МОСКОВСКИЙ ГОСУДАРСТВЕННЫЙ ТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ
им. Н.Э. Баумана

Кафедра «Систем обработки информации и управления»

ОТЧЕТ

Лабораторная работа №3
по курсу «Методы машинного обучения»

Тема: «Обработка пропусков в данных, кодирование категориальных признаков, масштабирование данных»

ИСПОЛНИТЕЛЬ: Ромичева Е.В.

группа ИУ5-22М ____________________
подпись

"__"_________2019 г.

ПРЕПОДАВАТЕЛЬ: Гапанюк Ю.Е.

____
подпись

"__"_________2019 г.

Москва - 2019

Цель работы

Цель лабораторной работы: изучение способов предварительной обработки данных для дальнейшего формирования моделей.

Задание

1. Выбрать набор данных (датасет), содержащий категориальные признаки и пропуски в данных. Для выполнения следующих пунктов можно использовать несколько различных наборов данных (один для обработки пропусков, другой для категориальных признаков и т.д.)
2. Для выбранного датасета (датасетов) на основе материалов лекции решить следующие задачи:

- обработку пропусков в данных (lab3_1);
- кодирование категориальных признаков (lab3_2);
- масштабирование данных (lab3_3);

lab3_1

March 6, 2019

```
In [1]: import numpy as np
        import pandas as pd
        import seaborn as sns
        import matplotlib.pyplot as plt
        %matplotlib inline
        sns.set(style="ticks")

In [22]: data = pd.read_csv(r'shanghaiData.csv', sep=",")
         data.head()

Out[22]:   world_rank                            university_name national_rank  \
         0          1                         Harvard University             1
         1          2                    University of Cambridge             1
         2          3                        Stanford University             2
         3          4         University of California, Berkeley             3
         4          5  Massachusetts Institute of Technology (MIT)             4

           total_score  alumni  award   hici     ns    pub   pcp  year
         0       100.0   100.0  100.0  100.0  100.0  100.0  72.4  2005
         1        73.6    99.8   93.4   53.3   56.6   70.9  66.9  2005
         2        73.4    41.1   72.2   88.5   70.9   72.3  65.0  2005
         3        72.8    71.8   76.0   69.4   73.9   72.2  52.7  2005
         4        70.1    74.0   80.6   66.7   65.8   64.3  53.0  2005

In [23]: for col in data.columns:
             #    -
             temp_null_count = data[data[col].isnull()].shape[0]
             print('{} - {}'.format(col, temp_null_count))

world_rank - 0
university_name - 1
national_rank - 1
total_score - 3796
alumni - 1
award - 2
hici - 2
ns - 22
pub - 2
```

```
pcp - 2
year - 0
```

In [24]:
```
data.shape
```

Out[24]:
```
(4897, 11)
```

In [25]:
```
data.dtypes
```

Out[25]:
```
world_rank          object
university_name     object
national_rank       object
total_score        float64
alumni             float64
award              float64
hici               float64
ns                 float64
pub                float64
pcp                float64
year                 int64
dtype: object
```

In [26]:
```
data.isnull().sum()
```

Out[26]:
```
world_rank            0
university_name       1
national_rank         1
total_score        3796
alumni                1
award                 2
hici                  2
ns                   22
pub                   2
pcp                   2
year                  0
dtype: int64
```

1.

1.1. -

In [7]:
```
#  ,
data1 = data.dropna(axis=1, how='any')
(data.shape, data1.shape)
```

Out[7]:
```
((2603, 14), (2603, 10))
```

In [8]:
```
#  ,
data2 = data.dropna(axis=0, how='any')
(data.shape, data2.shape)
```

```
Out[8]: ((2603, 14), (2362, 14))

In [9]: #
        #     ,
        data3 = data.fillna(0)
        data3.head()

Out[9]:   world_rank                        university_name                   country  \
        0          1                     Harvard University  United States of America
        1          2     California Institute of Technology  United States of America
        2          3  Massachusetts Institute of Technology  United States of America
        3          4                    Stanford University  United States of America
        4          5                   Princeton University  United States of America

          teaching international  research  citations income total_score  \
        0     99.7          72.4      98.7       98.8   34.5        96.1
        1     97.7          54.6      98.0       99.9   83.7        96.0
        2     97.8          82.3      91.4       99.9   87.5        95.6
        3     98.3          29.5      98.1       99.2   64.3        94.3
        4     90.9          70.3      95.4       99.9      -        94.2

          num_students  student_staff_ratio international_students female_male_ratio  \
        0       20,152                  8.9                    25%                 0
        1        2,243                  6.9                    27%           33 : 67
        2       11,074                  9.0                    33%           37 : 63
        3       15,596                  7.8                    22%           42 : 58
        4        7,929                  8.4                    27%           45 : 55

          year
        0 2011
        1 2011
        2 2011
        3 2011
        4 2011

In [27]: total_count = data.shape[0]
         print(' : {}'.format(total_count))

 : 4897
```

1.2. “ ” - (imputation)

1.2.1.

```
In [28]: #
         #
         num_cols = []
         for col in data.columns:
             #
```

```
        temp_null_count = data[data[col].isnull()].shape[0]
        dt = str(data[col].dtype)
        if temp_null_count>0 and (dt=='float64' or dt=='int64'):
            num_cols.append(col)
            temp_perc = round((temp_null_count / total_count) * 100.0, 2)
            print(' {}.   {}.    {}, {}%.'.format(col, dt, temp_null_count, temp_perc))
```

```
 total_score.   float64.    3796, 77.52%.
 alumni.   float64.    1, 0.02%.
 award.   float64.    2, 0.04%.
 hici.   float64.    2, 0.04%.
 ns.   float64.    22, 0.45%.
 pub.   float64.    2, 0.04%.
 pcp.   float64.    2, 0.04%.
```

In [29]:
```
data_num = data[num_cols]
data_num
```

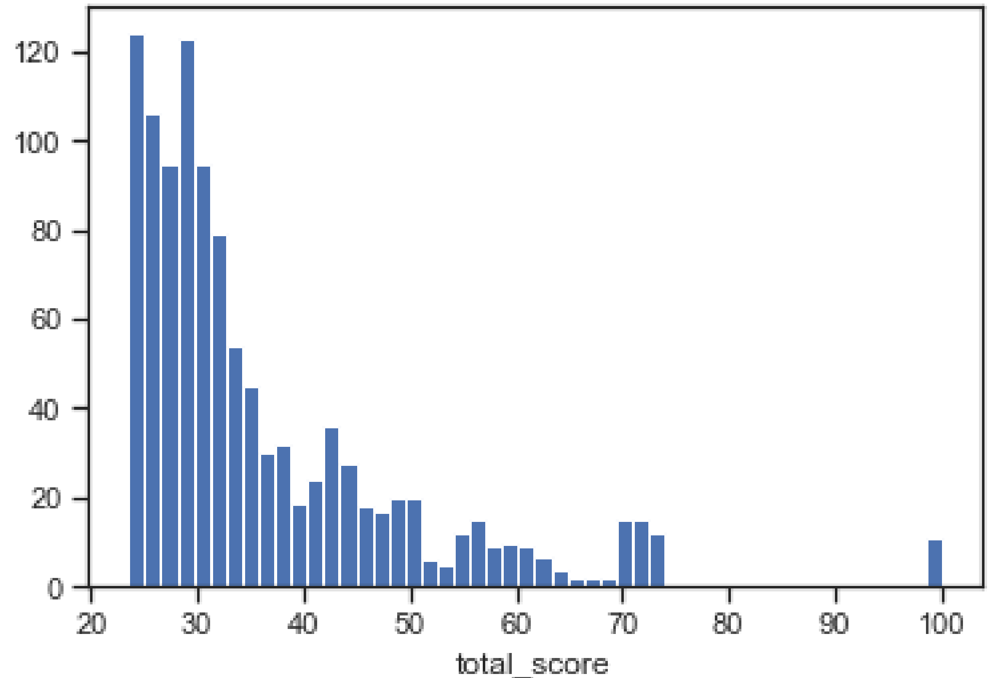
Out[29]:

	total_score	alumni	award	hici	ns	pub	pcp
0	100.0	100.0	100.0	100.0	100.0	100.0	72.4
1	73.6	99.8	93.4	53.3	56.6	70.9	66.9
2	73.4	41.1	72.2	88.5	70.9	72.3	65.0
3	72.8	71.8	76.0	69.4	73.9	72.2	52.7
4	70.1	74.0	80.6	66.7	65.8	64.3	53.0
5	67.1	59.2	68.6	59.8	65.8	52.5	100.0
6	62.3	79.4	60.6	56.1	54.2	69.5	45.4
7	60.9	63.4	76.8	60.9	48.7	48.5	59.1
8	60.1	75.6	81.9	50.3	44.7	56.4	42.2
9	59.7	64.3	59.1	48.4	55.6	68.4	53.2
10	56.9	52.1	44.5	60.3	57.2	63.9	49.3
11	54.6	46.5	52.4	55.0	48.8	66.3	39.8
12	51.0	17.7	34.7	59.8	56.5	64.5	46.6
13	50.6	27.3	32.8	56.7	50.1	75.6	34.3
14	50.2	35.5	35.1	56.7	42.9	71.8	39.1
15	49.2	43.0	36.3	52.1	46.3	68.7	29.0
16	48.4	28.8	32.4	53.9	47.1	73.8	27.2
17	47.8	0.0	37.6	55.6	57.9	58.8	45.2
18	46.9	51.4	28.3	41.6	52.2	67.7	24.9
19	46.7	36.0	14.4	38.5	52.1	86.5	34.7
20	44.9	43.0	0.0	61.9	43.0	76.5	30.9
21	43.8	39.7	34.1	34.2	37.0	72.3	31.1
22	43.7	20.8	38.1	40.8	38.2	64.6	40.3
23	43.1	28.1	19.7	39.3	38.9	76.7	41.9
24	42.8	41.6	37.4	44.4	34.1	58.0	26.0
25	42.6	30.7	32.9	37.7	41.5	60.5	38.8
26	41.7	40.2	37.0	35.1	41.1	43.4	52.4
27	40.7	25.1	26.6	38.5	46.5	53.9	39.9

