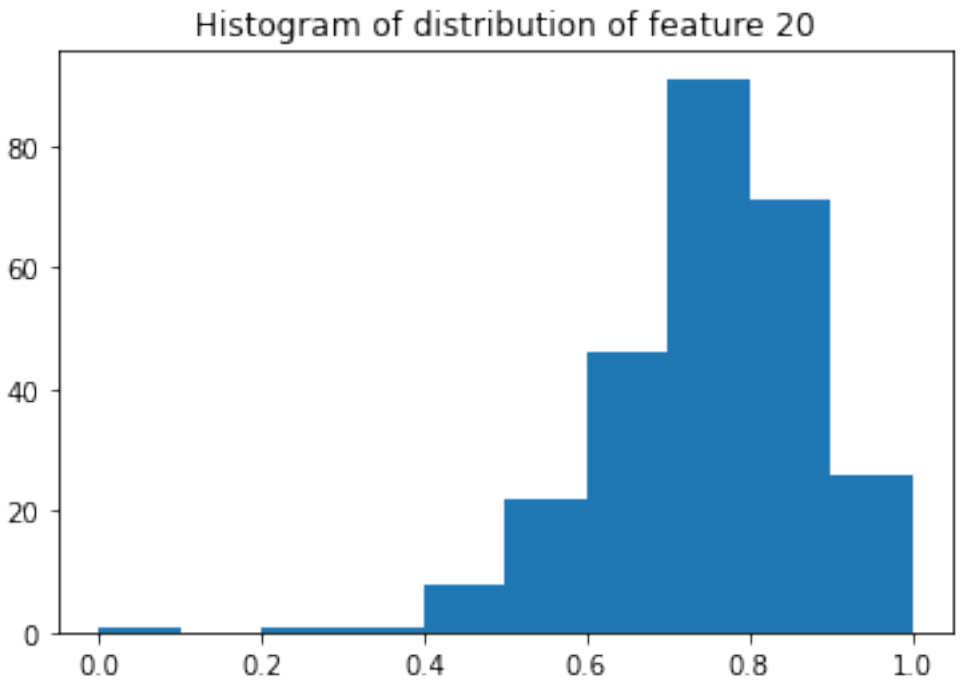

```
Normality test for feature 18:
        P-value: 5.0721309449648754e-37 Samples do not come from a normal
distribution.
```



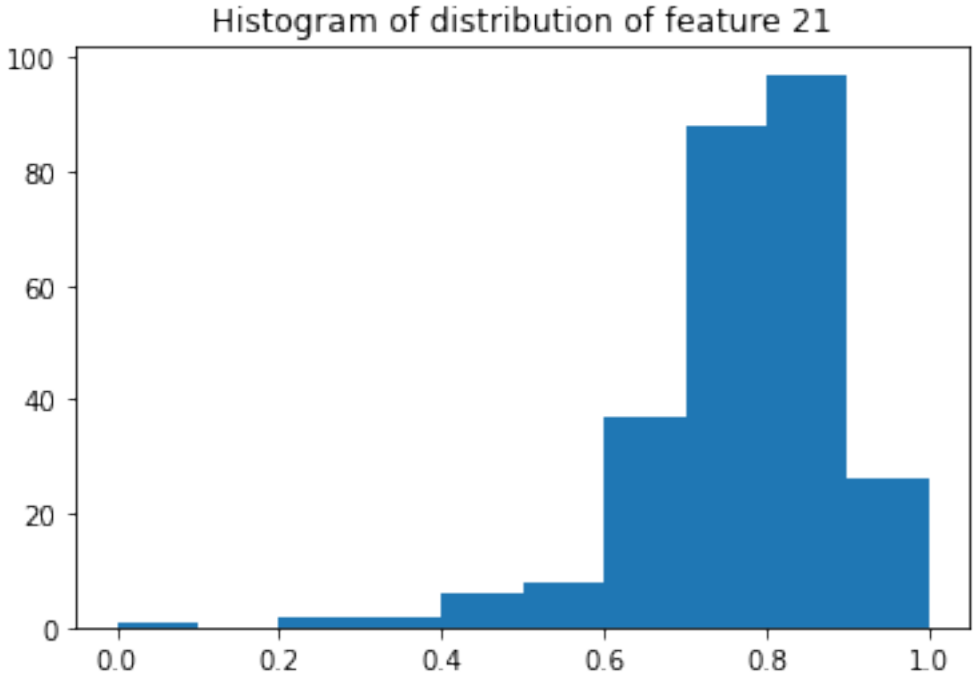
Normality test for feature 19:
P-value: 2.1148738984059247e-09 Samples do not come from a normal distribution.



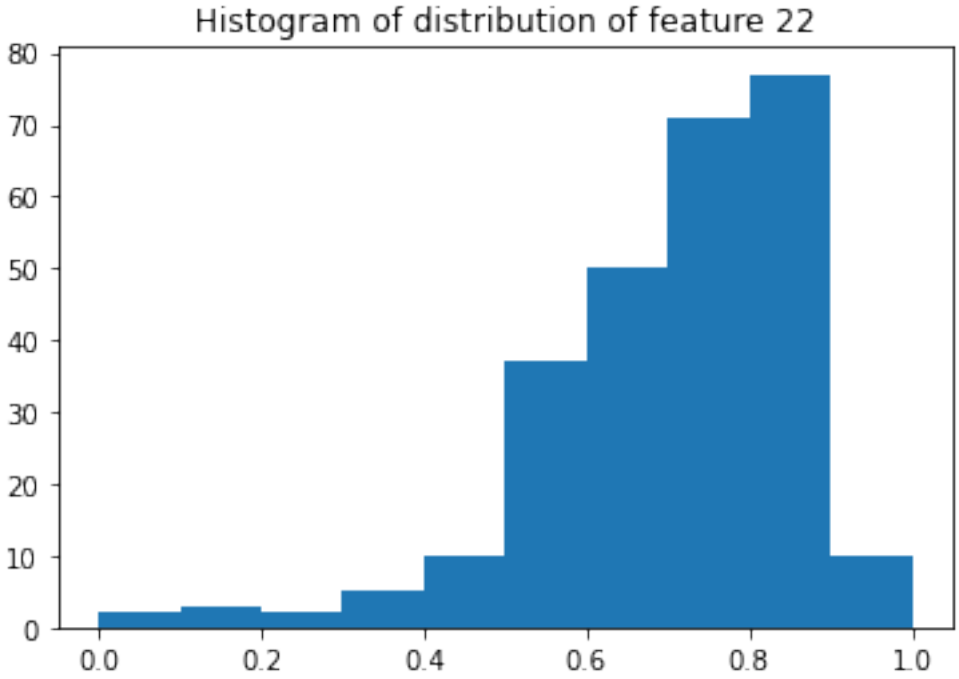
Normality test for feature 20:
P-value: 7.101699312967085e-17 Samples do not come from a normal distribution.



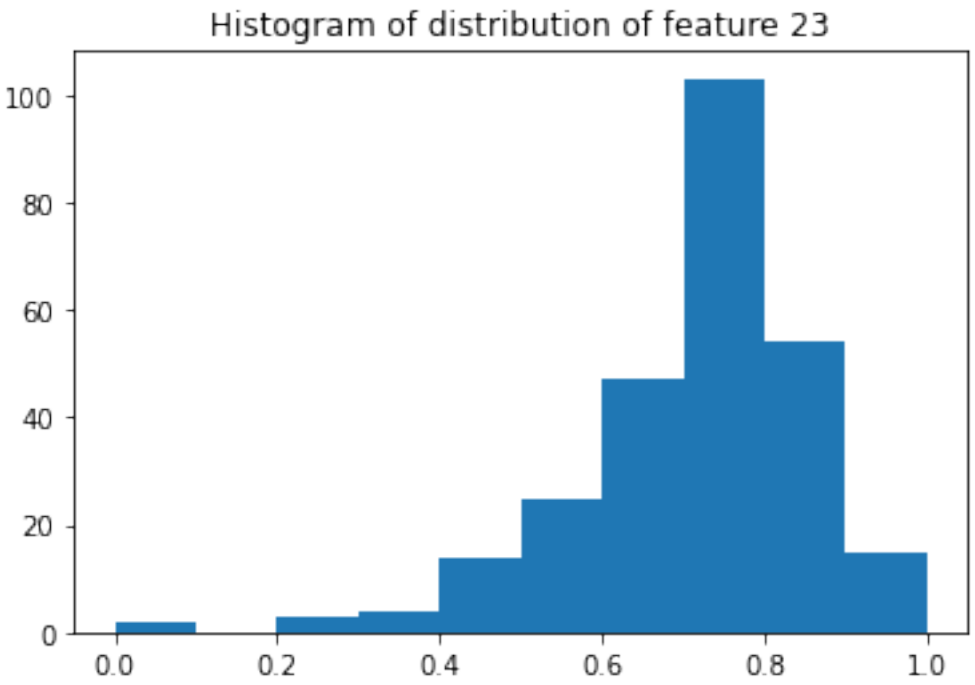

```
Normality test for feature 21:
        P-value: 3.9283389279759544e-28 Samples do not come from a normal
distribution.
```



