Лабораторная работа №2. Предобработка данных и проектирование признаков.

Используемый набор данных: Energy efficiency (https://archive.ics.uci.edu/ml/datasets/Energy+efficiency).

In [1]:

```
from IPython.display import display
import numpy as np
import pandas as pd
from sklearn import preprocessing
import os
import requests
import xlrd # для pd.read_excel()

%matplotlib inline

pd.options.display.max_columns = None
```

In [2]:

```
def downloadFile(url, filePath):
    if not os.path.exists(filePath):
        req = requests.get(url)
        f = open(filePath, "wb")
        f.write(req.content)
        f.close

url = "https://archive.ics.uci.edu/ml/machine-learning-databases/00242"
downloadFile(url + "/ENB2012_data.xlsx", "dataset/ENB2012_data.xlsx")
```

In [3]:

```
headers = ["Relative Compactness", "Surface Area", "Wall Area", "Roof Area", "Overall H
eight", "Orientation",
           "Glazing Area",
           "Glazing Area Distribution", "Heating Load", "Cooling Load"]
data = pd.read_excel("dataset/ENB2012_data.xlsx", names=headers)
data.sample(40)
```

Out[3]:

	Relative Compactness	Surface Area	Wall Area	Roof Area	Overall Height	Orientation	Glazing Area	Glazing Area Distribution	Heatin Loa
710	0.66	759.5	318.5	220.50	3.5	4	0.40	4	15.0
696	0.74	686.0	245.0	220.50	3.5	2	0.40	4	14.3
569	0.64	784.0	343.0	220.50	3.5	3	0.40	1	19.3
418	0.69	735.0	294.0	220.50	3.5	4	0.25	3	12.3
723	0.98	514.5	294.0	110.25	7.0	5	0.40	5	32.7
550	0.76	661.5	416.5	122.50	7.0	4	0.40	1	40.7
694	0.76	661.5	416.5	122.50	7.0	4	0.40	4	40.6
221	0.71	710.5	269.5	220.50	3.5	3	0.10	4	10.6
720	0.98	514.5	294.0	110.25	7.0	2	0.40	5	32.8
488	0.86	588.0	294.0	147.00	7.0	2	0.25	5	29.7
425	0.64	784.0	343.0	220.50	3.5	3	0.25	3	16.9
111	0.82	612.5	318.5	147.00	7.0	5	0.10	2	22.7
458	0.74	686.0	245.0	220.50	3.5	4	0.25	4	12.1
607	0.71	710.5	269.5	220.50	3.5	5	0.40	2	14.7
282	0.64	784.0	343.0	220.50	3.5	4	0.10	5	15.1
106	0.86	588.0	294.0	147.00	7.0	4	0.10	2	26.3
22	0.76	661.5	416.5	122.50	7.0	4	0.00	0	24.7
493	0.82	612.5	318.5	147.00	7.0	3	0.25	5	25.1
56	0.86	588.0	294.0	147.00	7.0	2	0.10	1	26.2
70	0.76	661.5	416.5	122.50	7.0	4	0.10	1	32.9
486	0.90	563.5	318.5	122.50	7.0	4	0.25	5	31.5
616	0.64	784.0	343.0	220.50	3.5	2	0.40	2	19.2
445	0.82	612.5	318.5	147.00	7.0	3	0.25	4	24.9
563	0.69	735.0	294.0	220.50	3.5	5	0.40	1	14.4
734	0.82	612.5	318.5	147.00	7.0	4	0.40	5	29.0
10	0.86	588.0	294.0	147.00	7.0	4	0.00	0	19.3
140	0.62	808.5	367.5	220.50	3.5	2	0.10	2	12.8
190	0.62	808.5	367.5	220.50	3.5	4	0.10	3	12.7
209	0.79	637.0	343.0	147.00	7.0	3	0.10	4	35.8
725	0.90	563.5	318.5	122.50	7.0	3	0.40	5	35.0
664	0.64	784.0	343.0	220.50	3.5	2	0.40	3	18.4
685	0.82	612.5	318.5	147.00	7.0	3	0.40	4	28.0
677	0.90	563.5	318.5	122.50	7.0	3	0.40	4	35.7
596	0.76	661.5	416.5	122.50	7.0	2	0.40	2	40.7
85	0.66	759.5	318.5	220.50	3.5	3	0.10	1	11.6
232	0.64	784.0	343.0	220.50	3.5	2	0.10	4	15.4

	Relative Compactness	Surface Area	Wall Area	Roof Area	Overall Height	Orientation	Glazing Area	Glazing Area Distribution	Heatin Loa
201	0.86	588.0	294.0	147.00	7.0	3	0.10	4	25.3
246	0.90	563.5	318.5	122.50	7.0	4	0.10	5	28.0
277	0.66	759.5	318.5	220.50	3.5	3	0.10	5	11.3
740	0.76	661.5	416.5	122.50	7.0	2	0.40	5	38.8