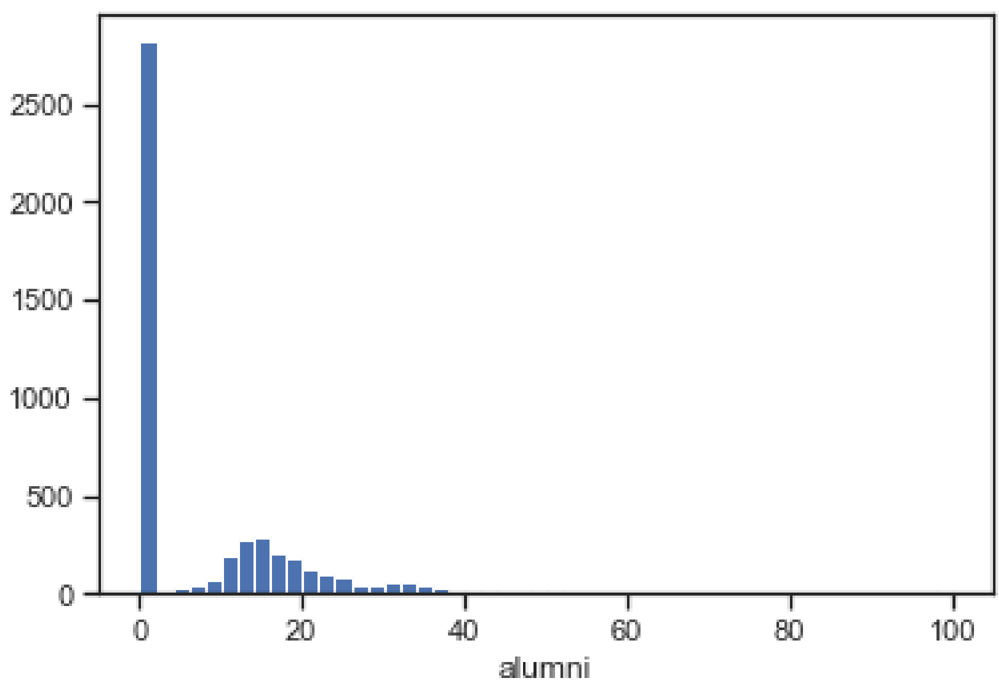
```
28          38.8   33.8   25.0   43.0   35.3   55.4   26.3
29          38.2   22.6   59.8   28.3   44.1   24.0   35.9
...          ...    ...    ...    ...    ...    ...    ...
4867         NaN    0.0    0.0    0.0    9.3   34.0   17.1
4868         NaN    0.0    0.0    3.6   10.3   26.7   14.1
4869         NaN    0.0    0.0   12.1   11.4   22.2   13.5
4870         NaN    0.0    0.0    3.6    8.4   32.8   16.6
4871         NaN    0.0    0.0    0.0    7.7   35.1   14.2
4872         NaN    0.0    0.0   17.4    6.5   17.8   17.3
4873         NaN    0.0    0.0    5.0    5.6   30.9   21.4
4874         NaN    0.0    0.0    3.6   16.3   26.2   12.7
4875         NaN    0.0    0.0    5.1   10.0   28.0   14.0
4876         NaN    0.0    0.0    6.3    6.6   28.2   14.8
4877         NaN    0.0    0.0   13.6    2.1   26.7   19.6
4878         NaN    0.0    0.0    5.0    5.7   30.7   19.7
4879         NaN    0.0    0.0    3.6    7.5   29.3   18.5
4880         NaN    0.0    0.0    0.0    9.8   33.3   16.8
4881         NaN    0.0    0.0    3.6   13.9   27.7   15.2
4882         NaN    0.0    0.0    3.6    9.2   28.1   11.2
4883         NaN    0.0    0.0   15.2    6.1   21.1   16.0
4884         NaN    0.0    0.0    0.0    8.8   33.7   19.2
4885         NaN    0.0    0.0    8.6    8.4   25.0   13.5
4886         NaN    0.0    0.0    7.1    6.1   31.1   13.2
4887         NaN    0.0    0.0    7.1    3.3   30.6   15.7
4888         NaN    0.0    0.0    0.0    7.5   33.7   11.3
4889         NaN    0.0    0.0    8.6    4.9   27.0   18.0
4890         NaN    0.0   13.3    3.6    3.4   21.8   12.8
4891         NaN    0.0    0.0    3.6    7.1   36.1   13.5
4892         NaN    0.0    0.0    5.0   10.9   25.1   20.1
4893         NaN    0.0    0.0    7.6    5.1   33.3   13.1
4894         NaN   13.6    0.0    3.6   10.8   25.1   15.5
4895         NaN    0.0    0.0    0.0   12.2   28.8   22.9
4896         NaN    0.0    0.0   14.9    7.5   25.0   11.9

[4897 rows x 7 columns]
```

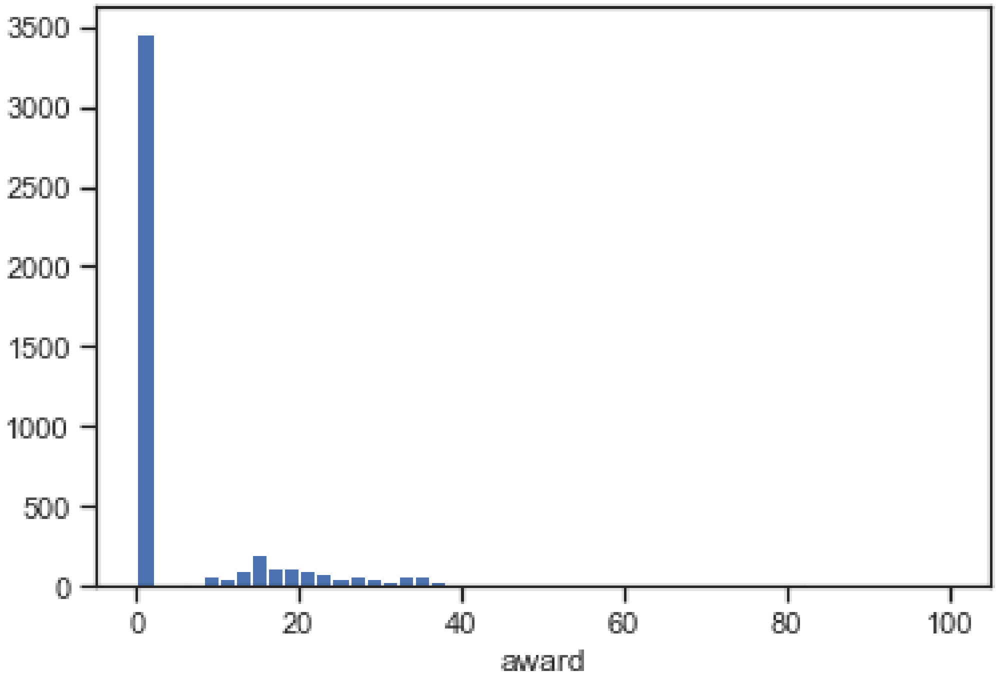
In [30]:
```
#
for col in data_num:
    plt.hist(data[col], 50)
    plt.xlabel(col)
    plt.show()
```