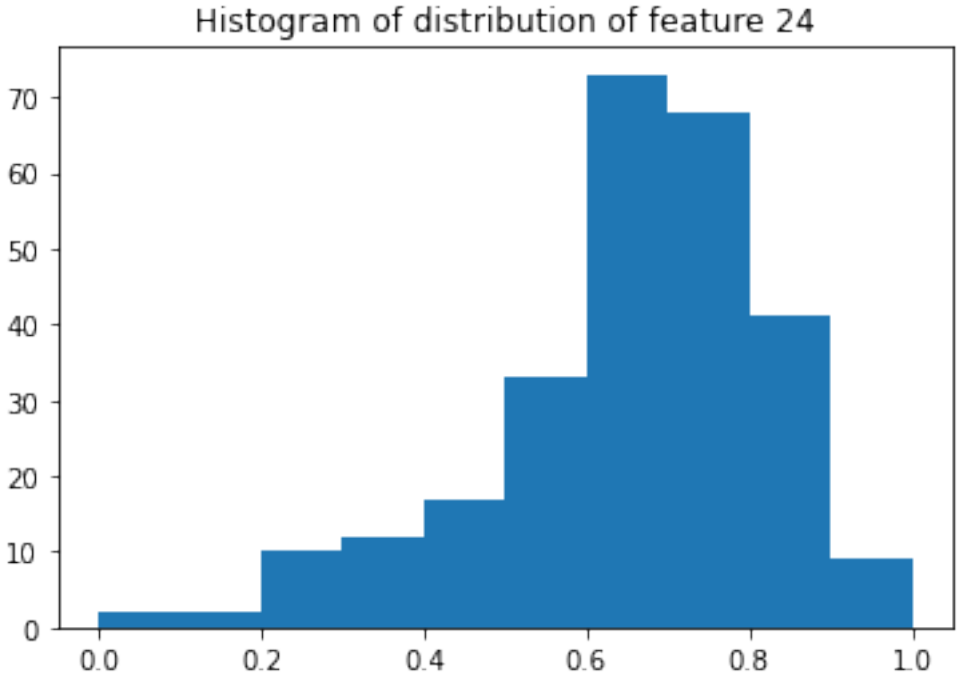
Normality test for feature 22:
        P-value: 1.397447512833632e-17 Samples do not come from a normal distribution.



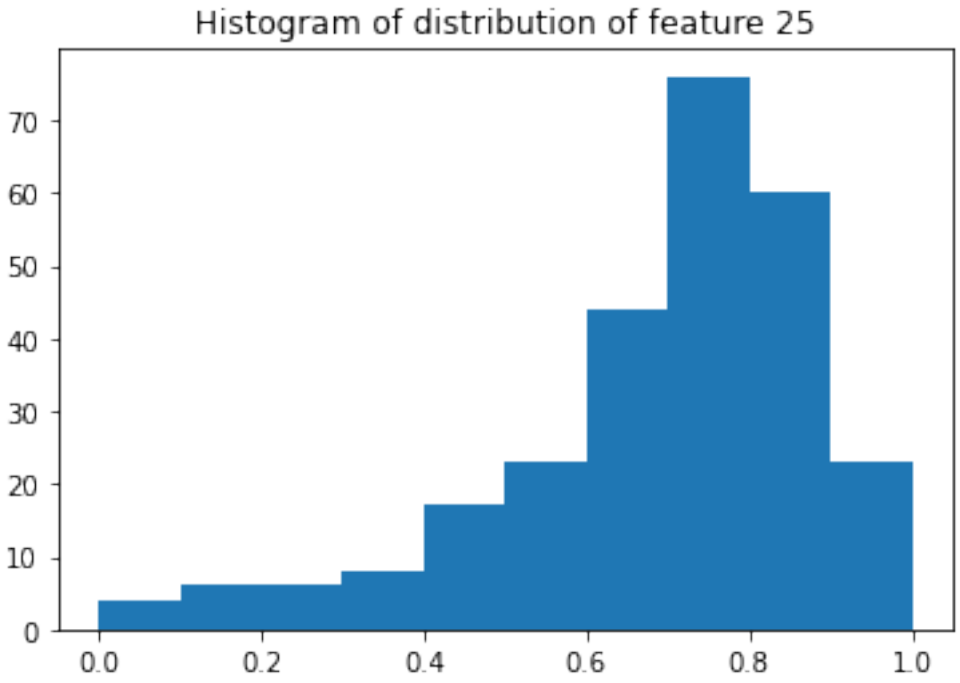
Normality test for feature 23:
        P-value: 3.797283987540923e-17 Samples do not come from a normal distribution.



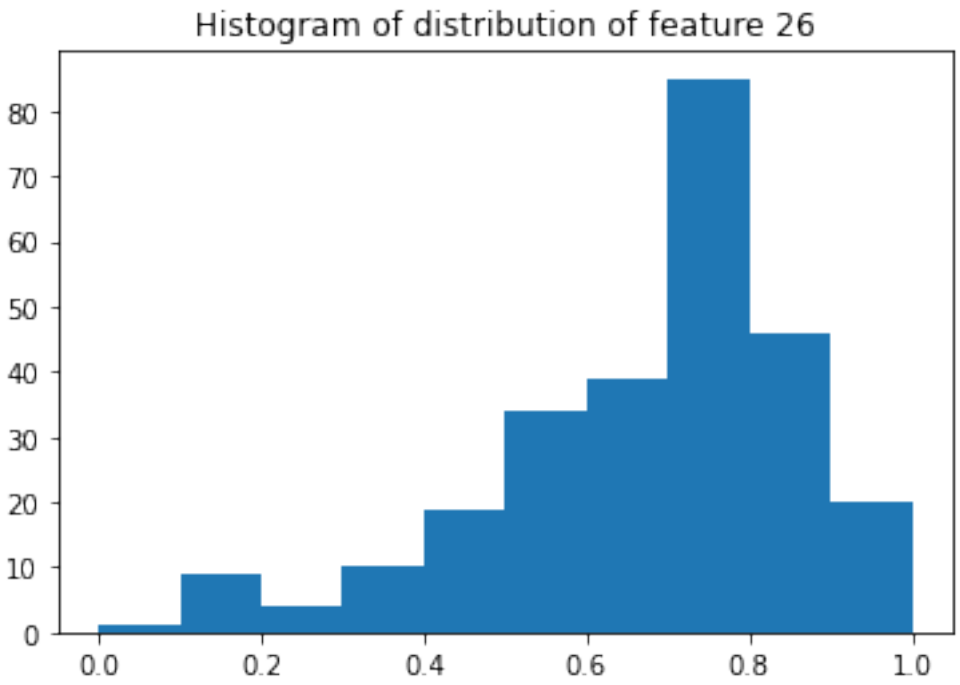

```
Normality test for feature 24:
        P-value: 1.7072787475610888e-10 Samples do not come from a normal
distribution.
```



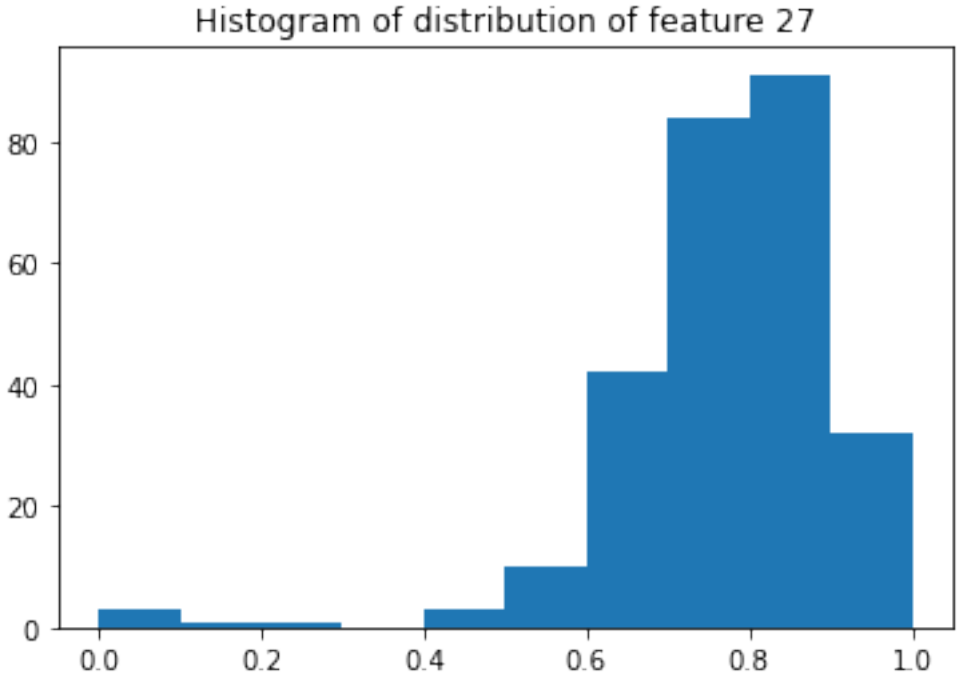
```
Normality test for feature 25:
        P-value: 3.103618993795135e-12  Samples do not come from a normal
distribution.
```