In [4]:

```
data.dtypes
```

Out[4]:

```
Relative Compactness         float64
Surface Area                 float64
Wall Area                    float64
Roof Area                    float64
Overall Height               float64
Orientation                    int64
Glazing Area                 float64
Glazing Area Distribution      int64
Heating Load                 float64
Cooling Load                 float64
dtype: object
```

In [5]:

```
display(data.describe())
display(data.isna().sum())
```

	Relative Compactness	Surface Area	Wall Area	Roof Area	Overall Height	Orientation	Glazing Area
count	768.000000	768.000000	768.000000	768.000000	768.00000	768.000000	768.000000
mean	0.764167	671.708333	318.500000	176.604167	5.25000	3.500000	0.234375
std	0.105777	88.086116	43.626481	45.165950	1.75114	1.118763	0.133221
min	0.620000	514.500000	245.000000	110.250000	3.50000	2.000000	0.000000
25%	0.682500	606.375000	294.000000	140.875000	3.50000	2.750000	0.100000
50%	0.750000	673.750000	318.500000	183.750000	5.25000	3.500000	0.250000
75%	0.830000	741.125000	343.000000	220.500000	7.00000	4.250000	0.400000
max	0.980000	808.500000	416.500000	220.500000	7.00000	5.000000	0.400000

```
Relative Compactness         0
Surface Area                 0
Wall Area                    0
Roof Area                    0
Overall Height               0
Orientation                  0
Glazing Area                 0
Glazing Area Distribution    0
Heating Load                 0
Cooling Load                 0
dtype: int64
```

Добавим новый признак: объем здания (*Volume*). Он определяется как произведение высоты здания и площади его крыши.

In [6]:

```
data["Volume"] = data["Roof Area"] * data["Overall Height"]
data.sample(40)
```

Out[6]:

	Relative Compactness	Surface Area	Wall Area	Roof Area	Overall Height	Orientation	Glazing Area	Glazing Area Distribution	Heatin Loa
284	0.62	808.5	367.5	220.50	3.5	2	0.10	5	12.59
236	0.62	808.5	367.5	220.50	3.5	2	0.10	4	12.85
26	0.74	686.0	245.0	220.50	3.5	4	0.00	0	6.01
183	0.66	759.5	318.5	220.50	3.5	5	0.10	3	11.61
499	0.79	637.0	343.0	147.00	7.0	5	0.25	5	38.65
253	0.82	612.5	318.5	147.00	7.0	3	0.10	5	23.89
331	0.64	784.0	343.0	220.50	3.5	5	0.25	1	17.37
12	0.82	612.5	318.5	147.00	7.0	2	0.00	0	17.05
167	0.76	661.5	416.5	122.50	7.0	5	0.10	3	33.24
437	0.90	563.5	318.5	122.50	7.0	3	0.25	4	31.69
58	0.86	588.0	294.0	147.00	7.0	4	0.10	1	26.37
32	0.69	735.0	294.0	220.50	3.5	2	0.00	0	6.85
152	0.86	588.0	294.0	147.00	7.0	2	0.10	3	25.41
591	0.82	612.5	318.5	147.00	7.0	5	0.40	2	28.01
24	0.74	686.0	245.0	220.50	3.5	2	0.00	0	6.07
683	0.86	588.0	294.0	147.00	7.0	5	0.40	4	32.75
49	0.98	514.5	294.0	110.25	7.0	3	0.10	1	24.63
708	0.66	759.5	318.5	220.50	3.5	2	0.40	4	15.34
339	0.98	514.5	294.0	110.25	7.0	5	0.25	2	28.60
710	0.66	759.5	318.5	220.50	3.5	4	0.40	4	15.09
441	0.86	588.0	294.0	147.00	7.0	3	0.25	4	28.42
689	0.79	637.0	343.0	147.00	7.0	3	0.40	4	41.73
371	0.69	735.0	294.0	220.50	3.5	5	0.25	2	12.95
289	0.98	514.5	294.0	110.25	7.0	3	0.25	1	28.15
36	0.66	759.5	318.5	220.50	3.5	2	0.00	0	7.18
561	0.69	735.0	294.0	220.50	3.5	3	0.40	1	14.70
369	0.69	735.0	294.0	220.50	3.5	3	0.25	2	12.87
118	0.76	661.5	416.5	122.50	7.0	4	0.10	2	33.12
30	0.71	710.5	269.5	220.50	3.5	4	0.00	0	6.36
202	0.86	588.0	294.0	147.00	7.0	4	0.10	4	26.33
4	0.90	563.5	318.5	122.50	7.0	2	0.00	0	20.84
81	0.69	735.0	294.0	220.50	3.5	3	0.10	1	11.13
761	0.64	784.0	343.0	220.50	3.5	3	0.40	5	18.19
348	0.82	612.5	318.5	147.00	7.0	2	0.25	2	25.74
42	0.64	784.0	343.0	220.50	3.5	4	0.00	0	10.77
386	0.98	514.5	294.0	110.25	7.0	4	0.25	3	28.17