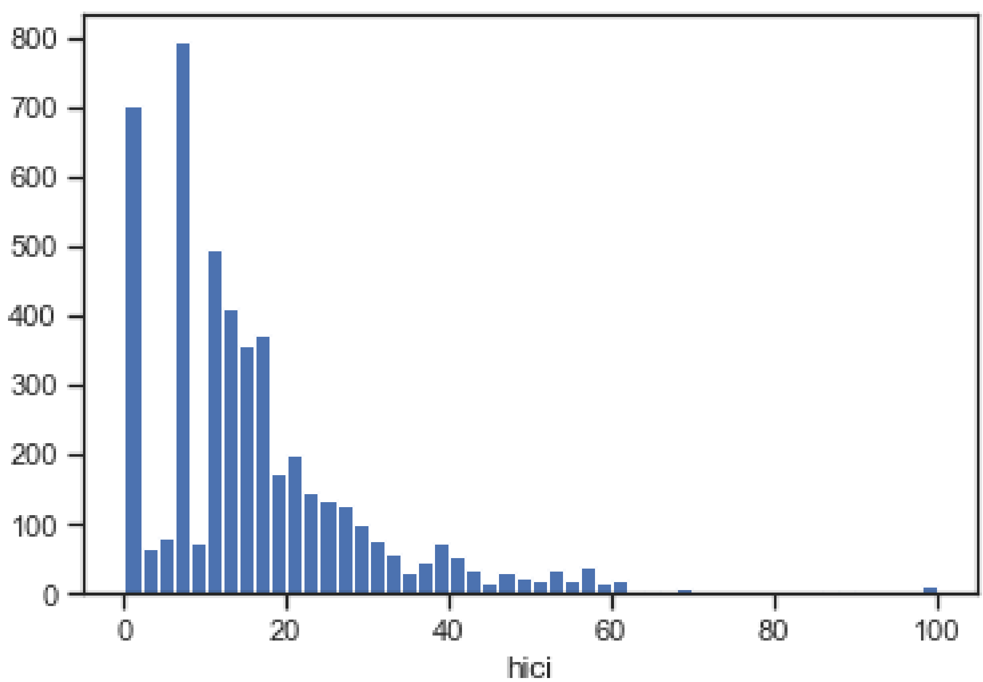


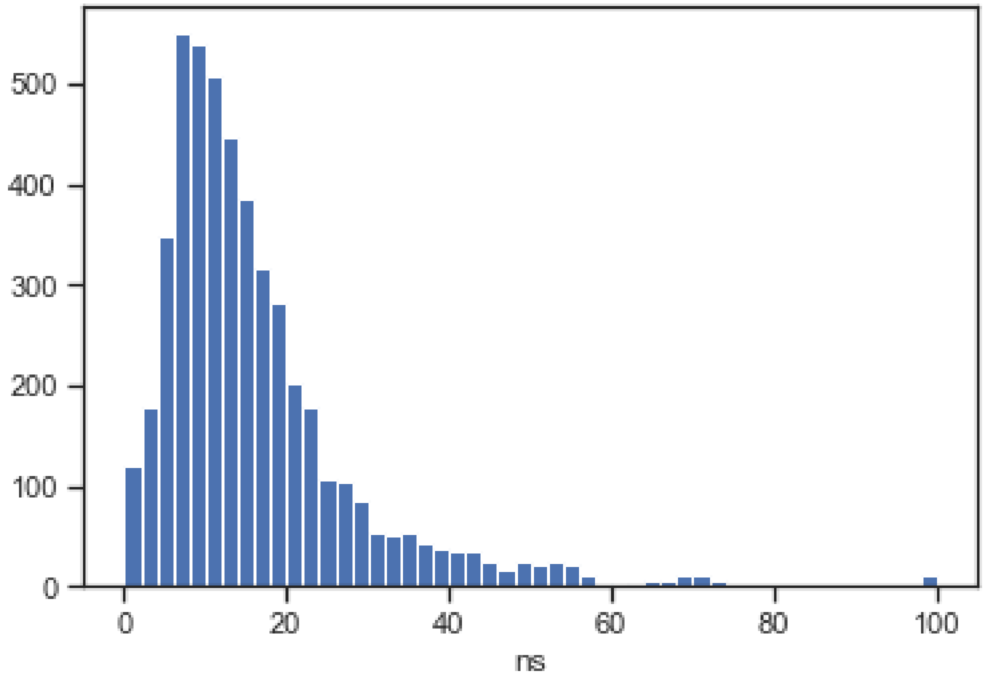
120
100
80
60
40
20
0
20
30
40
50
60
70
80
90
100
total_score



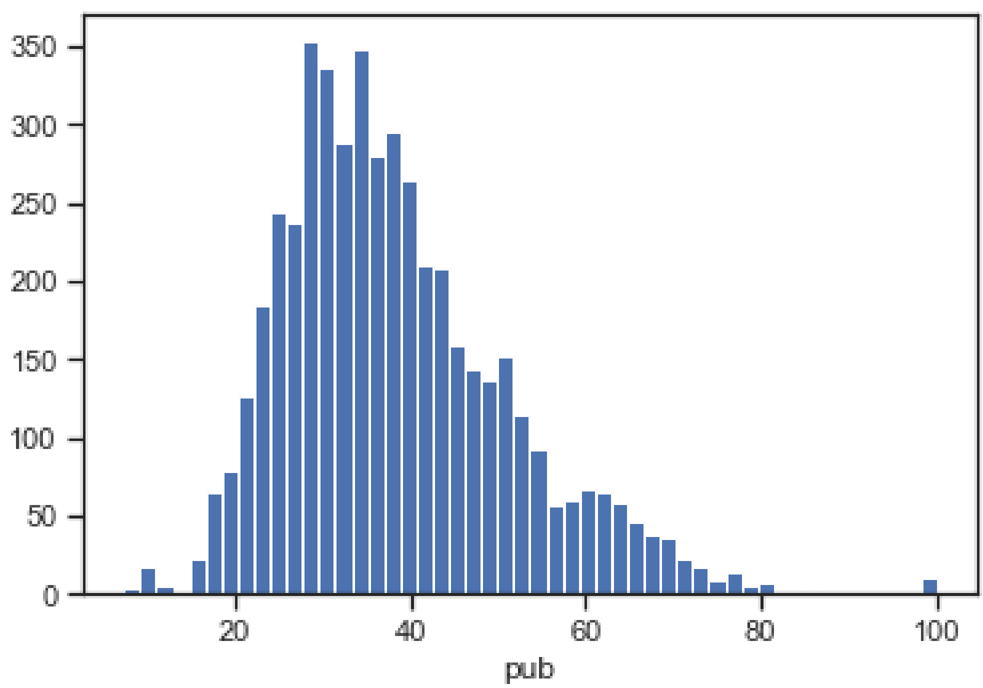
2500
2000
1500
1000
500
0
0
20
40
60
80
100
alumni


3500
3000
2500
2000
1500
1000
500
0
0
20
40
60
80
100
award

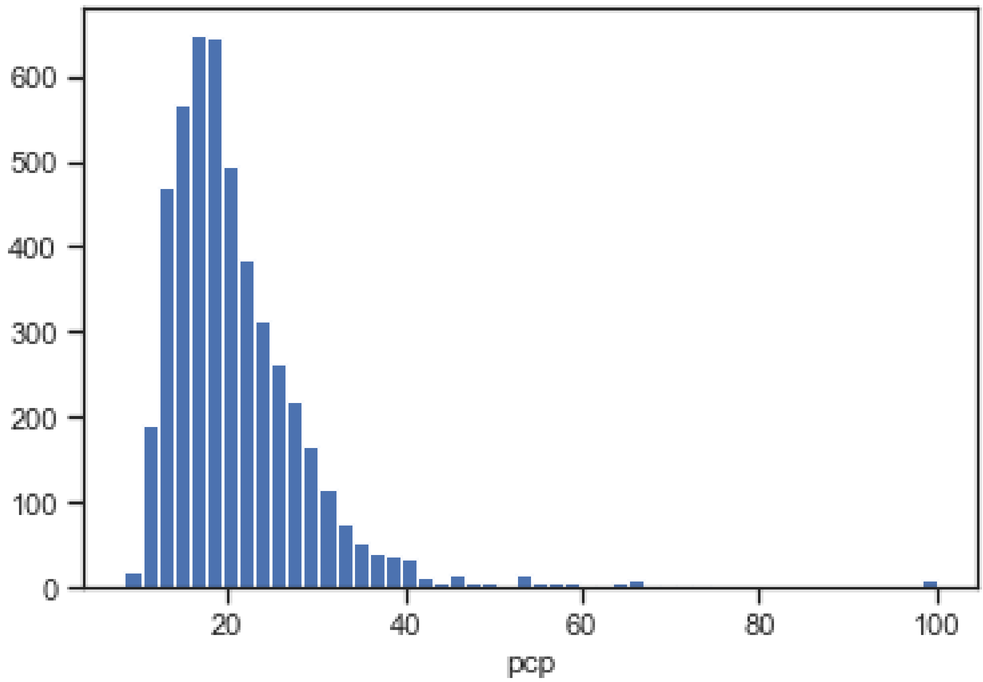

800
700
600
500
400
300
200
100
0
0
20
40
60
80
100
hici



500
400
300
200
100
0
0
20
40
60
80
100
ns



350
300
250
200
150
100
50
0
20
40
60
80
100
pub



In [31]:
```
#       MasVnrArea
data[data['total_score'].isnull()]
```

Out[31]:
```
      world_rank                            university_name national_rank  \
100      101-152                          Aarhus University             2
101      101-152           Arizona State University - Tempe         54-71
102      101-152                  Baylor College of Medicine         54-71
103      101-152              Catholic University of Leuven           1-4
104      101-152             Catholic University of Louvain           1-4
105      101-152                           College of France             5
106      101-152                          Dartmouth College         54-71
107      101-152                           Emory University         54-71
108      101-152            Georgia Institute of Technology         54-71
109      101-152                            Ghent University           1-4
110      101-152                        Hokkaido University           6-9
111      101-152                          Kyushu University           6-9
112      101-152                        Mayo Medical School         54-71
113      101-152                          Nagoya University           6-9
114      101-152           National University of Singapore             1
115      101-152   North Carolina State University - Raleigh         54-71
116      101-152                     Oregon State University         54-71
117      101-152                   Seoul National University             1
118      101-152 State University of New York at Stony Brook         54-71
119      101-152       Technion-Israel Institute of Technology           2-4
```