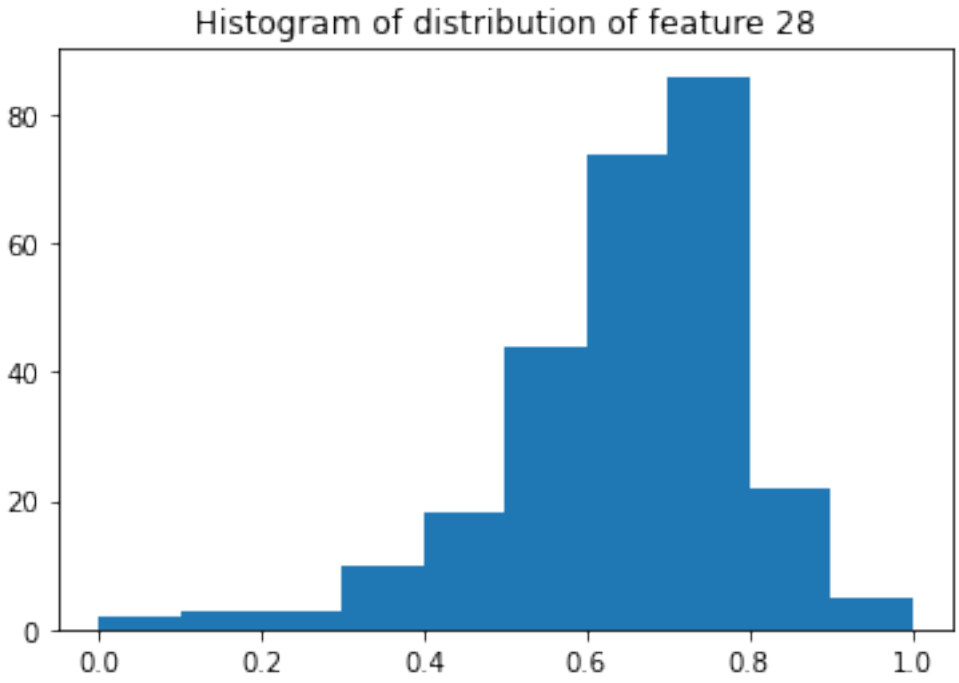

```
Normality test for feature 26:
        P-value: 2.472966175011184e-09  Samples do not come from a normal
distribution.
```



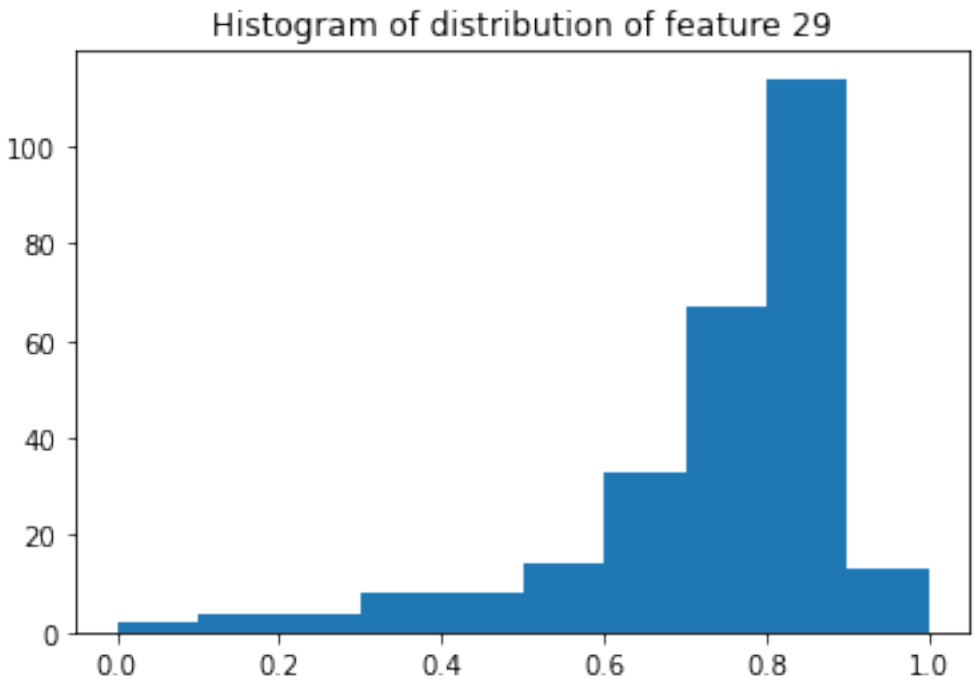

```
Normality test for feature 27:
        P-value: 2.463969335053877e-35  Samples do not come from a normal
distribution.
```



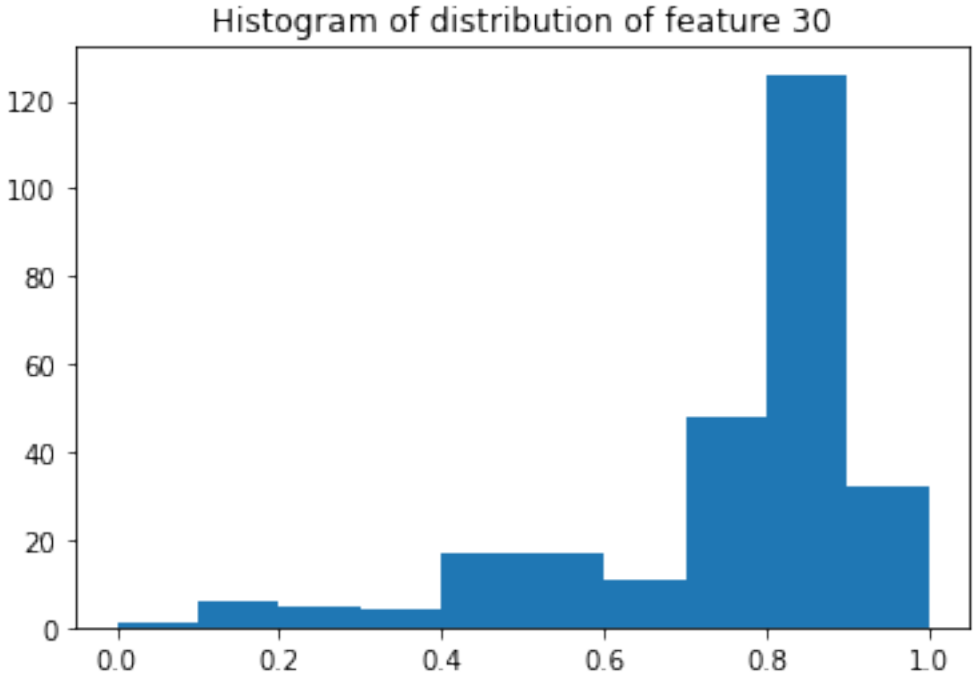
```
Normality test for feature 28:
        P-value: 9.204132681866072e-14  Samples do not come from a normal
distribution.
```