	Relative Compactness	Surface Area	Wall Area	Roof Area	Overall Height	Orientation	Glazing Area	Glazing Area Distribution	Heatin Loa
273	0.69	735.0	294.0	220.50	3.5	3	0.10	5	11.14
165	0.76	661.5	416.5	122.50	7.0	3	0.10	3	33.28
229	0.66	759.5	318.5	220.50	3.5	3	0.10	4	11.42
41	0.64	784.0	343.0	220.50	3.5	3	0.00	0	10.54

In [7]:

```
display(data.dtypes)
display(data.isna().sum())
```

```
Relative Compactness         float64
Surface Area                 float64
Wall Area                    float64
Roof Area                    float64
Overall Height               float64
Orientation                    int64
Glazing Area                 float64
Glazing Area Distribution      int64
Heating Load                 float64
Cooling Load                 float64
Volume                       float64
dtype: object

Relative Compactness         0
Surface Area                 0
Wall Area                    0
Roof Area                    0
Overall Height               0
Orientation                  0
Glazing Area                 0
Glazing Area Distribution    0
Heating Load                 0
Cooling Load                 0
Volume                       0
dtype: int64
```

In [8]:

```
data.describe()
```

Out[8]:

	Relative Compactness	Surface Area	Wall Area	Roof Area	Overall Height	Orientation	Glazing Area
count	768.000000	768.000000	768.000000	768.000000	768.00000	768.000000	768.000000
mean	0.764167	671.708333	318.500000	176.604167	5.25000	3.500000	0.234375
std	0.105777	88.086116	43.626481	45.165950	1.75114	1.118763	0.133221
min	0.620000	514.500000	245.000000	110.250000	3.50000	2.000000	0.000000
25%	0.682500	606.375000	294.000000	140.875000	3.50000	2.750000	0.100000
50%	0.750000	673.750000	318.500000	183.750000	5.25000	3.500000	0.250000
75%	0.830000	741.125000	343.000000	220.500000	7.00000	4.250000	0.400000
max	0.980000	808.500000	416.500000	220.500000	7.00000	5.000000	0.400000

Признаки различаются по масштабу. Выполним нормализацию.

In [9]:

```
data_norm = pd.DataFrame(preprocessing.normalize(data), columns=headers+["Volume"])
data_norm
```

Out[9]:

	Relative Compactness	Surface Area	Wall Area	Roof Area	Overall Height	Orientation	Glazing Area	Glazin Are Distributio
0	0.001000	0.525205	0.300117	0.112544	0.007146	0.002042	0.000000	0.00000
1	0.001000	0.525204	0.300116	0.112544	0.007146	0.003062	0.000000	0.00000
2	0.001000	0.525202	0.300115	0.112543	0.007146	0.004083	0.000000	0.00000
3	0.001000	0.525199	0.300114	0.112543	0.007146	0.005104	0.000000	0.00000
4	0.000832	0.520828	0.294381	0.113223	0.006470	0.001849	0.000000	0.00000
...	...	...	...	...	...	...	...	.
763	0.000545	0.668023	0.292260	0.187882	0.002982	0.004260	0.000341	0.00426
764	0.000518	0.675265	0.306938	0.184163	0.002923	0.001670	0.000334	0.00417
765	0.000518	0.675262	0.306937	0.184162	0.002923	0.002506	0.000334	0.00417
766	0.000518	0.675264	0.306938	0.184163	0.002923	0.003341	0.000334	0.00417
767	0.000518	0.675266	0.306939	0.184163	0.002923	0.004176	0.000334	0.00417

768 rows × 11 columns

In [10]:

```
data_norm.describe()
```

Out[10]:

	Relative Compactness	Surface Area	Wall Area	Roof Area	Overall Height	Orientation	Glazing Area
count	768.000000	768.000000	768.000000	768.000000	768.000000	768.000000	768.000000
mean	0.000670	0.584489	0.276841	0.153824	0.004572	0.003055	0.000205
std	0.000126	0.074275	0.034463	0.040308	0.001554	0.001003	0.000111
min	0.000518	0.477707	0.225982	0.104860	0.002923	0.001577	0.000000
25%	0.000602	0.515999	0.252259	0.115251	0.003089	0.002285	0.000085
50%	0.000646	0.599690	0.273785	0.151824	0.004375	0.003108	0.000213
75%	0.000687	0.654017	0.295769	0.192680	0.005765	0.003960	0.000325
max	0.001000	0.675350	0.356768	0.203417	0.007146	0.005104	0.000408