120	101-152	Tel Aviv University	2-4
121	101-152	The University of Georgia	54-71
122	101-152	The University of Glasgow	12-15
123	101-152	The University of Queensland	3-4
124	101-152	University Libre Bruxelles	1-4
125	101-152	University of Alberta	5
126	101-152	University of Amsterdam	3-4
127	101-152	University of Bonn	6-11
128	101-152	University of California, Riverside	54-71
129	101-152	University of California, Santa Cruz	54-71
...	...	...	...
4867	401-500	University of Jena	29-39
4868	401-500	University of Jyvaskyla	4-6
4869	401-500	University of Konstanz	29-39
4870	401-500	University of KwaZulu-Natal	3-4
4871	401-500	University of Ljubljana	1
4872	401-500	University of Maryland, Baltimore County	126-146
4873	401-500	University of Milan - Bicocca	11-20
4874	401-500	University of Nice Sophia Antipolis	19-22
4875	401-500	University of Oklahoma - Norman	126-146
4876	401-500	University of Palermo	11-20
4877	401-500	University of Parma	11-20
4878	401-500	University of Pavia	11-20
4879	401-500	University of Perugia	11-20
4880	401-500	University of Quebec	19-20
4881	401-500	University of Regensburg	29-39
4882	401-500	University of Rennes 1	19-22
4883	401-500	University of Rhode Island	126-146
4884	401-500	University of Roma - Tor Vergata	11-20
4885	401-500	University of Rostock	29-39
4886	401-500	University of Santiago Compostela	9-13
4887	401-500	University of Science, Malaysia	2
4888	401-500	University of Seville	9-13
4889	401-500	University of Surrey	34-37
4890	401-500	University of Szeged	1-2
4891	401-500	University of the Basque Country	9-13
4892	401-500	University of Trieste	11-20
4893	401-500	University of Zaragoza	9-13
4894	401-500	Utah State University	126-146
4895	401-500	Vienna University of Technology	4-6
4896	401-500	Wake Forest University	126-146

	total_score	alumni	award	hici	ns	pub	pcp	year
100	NaN	15.4	19.3	7.9	22.3	41.6	22.4	2005
101	NaN	0.0	14.4	20.8	26.3	41.9	17.5	2005
102	NaN	0.0	0.0	17.6	34.5	44.0	24.9	2005
103	NaN	0.0	0.0	19.2	16.0	48.7	23.1	2005
104	NaN	14.0	13.9	13.6	8.3	44.7	26.9	2005

105	NaN	15.4	37.4	11.1	11.7	16.9	19.3	2005
106	NaN	24.3	0.0	20.8	22.0	33.0	29.1	2005
107	NaN	0.0	0.0	28.3	19.0	48.4	21.6	2005
108	NaN	16.6	0.0	23.6	19.0	43.9	25.8	2005
109	NaN	8.9	15.8	15.7	9.1	48.8	27.2	2005
110	NaN	0.0	0.0	15.7	14.1	53.8	21.5	2005
111	NaN	0.0	0.0	13.6	21.3	52.8	21.6	2005
112	NaN	0.0	0.0	27.2	6.2	50.2	24.4	2005
113	NaN	0.0	14.4	15.7	20.5	52.3	25.1	2005
114	NaN	0.0	0.0	15.7	13.8	56.7	25.7	2005
115	NaN	0.0	0.0	29.4	17.8	44.3	19.0	2005
116	NaN	15.4	0.0	24.8	25.7	36.6	27.1	2005
117	NaN	0.0	0.0	7.9	14.6	61.2	26.9	2005
118	NaN	0.0	0.0	17.6	31.4	40.7	20.5	2005
119	NaN	18.8	23.5	13.6	14.1	42.1	22.8	2005
120	NaN	0.0	0.0	24.8	20.1	54.7	26.9	2005
121	NaN	0.0	0.0	28.3	21.1	46.4	18.4	2005
122	NaN	10.9	0.0	19.2	17.1	44.4	22.1	2005
123	NaN	16.6	0.0	7.9	19.9	50.1	18.9	2005
124	NaN	28.1	19.3	0.0	12.8	37.8	29.1	2005
125	NaN	15.4	0.0	17.6	17.4	55.1	26.1	2005
126	NaN	8.9	0.0	19.2	22.2	50.1	23.1	2005
127	NaN	19.8	20.4	15.7	11.3	41.3	22.0	2005
128	NaN	0.0	0.0	28.3	25.9	36.5	27.0	2005
129	NaN	0.0	0.0	28.3	28.5	31.1	29.2	2005
...	...	...	...	...	...	...	...	...
4867	NaN	0.0	0.0	0.0	9.3	34.0	17.1	2015
4868	NaN	0.0	0.0	3.6	10.3	26.7	14.1	2015
4869	NaN	0.0	0.0	12.1	11.4	22.2	13.5	2015
4870	NaN	0.0	0.0	3.6	8.4	32.8	16.6	2015
4871	NaN	0.0	0.0	0.0	7.7	35.1	14.2	2015
4872	NaN	0.0	0.0	17.4	6.5	17.8	17.3	2015
4873	NaN	0.0	0.0	5.0	5.6	30.9	21.4	2015
4874	NaN	0.0	0.0	3.6	16.3	26.2	12.7	2015
4875	NaN	0.0	0.0	5.1	10.0	28.0	14.0	2015
4876	NaN	0.0	0.0	6.3	6.6	28.2	14.8	2015
4877	NaN	0.0	0.0	13.6	2.1	26.7	19.6	2015
4878	NaN	0.0	0.0	5.0	5.7	30.7	19.7	2015
4879	NaN	0.0	0.0	3.6	7.5	29.3	18.5	2015
4880	NaN	0.0	0.0	0.0	9.8	33.3	16.8	2015
4881	NaN	0.0	0.0	3.6	13.9	27.7	15.2	2015
4882	NaN	0.0	0.0	3.6	9.2	28.1	11.2	2015
4883	NaN	0.0	0.0	15.2	6.1	21.1	16.0	2015
4884	NaN	0.0	0.0	0.0	8.8	33.7	19.2	2015
4885	NaN	0.0	0.0	8.6	8.4	25.0	13.5	2015
4886	NaN	0.0	0.0	7.1	6.1	31.1	13.2	2015
4887	NaN	0.0	0.0	7.1	3.3	30.6	15.7	2015
4888	NaN	0.0	0.0	0.0	7.5	33.7	11.3	2015

```
4889          NaN     0.0    0.0   8.6   4.9  27.0  18.0  2015
4890          NaN     0.0   13.3   3.6   3.4  21.8  12.8  2015
4891          NaN     0.0    0.0   3.6   7.1  36.1  13.5  2015
4892          NaN     0.0    0.0   5.0  10.9  25.1  20.1  2015
4893          NaN     0.0    0.0   7.6   5.1  33.3  13.1  2015
4894          NaN    13.6    0.0   3.6  10.8  25.1  15.5  2015
4895          NaN     0.0    0.0   0.0  12.2  28.8  22.9  2015
4896          NaN     0.0    0.0  14.9   7.5  25.0  11.9  2015

[3796 rows x 11 columns]
```

In [33]:
```
#
flt_index = data[data['total_score'].isnull()].index
flt_index
```

Out[33]:
```
Int64Index([ 100,  101,  102,  103,  104,  105,  106,  107,  108,  109,
            ...
            4887, 4888, 4889, 4890, 4891, 4892, 4893, 4894, 4895, 4896],
           dtype='int64', length=3796)
```

In [34]:
```
#
data[data.index.isin(flt_index)]
```

Out[34]:
```
    world_rank                          university_name national_rank  \
100    101-152                         Aarhus University             2
101    101-152             Arizona State University - Tempe         54-71
102    101-152                  Baylor College of Medicine         54-71
103    101-152                Catholic University of Leuven           1-4
104    101-152               Catholic University of Louvain           1-4
105    101-152                           College of France             5
106    101-152                           Dartmouth College         54-71
107    101-152                            Emory University         54-71
108    101-152               Georgia Institute of Technology         54-71
109    101-152                            Ghent University           1-4
110    101-152                         Hokkaido University           6-9
111    101-152                           Kyushu University           6-9
112    101-152                         Mayo Medical School         54-71
113    101-152                           Nagoya University           6-9
114    101-152           National University of Singapore             1
115    101-152     North Carolina State University - Raleigh         54-71
116    101-152                     Oregon State University         54-71
117    101-152                    Seoul National University             1
118    101-152  State University of New York at Stony Brook         54-71
119    101-152       Technion-Israel Institute of Technology           2-4
120    101-152                         Tel Aviv University           2-4
121    101-152                    The University of Georgia         54-71
122    101-152                    The University of Glasgow         12-15
123    101-152                 The University of Queensland           3-4
124    101-152                  University Libre Bruxelles           1-4
```