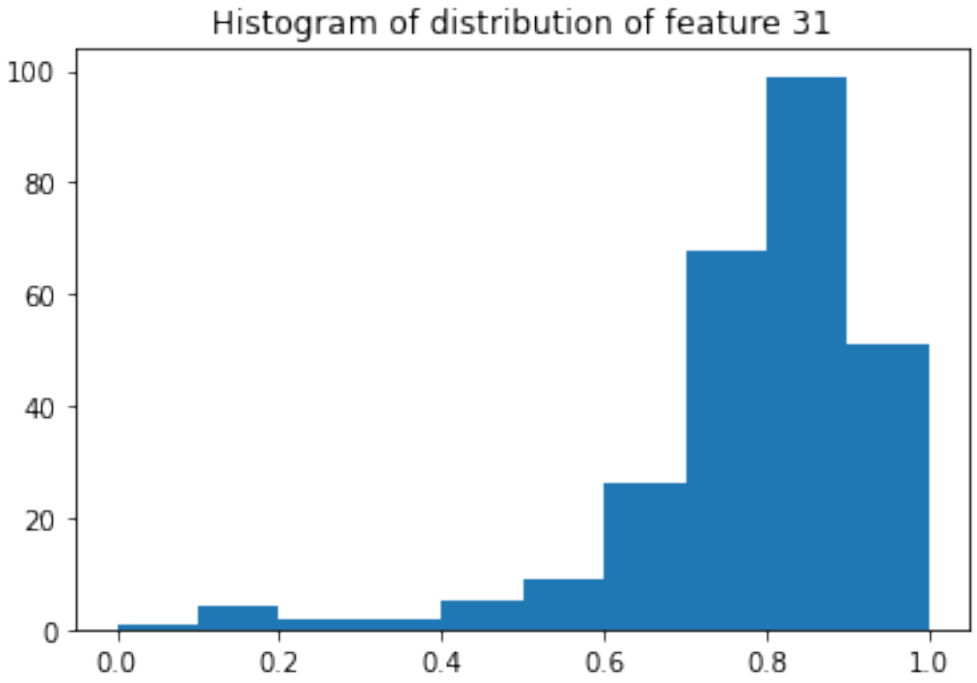

```
Normality test for feature 29:
        P-value: 3.211287389340235e-24  Samples do not come from a normal
distribution.
```



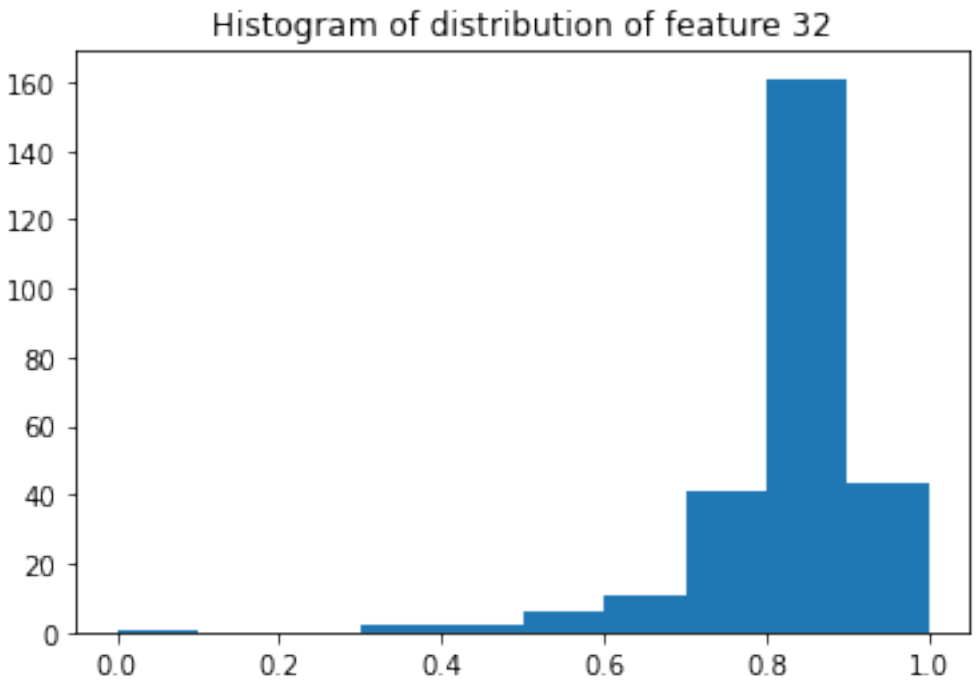

```
Normality test for feature 30:
        P-value: 2.1277620647960742e-20 Samples do not come from a normal
distribution.
```



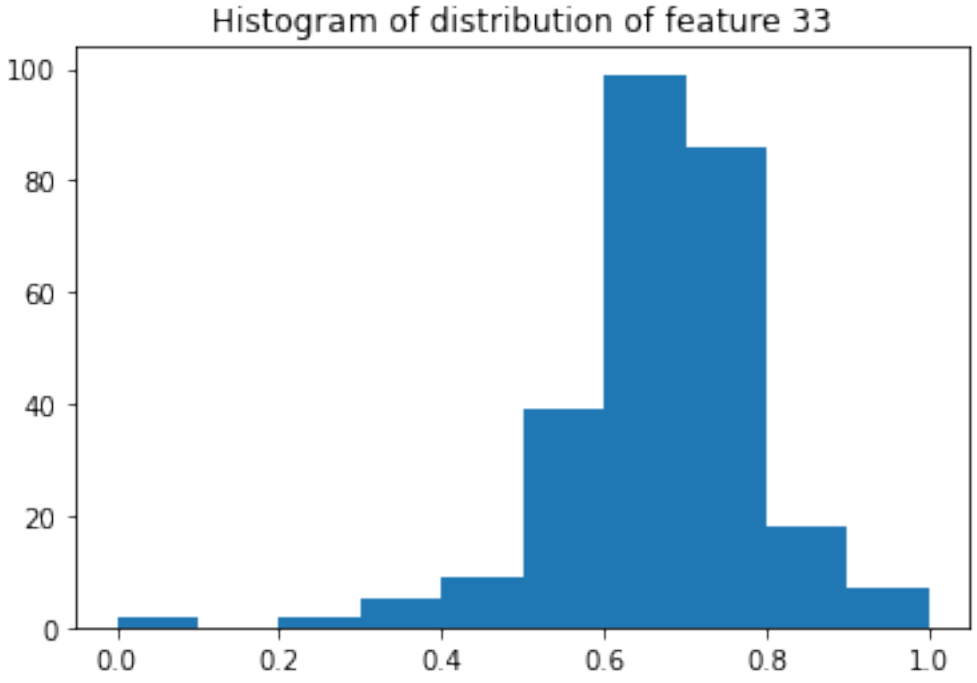
Normality test for feature 31:
P-value: 2.2163073653823443e-31 Samples do not come from a normal distribution.



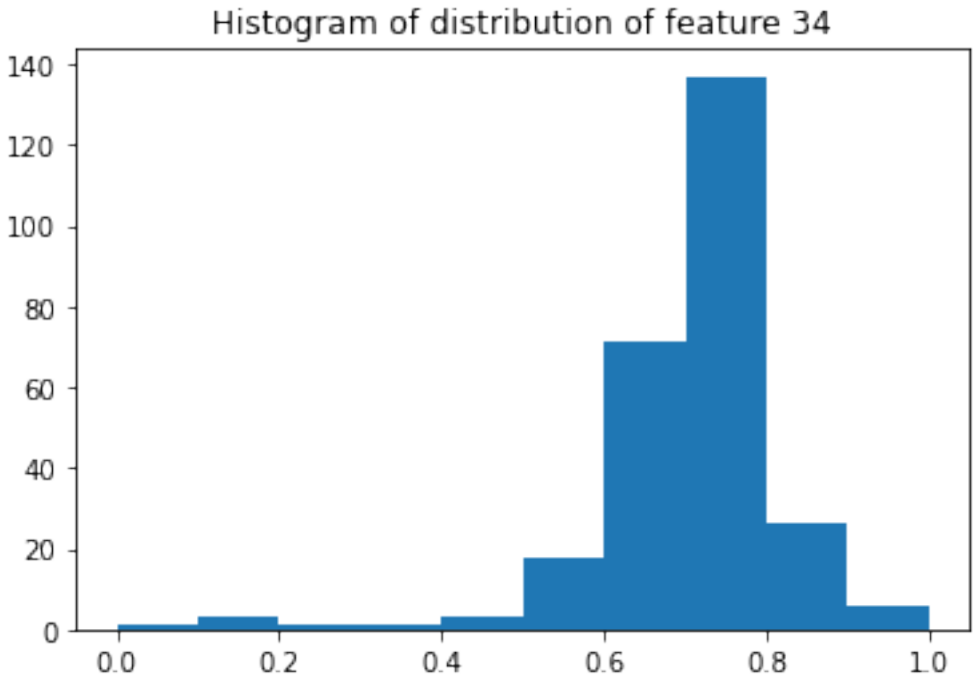
Normality test for feature 32:
P-value: 3.243030847198031e-46 Samples do not come from a normal distribution.