125	101-152	University of Alberta	5
126	101-152	University of Amsterdam	3-4
127	101-152	University of Bonn	6-11
128	101-152	University of California, Riverside	54-71
129	101-152	University of California, Santa Cruz	54-71
...	...	...	...
4867	401-500	University of Jena	29-39
4868	401-500	University of Jyvaskyla	4-6
4869	401-500	University of Konstanz	29-39
4870	401-500	University of KwaZulu-Natal	3-4
4871	401-500	University of Ljubljana	1
4872	401-500	University of Maryland, Baltimore County	126-146
4873	401-500	University of Milan - Bicocca	11-20
4874	401-500	University of Nice Sophia Antipolis	19-22
4875	401-500	University of Oklahoma - Norman	126-146
4876	401-500	University of Palermo	11-20
4877	401-500	University of Parma	11-20
4878	401-500	University of Pavia	11-20
4879	401-500	University of Perugia	11-20
4880	401-500	University of Quebec	19-20
4881	401-500	University of Regensburg	29-39
4882	401-500	University of Rennes 1	19-22
4883	401-500	University of Rhode Island	126-146
4884	401-500	University of Roma - Tor Vergata	11-20
4885	401-500	University of Rostock	29-39
4886	401-500	University of Santiago Compostela	9-13
4887	401-500	University of Science, Malaysia	2
4888	401-500	University of Seville	9-13
4889	401-500	University of Surrey	34-37
4890	401-500	University of Szeged	1-2
4891	401-500	University of the Basque Country	9-13
4892	401-500	University of Trieste	11-20
4893	401-500	University of Zaragoza	9-13
4894	401-500	Utah State University	126-146
4895	401-500	Vienna University of Technology	4-6
4896	401-500	Wake Forest University	126-146

	total_score	alumni	award	hici	ns	pub	pcp	year
100	NaN	15.4	19.3	7.9	22.3	41.6	22.4	2005
101	NaN	0.0	14.4	20.8	26.3	41.9	17.5	2005
102	NaN	0.0	0.0	17.6	34.5	44.0	24.9	2005
103	NaN	0.0	0.0	19.2	16.0	48.7	23.1	2005
104	NaN	14.0	13.9	13.6	8.3	44.7	26.9	2005
105	NaN	15.4	37.4	11.1	11.7	16.9	19.3	2005
106	NaN	24.3	0.0	20.8	22.0	33.0	29.1	2005
107	NaN	0.0	0.0	28.3	19.0	48.4	21.6	2005
108	NaN	16.6	0.0	23.6	19.0	43.9	25.8	2005
109	NaN	8.9	15.8	15.7	9.1	48.8	27.2	2005

110	NaN	0.0	0.0	15.7	14.1	53.8	21.5	2005
111	NaN	0.0	0.0	13.6	21.3	52.8	21.6	2005
112	NaN	0.0	0.0	27.2	6.2	50.2	24.4	2005
113	NaN	0.0	14.4	15.7	20.5	52.3	25.1	2005
114	NaN	0.0	0.0	15.7	13.8	56.7	25.7	2005
115	NaN	0.0	0.0	29.4	17.8	44.3	19.0	2005
116	NaN	15.4	0.0	24.8	25.7	36.6	27.1	2005
117	NaN	0.0	0.0	7.9	14.6	61.2	26.9	2005
118	NaN	0.0	0.0	17.6	31.4	40.7	20.5	2005
119	NaN	18.8	23.5	13.6	14.1	42.1	22.8	2005
120	NaN	0.0	0.0	24.8	20.1	54.7	26.9	2005
121	NaN	0.0	0.0	28.3	21.1	46.4	18.4	2005
122	NaN	10.9	0.0	19.2	17.1	44.4	22.1	2005
123	NaN	16.6	0.0	7.9	19.9	50.1	18.9	2005
124	NaN	28.1	19.3	0.0	12.8	37.8	29.1	2005
125	NaN	15.4	0.0	17.6	17.4	55.1	26.1	2005
126	NaN	8.9	0.0	19.2	22.2	50.1	23.1	2005
127	NaN	19.8	20.4	15.7	11.3	41.3	22.0	2005
128	NaN	0.0	0.0	28.3	25.9	36.5	27.0	2005
129	NaN	0.0	0.0	28.3	28.5	31.1	29.2	2005
...	...	...	...	...	...	...	...	...
4867	NaN	0.0	0.0	0.0	9.3	34.0	17.1	2015
4868	NaN	0.0	0.0	3.6	10.3	26.7	14.1	2015
4869	NaN	0.0	0.0	12.1	11.4	22.2	13.5	2015
4870	NaN	0.0	0.0	3.6	8.4	32.8	16.6	2015
4871	NaN	0.0	0.0	0.0	7.7	35.1	14.2	2015
4872	NaN	0.0	0.0	17.4	6.5	17.8	17.3	2015
4873	NaN	0.0	0.0	5.0	5.6	30.9	21.4	2015
4874	NaN	0.0	0.0	3.6	16.3	26.2	12.7	2015
4875	NaN	0.0	0.0	5.1	10.0	28.0	14.0	2015
4876	NaN	0.0	0.0	6.3	6.6	28.2	14.8	2015
4877	NaN	0.0	0.0	13.6	2.1	26.7	19.6	2015
4878	NaN	0.0	0.0	5.0	5.7	30.7	19.7	2015
4879	NaN	0.0	0.0	3.6	7.5	29.3	18.5	2015
4880	NaN	0.0	0.0	0.0	9.8	33.3	16.8	2015
4881	NaN	0.0	0.0	3.6	13.9	27.7	15.2	2015
4882	NaN	0.0	0.0	3.6	9.2	28.1	11.2	2015
4883	NaN	0.0	0.0	15.2	6.1	21.1	16.0	2015
4884	NaN	0.0	0.0	0.0	8.8	33.7	19.2	2015
4885	NaN	0.0	0.0	8.6	8.4	25.0	13.5	2015
4886	NaN	0.0	0.0	7.1	6.1	31.1	13.2	2015
4887	NaN	0.0	0.0	7.1	3.3	30.6	15.7	2015
4888	NaN	0.0	0.0	0.0	7.5	33.7	11.3	2015
4889	NaN	0.0	0.0	8.6	4.9	27.0	18.0	2015
4890	NaN	0.0	13.3	3.6	3.4	21.8	12.8	2015
4891	NaN	0.0	0.0	3.6	7.1	36.1	13.5	2015
4892	NaN	0.0	0.0	5.0	10.9	25.1	20.1	2015
4893	NaN	0.0	0.0	7.6	5.1	33.3	13.1	2015

```
4894          NaN    13.6    0.0    3.6   10.8   25.1   15.5   2015
4895          NaN     0.0    0.0    0.0   12.2   28.8   22.9   2015
4896          NaN     0.0    0.0   14.9    7.5   25.0   11.9   2015

[3796 rows x 11 columns]
```

In [36]:
```
#
data_num[data_num.index.isin(flt_index)]['total_score']
```

Out[36]:
```
100     NaN
101     NaN
102     NaN
103     NaN
104     NaN
105     NaN
106     NaN
107     NaN
108     NaN
109     NaN
110     NaN
111     NaN
112     NaN
113     NaN
114     NaN
115     NaN
116     NaN
117     NaN
118     NaN
119     NaN
120     NaN
121     NaN
122     NaN
123     NaN
124     NaN
125     NaN
126     NaN
127     NaN
128     NaN
129     NaN
         ..
4867    NaN
4868    NaN
4869    NaN
4870    NaN
4871    NaN
4872    NaN
4873    NaN
4874    NaN
```

```
4875   NaN
4876   NaN
4877   NaN
4878   NaN
4879   NaN
4880   NaN
4881   NaN
4882   NaN
4883   NaN
4884   NaN
4885   NaN
4886   NaN
4887   NaN
4888   NaN
4889   NaN
4890   NaN
4891   NaN
4892   NaN
4893   NaN
4894   NaN
4895   NaN
4896   NaN
Name: total_score, Length: 3796, dtype: float64
```

In [37]:
```
data_num_MasVnrArea = data_num[['total_score']]
data_num_MasVnrArea.head()
```

Out[37]:
```
   total_score
0        100.0
1         73.6
2         73.4
3         72.8
4         70.1
```

In [38]:
```
from sklearn.impute import SimpleImputer
from sklearn.impute import MissingIndicator
```

In [39]:
```
#
indicator = MissingIndicator()
mask_missing_values_only = indicator.fit_transform(data_num_MasVnrArea)
mask_missing_values_only
```

Out[39]:
```
array([[False],
       [False],
       [False],
       ...,
       [ True],
       [ True],
       [ True]])
```

```
In [40]: strategies=['mean', 'median','most_frequent']

In [41]: def test_num_impute(strategy_param):
             imp_num = SimpleImputer(strategy=strategy_param)
             data_num_imp = imp_num.fit_transform(data_num_MasVnrArea)
             return data_num_imp[mask_missing_values_only]

In [42]: strategies[0], test_num_impute(strategies[0])

Out[42]: ('mean', array([36.38346957, 36.38346957, 36.38346957, ..., 36.38346957,
                 36.38346957, 36.38346957]))

In [44]: strategies[1], test_num_impute(strategies[1])

Out[44]: ('median', array([31.3, 31.3, 31.3, ..., 31.3, 31.3, 31.3]))

In [45]: strategies[2], test_num_impute(strategies[2])

Out[45]: ('most_frequent', array([24.9, 24.9, 24.9, ..., 24.9, 24.9, 24.9]))

In [46]: #   ,
         def test_num_impute_col(dataset, column, strategy_param):
             temp_data = dataset[[column]]

             indicator = MissingIndicator()
             mask_missing_values_only = indicator.fit_transform(temp_data)

             imp_num = SimpleImputer(strategy=strategy_param)
             data_num_imp = imp_num.fit_transform(temp_data)

             filled_data = data_num_imp[mask_missing_values_only]

             return column, strategy_param, filled_data.size, filled_data[0], filled_data[fille

In [49]: data[['hici']].describe()

Out[49]:               hici
         count  4895.000000
         mean     16.221491
         std      14.382710
         min       0.000000
         25%       7.300000
         50%      12.600000
         75%      21.700000
         max     100.000000

In [50]: test_num_impute_col(data, 'hici', strategies[0])

Out[50]: ('hici', 'mean', 2, 16.22149131767109, 16.22149131767109)
```

```
In [51]: test_num_impute_col(data, 'hici', strategies[1])

Out[51]: ('hici', 'median', 2, 12.6, 12.6)

In [52]: test_num_impute_col(data, 'hici', strategies[2])

Out[52]: ('hici', 'most_frequent', 2, 0.0, 0.0)
```

1.2.2.

```
In [53]: #
         #
         cat_cols = []
         for col in data.columns:
             #
             temp_null_count = data[data[col].isnull()].shape[0]
             dt = str(data[col].dtype)
             if temp_null_count>0 and (dt=='object'):
                 cat_cols.append(col)
                 temp_perc = round((temp_null_count / total_count) * 100.0, 2)
                 print(' {}.   {}.    {}, {}%.'.format(col, dt, temp_null_count, temp_perc))

 university_name.   object.    1, 0.02%.
 national_rank.   object.    1, 0.02%.


In [54]: cat_temp_data = data[['national_rank']]
         cat_temp_data.head()

Out[54]:   national_rank
         0             1
         1             1
         2             2
         3             3
         4             4

In [56]: cat_temp_data['national_rank'].unique()

Out[56]: 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', '26', '27', '28', '29', '30', '32', '33', '35', '37', '38',
                '39', '40', '41', '42', '43', '44', '45', '46', '47', '48', '49',
                '50', '51', '52', '53', '54-71', '1-4', '6-9', '2-4', '12-15',
                '3-4', '6-11', '2-3', '1-2', '72-90', '16-19', '5-7', '6-8', '5-6',
                '12-16', '4-5', '9-13', '9-17', '8-9', '91-119', '10-13', '20-30',
                '7-9', '17-23', '1-3', '120-140', '18-19', '14-24', '24-33',
                '14-19', '10-11', '3-7', '10-18', '3-5', '31-36', '5-9', '141-168',
                '25-34', '11-14', '34-40', '20-23', '19-23', '37-40', '20-21',
                '25', '31', '34', '36', '54', '55-69', '16-22', '70-87', '4-6',
```

```
       '88-118', '7-12', '9-16', '10-12', '23-33', '2-5', '119-140',
       '17-19', '13-20', '23-36', '13-17', '6-7', '8-14', '34-37',
       '38-43', '141-167', '18-21', '21-32', '15-23', '20-22', '55-70',
       '7-11', '16-23', '71-88', '89-117', '8-12', '8-17', '2-6', '15-22',
       '118-140', '13-18', '9-14', '38-42', '141-166', '19-33', '37-41',
       '15-20', '12-17', '71-90', '17-22', '12-14', '91-114', '7-18',
       '1-6', '15-24', '115-139', '25-35', '15-17', '10-14', '4-7',
       '34-38', '140-159', '39-42', '19-21', '13-22', '19-31', '8-18',
       '18-23', '36-40', '55', '56-70', '3-6', '91-112', '113-138',
       '25-36', '15-19', '7-8', '12-19', '8-13', '34-36', '139-152',
       '19-22', '14-21', '20-31', '9-18', '14-17', '6-10', '70-89',
       '90-111', '5-8', '112-137', '14-18', '11-17', '31-35', '8-10',
       '138-154', '36-38', '14-22', '18-25', '11-22', '34-39', '54-68',
       '11-15', '7-10', '69-89', '90-110', '2-7', '20-29', '111-137',
       '11-12', '11-16', '24-32', '30-33', '9-10', '138-151', '13-23',
       '8-11', '33-39', '54-67', '68-85', '86-109', '1.0', '2.0', '3.0',
       '4.0', '5.0', '6.0', '7.0', '8.0', '9.0', '10.0', '11.0', '12.0',
       '13.0', '14.0', '15.0', '16.0', '17.0', '18.0', '19.0', '20.0',
       '21.0', '22.0', '23.0', '24.0', '25.0', '26.0', '27.0', '28.0',
       '29.0', '30.0', '31.0', '32.0', '33.0', '35.0', '36.0', '37.0',
       '38.0', '39.0', '41.0', '42.0', '43.0', '44.0', '45.0', '46.0',
       '47.0', '49.0', '50.0', nan, '53-64', '65-77', '18-20', '78-104',
       '8-16', '21-29', '13-25', '105-125', '23-30', '9-12', '17-18',
       '126-146', '31-39', '26-32', '13-21', '52-65', '10-17', '66-78',
       '1-5', '79-102', '9-15', '9-11', '7-16', '22-28', '14-27',
       '103-125', '16-18', '29-33', '29-39', '28-32', '19-20', '11-20'],
      dtype=object)
```

In [57]:
```
cat_temp_data[cat_temp_data['national_rank'].isnull()].shape
```

Out[57]:
```
(1, 1)
```

In [58]:
```
#
imp2 = SimpleImputer(missing_values=np.nan, strategy='most_frequent')
data_imp2 = imp2.fit_transform(cat_temp_data)
data_imp2
```

Out[58]:
```
array([['1'],
       ['1'],
       ['2'],
       ...,
       ['126-146'],
       ['4-6'],
       ['126-146']], dtype=object)
```

In [59]:
```
#
np.unique(data_imp2)
```

Out[59]:
```
array(['1', '1-2', '1-3', '1-4', '1-5', '1-6', '1.0', '10', '10-11',
       '10-12', '10-13', '10-14', '10-17', '10-18', '10.0', '103-125',
```

```
       '105-125', '11', '11-12', '11-14', '11-15', '11-16', '11-17',
       '11-20', '11-22', '11.0', '111-137', '112-137', '113-138',
       '115-139', '118-140', '119-140', '12', '12-14', '12-15', '12-16',
       '12-17', '12-19', '12.0', '120-140', '126-146', '13', '13-17',
       '13-18', '13-20', '13-21', '13-22', '13-23', '13-25', '13.0',
       '138-151', '138-154', '139-152', '14', '14-17', '14-18', '14-19',
       '14-21', '14-22', '14-24', '14-27', '14.0', '140-159', '141-166',
       '141-167', '141-168', '15', '15-17', '15-19', '15-20', '15-22',
       '15-23', '15-24', '15.0', '16', '16-18', '16-19', '16-22', '16-23',
       '16.0', '17', '17-18', '17-19', '17-22', '17-23', '17.0', '18',
       '18-19', '18-20', '18-21', '18-23', '18-25', '18.0', '19', '19-20',
       '19-21', '19-22', '19-23', '19-31', '19-33', '19.0', '2', '2-3',
       '2-4', '2-5', '2-6', '2-7', '2.0', '20', '20-21', '20-22', '20-23',
       '20-29', '20-30', '20-31', '20.0', '21', '21-29', '21-32', '21.0',
       '22', '22-28', '22.0', '23', '23-30', '23-33', '23-36', '23.0',
       '24', '24-32', '24-33', '24.0', '25', '25-34', '25-35', '25-36',
       '25.0', '26', '26-32', '26.0', '27', '27.0', '28', '28-32', '28.0',
       '29', '29-33', '29-39', '29.0', '3', '3-4', '3-5', '3-6', '3-7',
       '3.0', '30', '30-33', '30.0', '31', '31-35', '31-36', '31-39',
       '31.0', '32', '32.0', '33', '33-39', '33.0', '34', '34-36',
       '34-37', '34-38', '34-39', '34-40', '35', '35.0', '36', '36-38',
       '36-40', '36.0', '37', '37-40', '37-41', '37.0', '38', '38-42',
       '38-43', '38.0', '39', '39-42', '39.0', '4', '4-5', '4-6', '4-7',
       '4.0', '40', '41', '41.0', '42', '42.0', '43', '43.0', '44',
       '44.0', '45', '45.0', '46', '46.0', '47', '47.0', '48', '49',
       '49.0', '5', '5-6', '5-7', '5-8', '5-9', '5.0', '50', '50.0', '51',
       '52', '52-65', '53', '53-64', '54', '54-67', '54-68', '54-71',
       '55', '55-69', '55-70', '56-70', '6', '6-10', '6-11', '6-7', '6-8',
       '6-9', '6.0', '65-77', '66-78', '68-85', '69-89', '7', '7-10',
       '7-11', '7-12', '7-16', '7-18', '7-8', '7-9', '7.0', '70-87',
       '70-89', '71-88', '71-90', '72-90', '78-104', '79-102', '8',
       '8-10', '8-11', '8-12', '8-13', '8-14', '8-16', '8-17', '8-18',
       '8-9', '8.0', '86-109', '88-118', '89-117', '9', '9-10', '9-11',
       '9-12', '9-13', '9-14', '9-15', '9-16', '9-17', '9-18', '9.0',
       '90-110', '90-111', '91-112', '91-114', '91-119'], dtype=object)
```