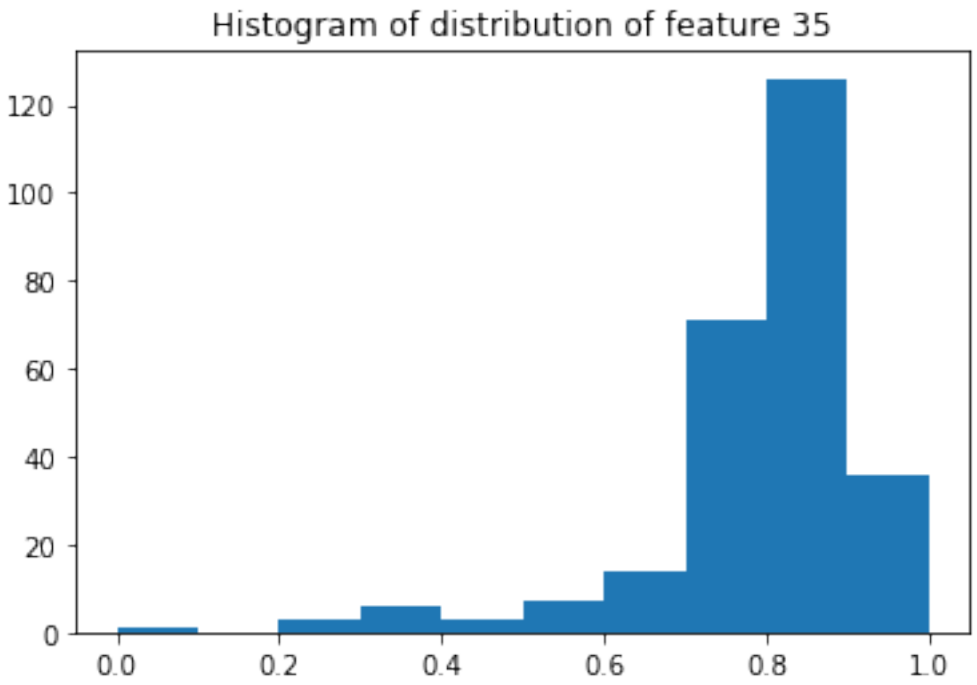

```
Normality test for feature 33:
        P-value: 2.3534425742796156e-18 Samples do not come from a normal
distribution.
```



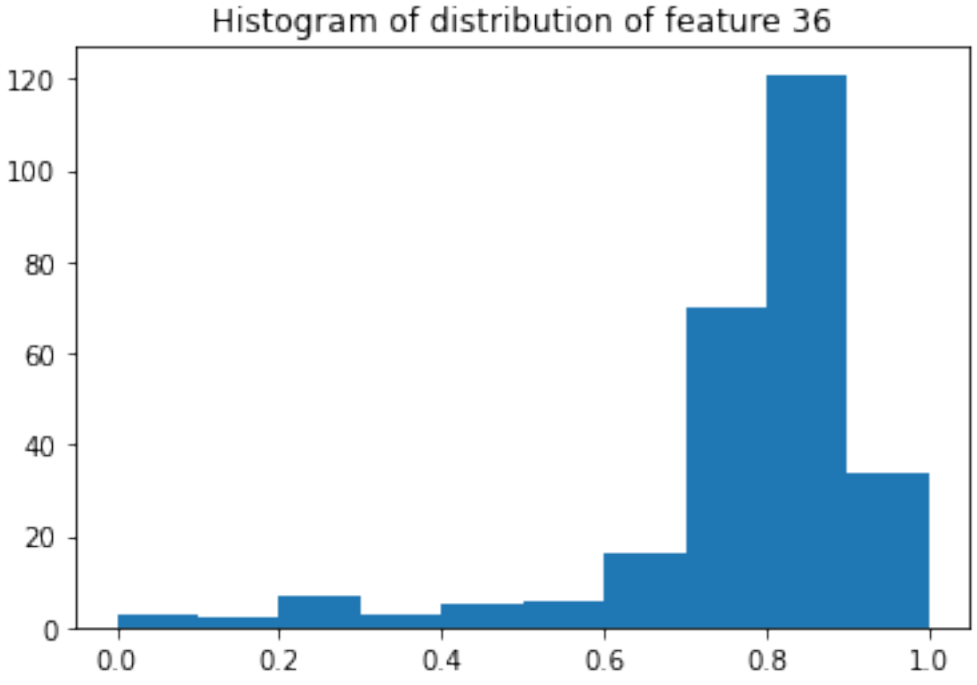
```
Normality test for feature 34:
        P-value: 7.83526865540055e-34   Samples do not come from a normal
distribution.
```



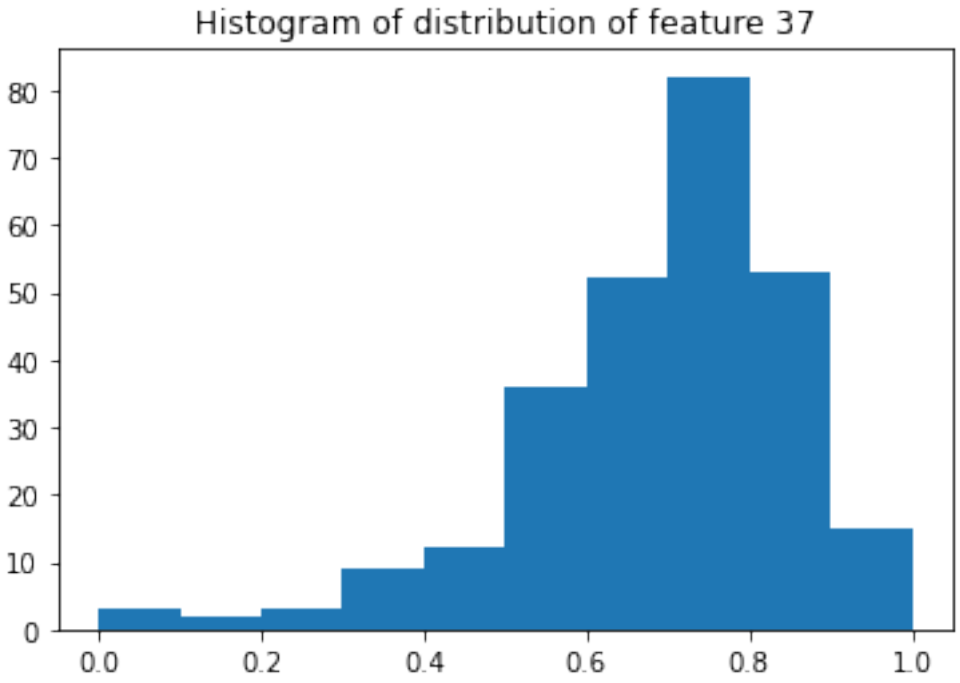

```
Normality test for feature 35:
        P-value: 1.325141100532452e-32  Samples do not come from a normal
distribution.
```



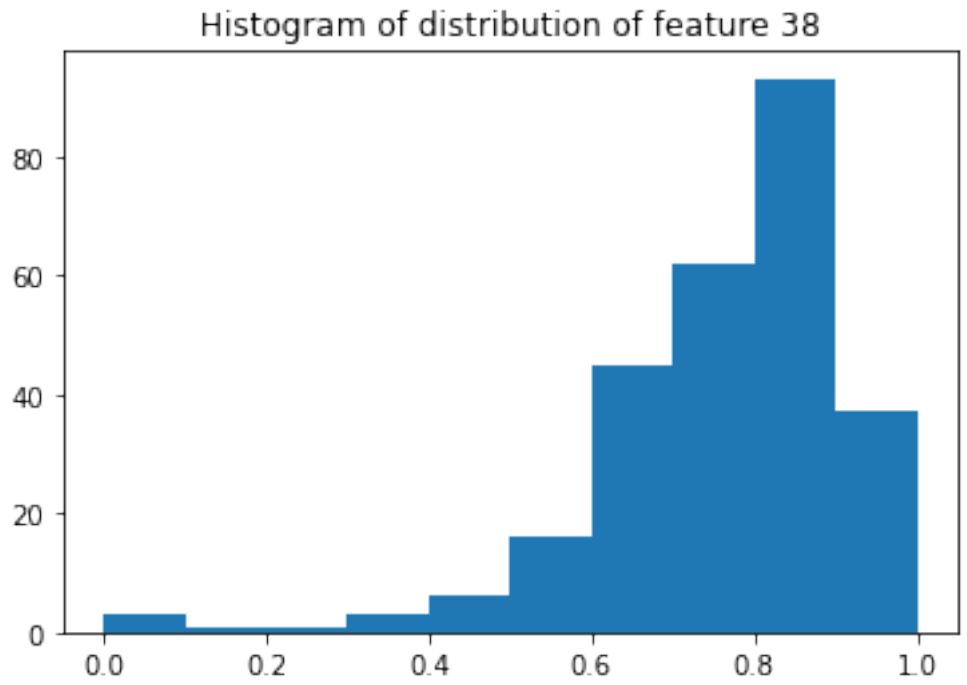

```
Normality test for feature 36:
        P-value: 3.2861666462959854e-32 Samples do not come from a normal
distribution.
```