In [60]:
```python
#
imp3 = SimpleImputer(missing_values=np.nan, strategy='constant', fill_value='!!!')
data_imp3 = imp3.fit_transform(cat_temp_data)
data_imp3
```

Out[60]:
```
array([['1'],
       ['1'],
       ['2'],
       ...,
       ['126-146'],
       ['4-6'],
       ['126-146']], dtype=object)
```

```
In [61]: np.unique(data_imp3)

Out[61]: array(['!!!', '1', '1-2', '1-3', '1-4', '1-5', '1-6', '1.0', '10',
               '10-11', '10-12', '10-13', '10-14', '10-17', '10-18', '10.0',
               '103-125', '105-125', '11', '11-12', '11-14', '11-15', '11-16',
               '11-17', '11-20', '11-22', '11.0', '111-137', '112-137', '113-138',
               '115-139', '118-140', '119-140', '12', '12-14', '12-15', '12-16',
               '12-17', '12-19', '12.0', '120-140', '126-146', '13', '13-17',
               '13-18', '13-20', '13-21', '13-22', '13-23', '13-25', '13.0',
               '138-151', '138-154', '139-152', '14', '14-17', '14-18', '14-19',
               '14-21', '14-22', '14-24', '14-27', '14.0', '140-159', '141-166',
               '141-167', '141-168', '15', '15-17', '15-19', '15-20', '15-22',
               '15-23', '15-24', '15.0', '16', '16-18', '16-19', '16-22', '16-23',
               '16.0', '17', '17-18', '17-19', '17-22', '17-23', '17.0', '18',
               '18-19', '18-20', '18-21', '18-23', '18-25', '18.0', '19', '19-20',
               '19-21', '19-22', '19-23', '19-31', '19-33', '19.0', '2', '2-3',
               '2-4', '2-5', '2-6', '2-7', '2.0', '20', '20-21', '20-22', '20-23',
               '20-29', '20-30', '20-31', '20.0', '21', '21-29', '21-32', '21.0',
               '22', '22-28', '22.0', '23', '23-30', '23-33', '23-36', '23.0',
               '24', '24-32', '24-33', '24.0', '25', '25-34', '25-35', '25-36',
               '25.0', '26', '26-32', '26.0', '27', '27.0', '28', '28-32', '28.0',
               '29', '29-33', '29-39', '29.0', '3', '3-4', '3-5', '3-6', '3-7',
               '3.0', '30', '30-33', '30.0', '31', '31-35', '31-36', '31-39',
               '31.0', '32', '32.0', '33', '33-39', '33.0', '34', '34-36',
               '34-37', '34-38', '34-39', '34-40', '35', '35.0', '36', '36-38',
               '36-40', '36.0', '37', '37-40', '37-41', '37.0', '38', '38-42',
               '38-43', '38.0', '39', '39-42', '39.0', '4', '4-5', '4-6', '4-7',
               '4.0', '40', '41', '41.0', '42', '42.0', '43', '43.0', '44',
               '44.0', '45', '45.0', '46', '46.0', '47', '47.0', '48', '49',
               '49.0', '5', '5-6', '5-7', '5-8', '5-9', '5.0', '50', '50.0', '51',
               '52', '52-65', '53', '53-64', '54', '54-67', '54-68', '54-71',
               '55', '55-69', '55-70', '56-70', '6', '6-10', '6-11', '6-7', '6-8',
               '6-9', '6.0', '65-77', '66-78', '68-85', '69-89', '7', '7-10',
               '7-11', '7-12', '7-16', '7-18', '7-8', '7-9', '7.0', '70-87',
               '70-89', '71-88', '71-90', '72-90', '78-104', '79-102', '8',
               '8-10', '8-11', '8-12', '8-13', '8-14', '8-16', '8-17', '8-18',
               '8-9', '8.0', '86-109', '88-118', '89-117', '9', '9-10', '9-11',
               '9-12', '9-13', '9-14', '9-15', '9-16', '9-17', '9-18', '9.0',
               '90-110', '90-111', '91-112', '91-114', '91-119'], dtype=object)

In [62]: data_imp3[data_imp3=='!!!'].size

Out[62]: 1
```

lab3_2

March 6, 2019

2.

```
In [1]: import numpy as np
        import pandas as pd
        import seaborn as sns
        import matplotlib.pyplot as plt
        %matplotlib inline
        sns.set(style="ticks")
        from sklearn.impute import SimpleImputer
        from sklearn.impute import MissingIndicator
        from sklearn.preprocessing import LabelEncoder, OneHotEncoder
```

```
In [2]: data = pd.read_csv(r'student-por.csv', sep=",")
        data.head()
```

```
Out[2]:   school sex  age address famsize Pstatus  Medu  Fedu     Mjob      Fjob  ...  \
        0     GP   F   18       U     GT3       A     4     4  at_home   teacher  ...
        1     GP   F   17       U     GT3       T     1     1  at_home     other  ...
        2     GP   F   15       U     LE3       T     1     1  at_home     other  ...
        3     GP   F   15       U     GT3       T     4     2   health  services  ...
        4     GP   F   16       U     GT3       T     3     3    other     other  ...

          famrel freetime  goout  Dalc  Walc health absences  G1  G2  G3
        0      4        3      4     1     1      3        4   0  11  11
        1      5        3      3     1     1      3        2   9  11  11
        2      4        3      2     2     3      3        6  12  13  12
        3      3        2      2     1     1      5        0  14  14  14
        4      4        3      2     1     2      5        0  11  13  13

        [5 rows x 33 columns]
```

2.1. - label encoding

```
In [3]: le = LabelEncoder()
        data_le = le.fit_transform(data['sex'])
```

```
In [4]: data['sex'].unique()
```

```
Out[4]: array(['F', 'M'], dtype=object)
```

In [5]: np.unique(data_le)

Out[5]: array([0, 1])

In [6]: le.inverse_transform([0, 1])

Out[6]: array(['F', 'M'], dtype=object)

2.2. - one-hot encoding

```
In [7]: ohe = OneHotEncoder()
        data_ohe = ohe.fit_transform(data[['sex']])
```

In [8]: data.shape

Out[8]: (649, 33)

In [9]: data_ohe.shape

Out[9]: (649, 2)

In [10]: data_ohe

```
Out[10]: <649x2 sparse matrix of type '<class 'numpy.float64'>'
                 with 649 stored elements in Compressed Sparse Row format>
```

In [11]: data_ohe.todense()[0:10]

```
Out[11]: matrix([[1., 0.],
                 [1., 0.],
                 [1., 0.],
                 [1., 0.],
                 [1., 0.],
                 [0., 1.],
                 [0., 1.],
                 [1., 0.],
                 [0., 1.],
                 [0., 1.]])
```

In [12]: data.head(10)

```
Out[12]:   school sex  age address famsize Pstatus  Medu  Fedu      Mjob      Fjob  \
         0     GP   F   18       U     GT3       A     4     4   at_home   teacher
         1     GP   F   17       U     GT3       T     1     1   at_home     other
         2     GP   F   15       U     LE3       T     1     1   at_home     other
         3     GP   F   15       U     GT3       T     4     2    health  services
         4     GP   F   16       U     GT3       T     3     3     other     other
         5     GP   M   16       U     LE3       T     4     3  services     other
         6     GP   M   16       U     LE3       T     2     2     other     other
         7     GP   F   17       U     GT3       A     4     4     other   teacher
```

```
8      GP    M    15        U      LE3        A      3      2  services      other
9      GP    M    15        U      GT3        T      3      4     other      other

   ... famrel freetime  goout  Dalc  Walc health absences  G1  G2  G3
0  ...      4        3      4     1     1      3        4   0  11  11
1  ...      5        3      3     1     1      3        2   9  11  11
2  ...      4        3      2     2     3      3        6  12  13  12
3  ...      3        2      2     1     1      5        0  14  14  14
4  ...      4        3      2     1     2      5        0  11  13  13
5  ...      5        4      2     1     2      5        6  12  12  13
6  ...      4        4      4     1     1      3        0  13  12  13
7  ...      4        1      4     1     1      1        2  10  13  13
8  ...      4        2      2     1     1      1        0  15  16  17
9  ...      5        5      1     1     1      5        0  12  12  13

[10 rows x 33 columns]
```