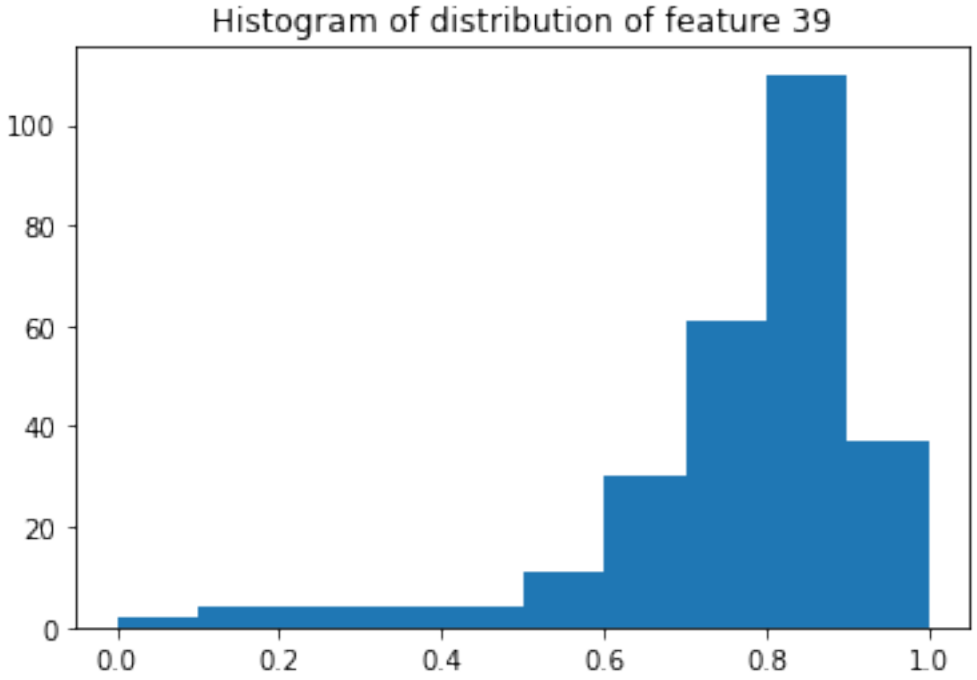
Normality test for feature 37:
P-value: 1.8215890226653356e-14 Samples do not come from a normal distribution.



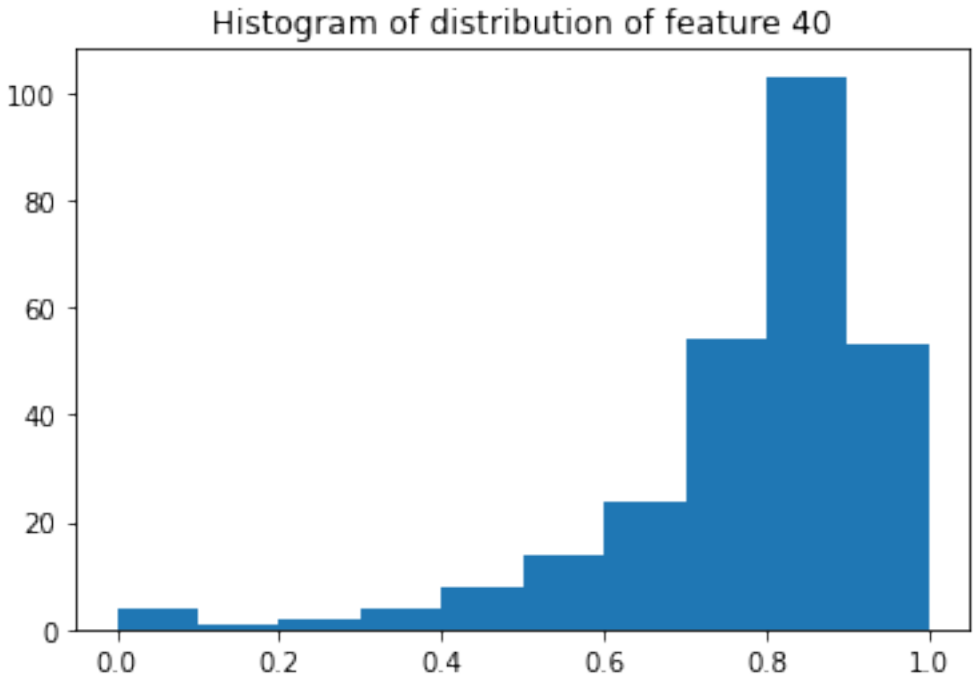
Normality test for feature 38:
P-value: 3.50797377328908e-24 Samples do not come from a normal distribution.



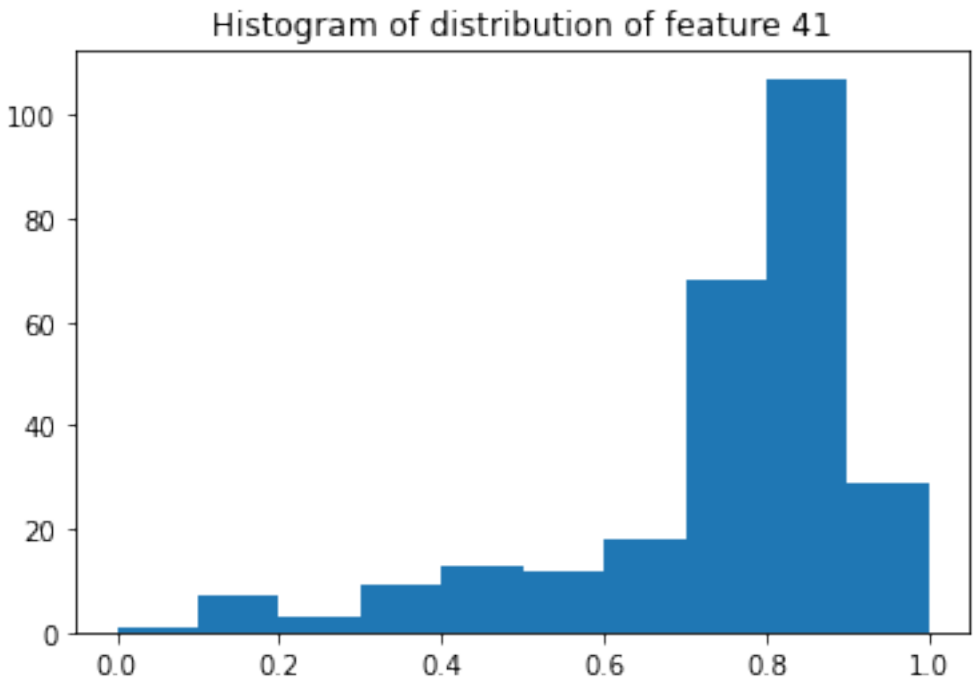

```
Normality test for feature 39:
        P-value: 1.1384125784058246e-27 Samples do not come from a normal
distribution.
```



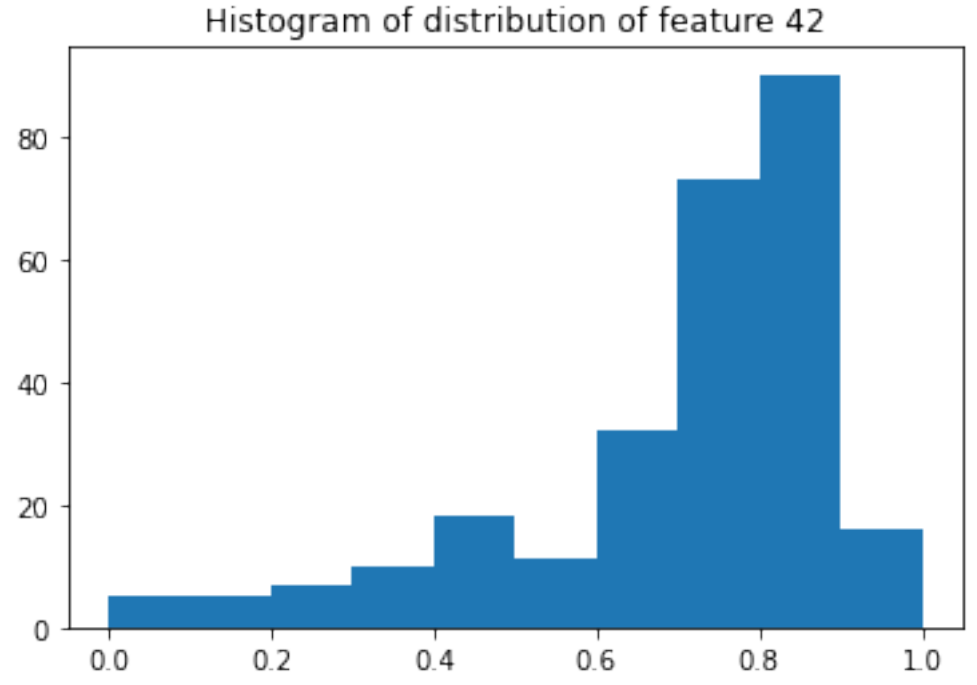
Normality test for feature 40:
P-value: 6.943205070233066e-26 Samples do not come from a normal distribution.