2.3. Pandas get_dummies -   one-hot ű

In [13]:
```
# pd.get_dummies(data).head()
pd.get_dummies(data.sex).head()
```

Out[13]:
```
   F  M
0  1  0
1  1  0
2  1  0
3  1  0
4  1  0
```

In [ ]:

lab3_3

March 6, 2019

```
In [2]: import numpy as np
        import pandas as pd
        import seaborn as sns
        import matplotlib.pyplot as plt
        %matplotlib inline
        sns.set(style="ticks")
        from sklearn.preprocessing import MinMaxScaler, StandardScaler, Normalizer
```

3.1. MinMax ű

```
In [15]: data = pd.read_csv(r'cwurData.csv', sep=",")
         data.head()
```

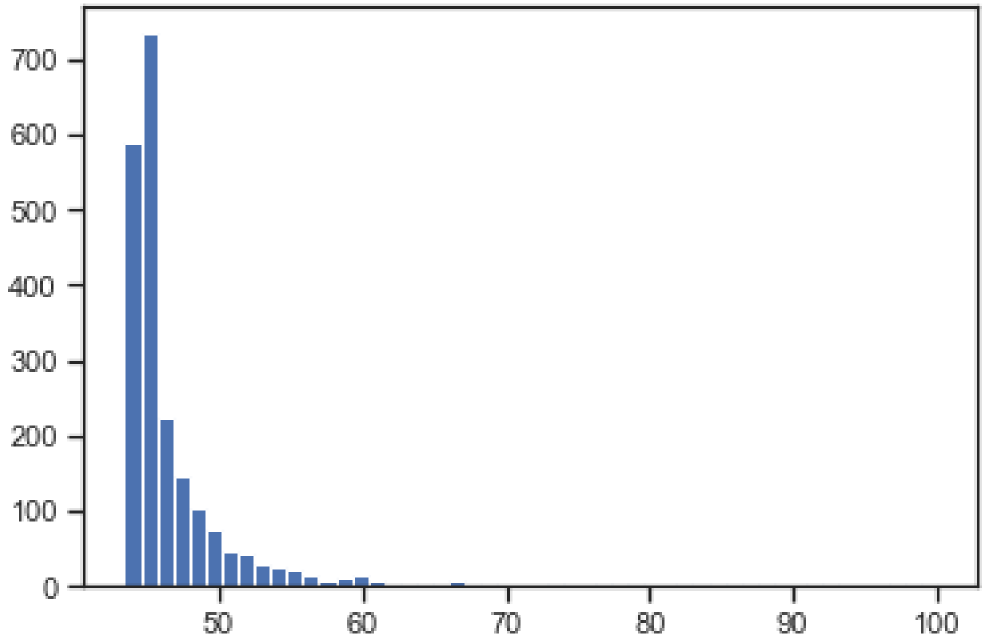
```
Out[15]:    world_rank                            institution         country  \
         0           1                     Harvard University             USA
         1           2  Massachusetts Institute of Technology             USA
         2           3                    Stanford University             USA
         3           4                University of Cambridge  United Kingdom
         4           5     California Institute of Technology             USA

            national_rank  quality_of_education  alumni_employment  quality_of_faculty  \
         0              1                     7                  9                   1
         1              2                     9                 17                   3
         2              3                    17                 11                   5
         3              1                    10                 24                   4
         4              4                     2                 29                   7

            publications  influence  citations  broad_impact  patents   score  year
         0             1          1          1           NaN        5  100.00  2012
         1            12          4          4           NaN        1   91.67  2012
         2             4          2          2           NaN       15   89.50  2012
         3            16         16         11           NaN       50   86.17  2012
         4            37         22         22           NaN       18   85.21  2012
```

```
In [16]: sc1 = MinMaxScaler()
         sc1_data = sc1.fit_transform(data[['score']])
```

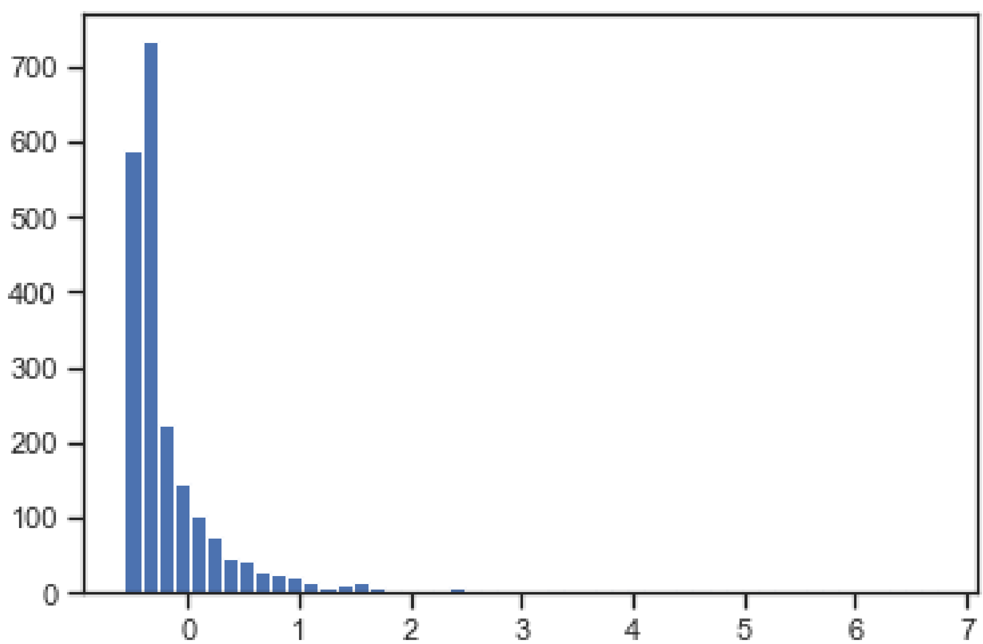
```
In [17]: plt.hist(data['score'], 50)
         plt.show()
```



3.2. Z- - StandardScalerű

```
In [18]: sc2 = StandardScaler()
         sc2_data = sc2.fit_transform(data[['score']])
```

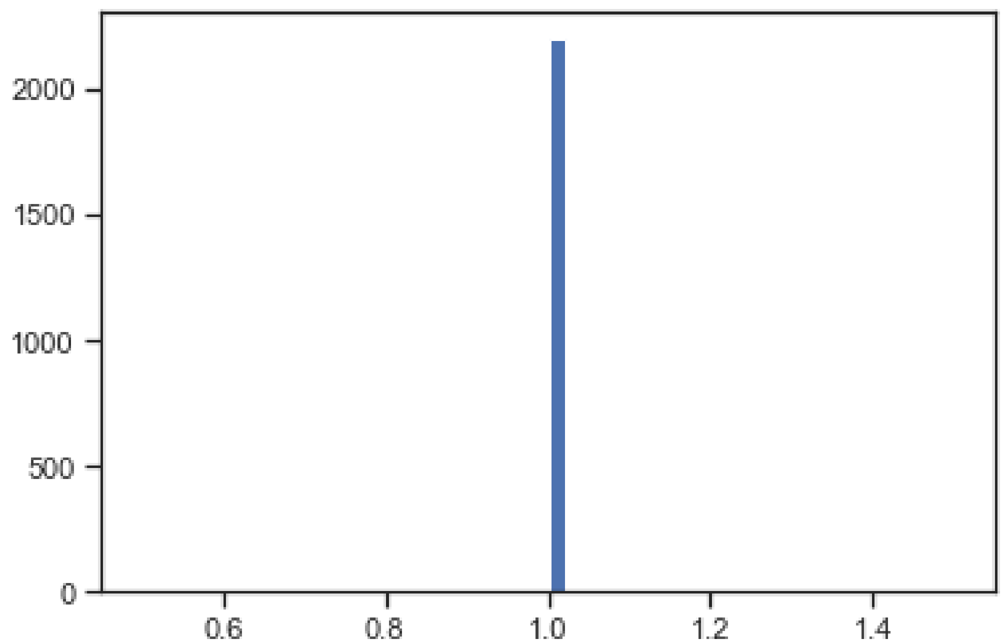
```
In [19]: plt.hist(sc2_data, 50)
         plt.show()
```



3.3.

```
In [20]: sc3 = Normalizer()
         sc3_data = sc3.fit_transform(data[['score']])
```

```
In [21]: plt.hist(sc3_data, 50)
         plt.show()
```




```
In [ ]:
```