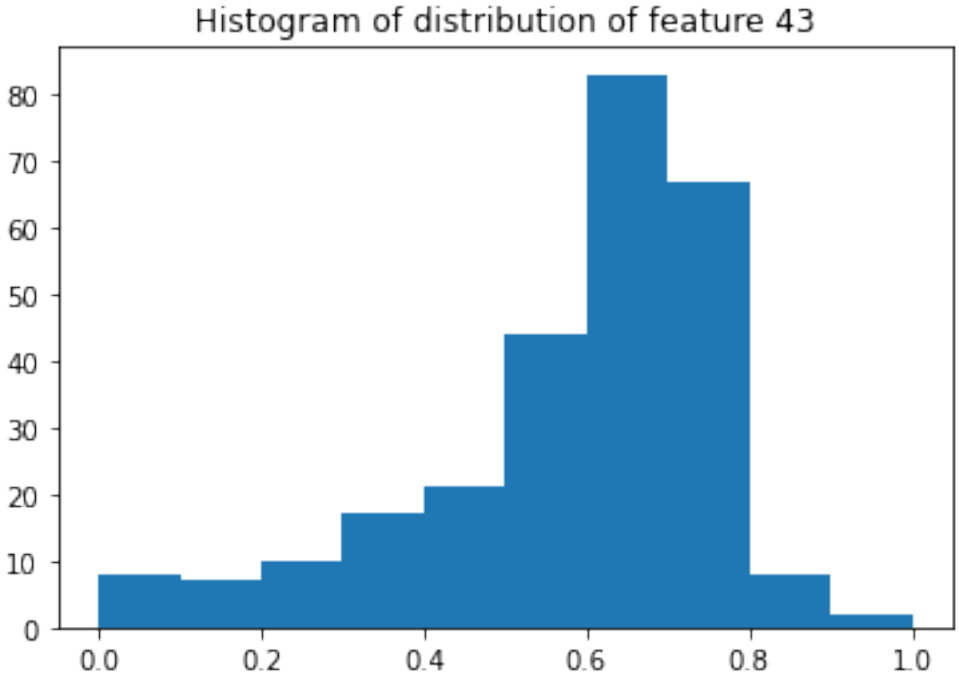
Normality test for feature 41:
P-value: 3.7092413705222476e-19 Samples do not come from a normal distribution.



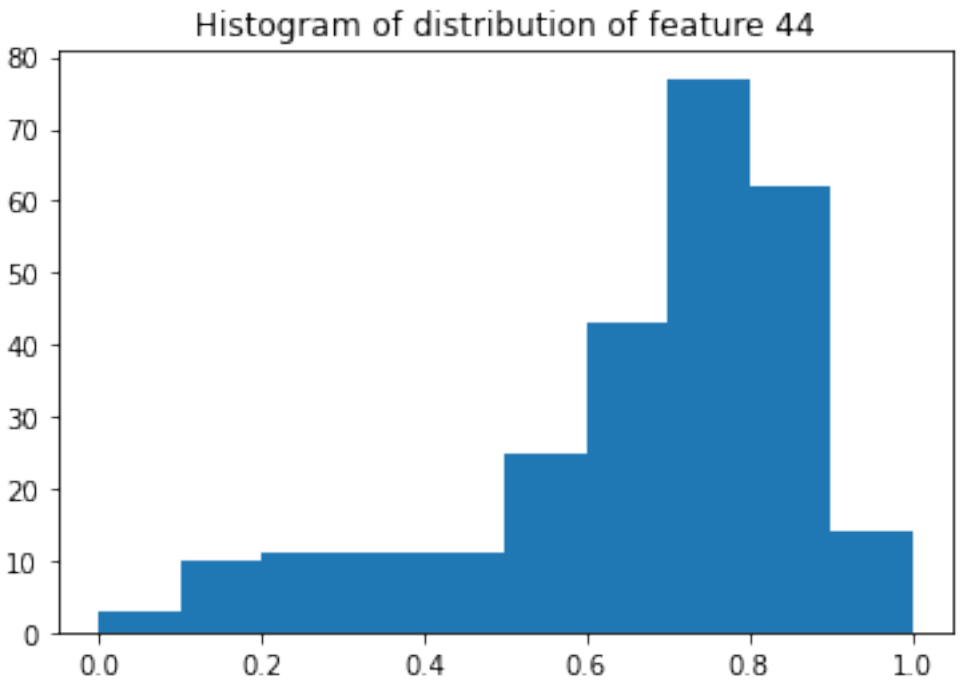

```
Normality test for feature 42:
        P-value: 2.741371386684411e-16  Samples do not come from a normal
distribution.
```



Normality test for feature 43:
P-value: 3.677381553003623e-12 Samples do not come from a normal distribution.



Normality test for feature 44:
P-value: 4.433189268261597e-12 Samples do not come from a normal distribution.



3 Perform a relief test on every feature

[5]:
```
sorted_w = relief_test(normalized)
```

```
0% 5% 10% 15% 20% 25% 30% 35% 40% 44% 49% 54% 59% 64% 69% 74% 79% 84% 89% 94%
99%

For a threshold 0.12239801227242092:
Feature 44 weight: 0.0348        Is considered irrelevant.
Feature 42 weight: 0.0337        Is considered irrelevant.
Feature 43 weight: 0.0292        Is considered irrelevant.
Feature 41 weight: 0.0272        Is considered irrelevant.
Feature 26 weight: 0.0263        Is considered irrelevant.
Feature 40 weight: 0.025         Is considered irrelevant.
Feature 25 weight: 0.0233        Is considered irrelevant.
Feature 30 weight: 0.0223        Is considered irrelevant.
Feature 15 weight: 0.0218        Is considered irrelevant.
Feature 39 weight: 0.0204        Is considered irrelevant.
Feature 16 weight: 0.0197        Is considered irrelevant.
Feature 24 weight: 0.0196        Is considered irrelevant.
Feature 6 weight: 0.0184         Is considered irrelevant.
Feature 19 weight: 0.018         Is considered irrelevant.
Feature 2 weight: 0.0167         Is considered irrelevant.
Feature 22 weight: 0.0167        Is considered irrelevant.
```

```
Feature 14 weight: 0.0164        Is considered irrelevant.
Feature 10 weight: 0.0162        Is considered irrelevant.
Feature 29 weight: 0.0161        Is considered irrelevant.
Feature 4 weight: 0.016  Is considered irrelevant.
Feature 8 weight: 0.0149         Is considered irrelevant.
Feature 13 weight: 0.0147        Is considered irrelevant.
Feature 3 weight: 0.0142         Is considered irrelevant.
Feature 21 weight: 0.0127        Is considered irrelevant.
Feature 28 weight: 0.0126        Is considered irrelevant.
Feature 33 weight: 0.0125        Is considered irrelevant.
Feature 36 weight: 0.0124        Is considered irrelevant.
Feature 5 weight: 0.0122         Is considered irrelevant.
Feature 31 weight: 0.012         Is considered irrelevant.
Feature 20 weight: 0.0114        Is considered irrelevant.
Feature 23 weight: 0.0112        Is considered irrelevant.
Feature 18 weight: 0.0111        Is considered irrelevant.
Feature 9 weight: 0.0093         Is considered irrelevant.
Feature 37 weight: 0.0092        Is considered irrelevant.
Feature 34 weight: 0.0088        Is considered irrelevant.
Feature 1 weight: 0.0081         Is considered irrelevant.
Feature 32 weight: 0.008         Is considered irrelevant.
Feature 7 weight: 0.0074         Is considered irrelevant.
Feature 17 weight: 0.0065        Is considered irrelevant.
Feature 12 weight: 0.0063        Is considered irrelevant.
Feature 35 weight: 0.0053        Is considered irrelevant.
Feature 27 weight: 0.0039        Is considered irrelevant.
Feature 11 weight: 0.0034        Is considered irrelevant.
Feature 38 weight: 0.0033        Is considered irrelevant.
```

4 Perform the Pearson Test on every feature

[6]:
```
pearson_coefs, to_compare = pearson_test(normalized)
pearson_coefs
```

[6]:

	1	2	3	4	5	6	\
1	1	0.594818	0.361351	0.361968	0.190589	0.211766	
2	0.594818	1	0.286115	0.464016	0.0728498	0.322178	
3	0.361351	0.286115	1	0.621208	0.182356	0.163781	
4	0.361968	0.464016	0.621208	1	0.116763	0.258593	
5	0.190589	0.0728498	0.182356	0.116763	1	0.673692	
6	0.211766	0.322178	0.163781	0.258593	0.673692	1	
7	0.573573	0.398424	0.287242	0.236824	0.283162	0.34198	
8	0.402308	0.53047	0.146218	0.253367	0.333614	0.611637	
9	0.502529	0.409771	0.25462	0.312069	0.219211	0.199263	
10	0.387242	0.602294	0.238437	0.361206	0.174396	0.306467	
..	…	…	…	…	…	…	
35	0.355329	0.333366	0.159536	0.153638	0.281186	0.386563	

```
36     0.291402     0.393538   0.146198  0.206242   0.240118    0.489232
37     0.256045     0.244496   0.249765  0.170886 -0.0188998  0.00785678
38     0.208062     0.450745   0.235431  0.323317 -0.0828012   0.0822074
39     0.404436     0.347584   0.286746  0.276272   0.344285    0.406786
40     0.293084      0.46494   0.236829  0.395085    0.20649    0.457176
41    0.0676561    0.0518904   0.164519   0.16444   0.479817    0.585095
42    0.0795237    0.0703992   0.130812  0.192109   0.435943    0.666987
43  5.52889e-05 -0.000144422  0.0851879   0.13421   0.370397    0.445398
44    0.0371708    0.0729498   0.104369  0.154437    0.36223    0.551609

           7         8          9          10  …        35        36  \
1    0.573573  0.402308   0.502529   0.387242  …  0.355329  0.291402
2    0.398424   0.53047   0.409771   0.602294  …  0.333366  0.393538
3    0.287242  0.146218    0.25462   0.238437  …  0.159536  0.146198
4    0.236824  0.253367   0.312069   0.361206  …  0.153638  0.206242
5    0.283162  0.333614   0.219211   0.174396  …  0.281186  0.240118
6     0.34198  0.611637   0.199263   0.306467  …  0.386563  0.489232
7           1  0.653655   0.538945   0.393628  …  0.606697   0.49992
8    0.653655         1   0.324386   0.436267  …  0.621734  0.709084
9    0.538945  0.324386          1    0.74155  …  0.519289   0.43642
10   0.393628  0.436267    0.74155          1  …  0.496886  0.564873
..        …         …          …          …  …         …         …
35   0.606697  0.621734   0.519289   0.496886  …         1  0.844275
36    0.49992  0.709084    0.43642   0.564873  …  0.844275         1
37    0.36054  0.179481   0.461254   0.339065  …  0.333128  0.290608
38   0.222742  0.285968   0.405574   0.526591  …  0.262324   0.40454
39   0.602381  0.589192   0.480088   0.421423  …  0.656742  0.641477
40   0.439435  0.609377    0.35461   0.467159  …  0.575375  0.662363
41    0.19694  0.369545   0.193161   0.173855  …  0.408514  0.461167
42   0.195516  0.396101   0.136932   0.157219  …   0.34472  0.431582
43  0.0678882  0.250077  0.0817281  0.0838936  …  0.237279  0.300401
44   0.131094  0.318606  0.0885651   0.134039  …      0.27  0.344596

           37         38        39        40         41         42  \
1    0.256045   0.208062  0.404436  0.293084  0.0676561  0.0795237
2    0.244496   0.450745  0.347584   0.46494  0.0518904  0.0703992
3    0.249765   0.235431  0.286746  0.236829   0.164519   0.130812
4    0.170886   0.323317  0.276272  0.395085    0.16444   0.192109
5   -0.0188998 -0.0828012  0.344285   0.20649   0.479817   0.435943
6   0.00785678  0.0822074  0.406786  0.457176   0.585095   0.666987
7     0.36054   0.222742  0.602381  0.439435    0.19694   0.195516
8    0.179481   0.285968  0.589192  0.609377   0.369545   0.396101
9    0.461254   0.405574  0.480088   0.35461   0.193161   0.136932
10   0.339065   0.526591  0.421423  0.467159   0.173855   0.157219
..        …          …         …         …          …          …
35   0.333128   0.262324  0.656742  0.575375   0.408514    0.34472
36   0.290608    0.40454  0.641477  0.662363   0.461167   0.431582
```

```
37           1   0.729553  0.426019  0.302821   0.188436  0.0734496
38    0.729553          1  0.397702  0.480046   0.154074  0.0568128
39    0.426019   0.397702         1   0.81387   0.575576   0.446725
40    0.302821   0.480046   0.81387         1   0.455036   0.450495
41    0.188436   0.154074  0.575576  0.455036          1   0.829906
42   0.0734496  0.0568128  0.446725  0.450495   0.829906          1
43    0.158684   0.127785  0.442958  0.380601   0.849442   0.780719
44    0.103282  0.0966623  0.421608  0.396498   0.801493   0.877703

              43         44
1   5.52889e-05  0.0371708
2  -0.000144422  0.0729498
3     0.0851879   0.104369
4       0.13421   0.154437
5      0.370397    0.36223
6      0.445398   0.551609
7     0.0678882   0.131094
8      0.250077   0.318606
9     0.0817281  0.0885651
10    0.0838936   0.134039
..          …         …
35     0.237279       0.27
36     0.300401   0.344596
37     0.158684   0.103282
38     0.127785  0.0966623
39     0.442958   0.421608
40     0.380601   0.396498
41     0.849442   0.801493
42     0.780719   0.877703
43            1   0.883061
44     0.883061          1

[44 rows x 44 columns]
```

5 Calculate a ponderated irrelevance score for each feature

The score of feature i is a combination between the sum of the absolute value of all Pearson coefficients feature i and its relief score:

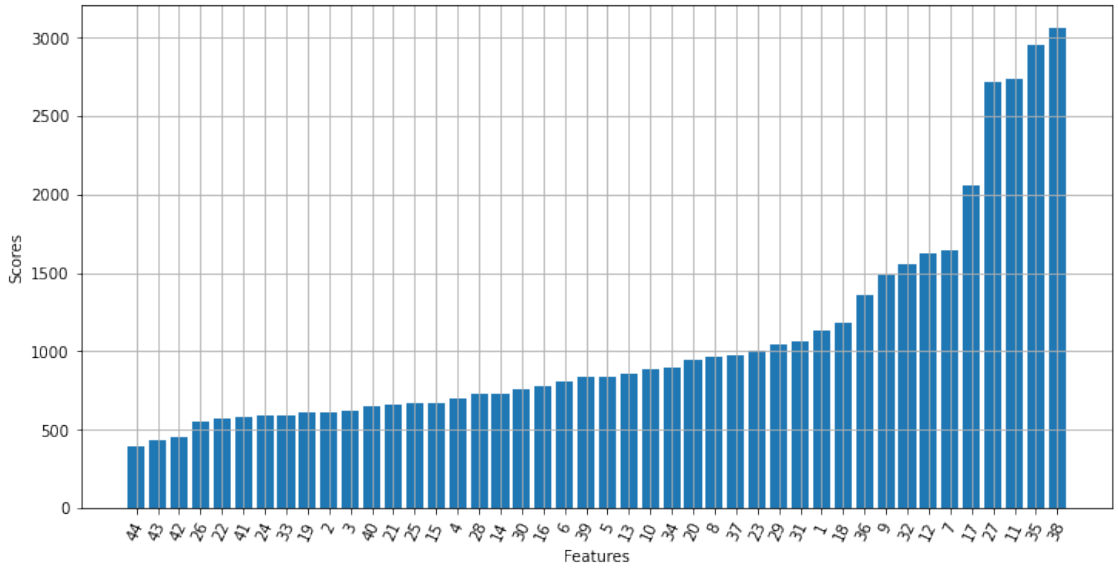
$score_i = \frac{Pearson_{total_i}}{relief_i}$

I consider this function because of the fact that Pearson coefficients are closer to 0 for uncorrelated features, while the relief scores are closer to 1 for statistically relevant features.

This means that for a higher $score_i$, feature i is more likely to be statistically irrelevant and/or more strongly correlated to other features.

[ ]:
```
scores = score_features(normalized, sorted_w, to_compare)
```

[9]:
```
x = [str(int(s[0])) for s in scores]
y = [s[1] for s in scores]
plt.figure(figsize=(12, 6))
plt.bar(x, y)
plt.grid()
plt.xlabel('Features')
plt.xticks(rotation=65)
plt.ylabel('Scores')
plt.show()
```



6 Select features to keep

By observing the plot above, I notice a sharp jump in scores for the last 10 of them. I remove these features from the set and keep the 38 features with the lowest scores.

[8]:
```
selected_features = [s[0] for s in scores[:-10]]
len(selected_features)
```

[8]: 34