Министерство науки и высшего образования Российской Федерации Федеральное государственное бюджетное образовательное учреждение высшего образования «Московский государственный технический университет имени Н.Э. Баумана (национальный исследовательский университет)» (МГТУ им. Н.Э. Баумана)

Рубежный контроль №1

по курсу «Теория машинного обучения»

Вариант 16

Выполнил

студент группы ИУ5-64Б

Сысойкин Е.М.

Москва, 2020

0.1 РК1; ТМО; Сысойкин Егор; ИУ5-64Б

Для заданного набора данных проведите обработку пропусков в данных для одного категориального и одного количественного признака. Какие способы обработки пропусков в данных для категориальных и количественных признаков Вы использовали? Какие признаки Вы будете использовать для дальнейшего построения моделей машинного обучения и почему?

Дополнительно: для произвольного столбца датасета построить скрипичную диаграмму (Violin plot).

https://www.kaggle.com/mathan/fifa-2018-match-statistics

[1]:
```
import numpy as np
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
```

[2]:
```
data = pd.read_csv("data/fifa.csv", sep=',')
data.dtypes
```

[2]:
```
Date                          object
Team                          object
Opponent                      object
Goal Scored                    int64
Ball Possession %              int64
Attempts                       int64
On-Target                      int64
Off-Target                     int64
Blocked                        int64
Corners                        int64
Offsides                       int64
Free Kicks                     int64
Saves                          int64
Pass Accuracy %                int64
Passes                         int64
Distance Covered (Kms)         int64
```

```
Fouls Committed             int64
Yellow Card                 int64
Yellow & Red                int64
Red                         int64
Man of the Match           object
1st Goal                  float64
Round                      object
PSO                        object
Goals in PSO                int64
Own goals                 float64
Own goal Time             float64
dtype: object
```

[3]:
```
data.shape
```

[3]:
```
(128, 27)
```

[4]:
```
data.isnull().sum()
```

[4]:
```
Date                      0
Team                      0
Opponent                  0
Goal Scored               0
Ball Possession %         0
Attempts                  0
On-Target                 0
Off-Target                0
Blocked                   0
Corners                   0
Offsides                  0
Free Kicks                0
Saves                     0
Pass Accuracy %           0
Passes                    0
Distance Covered (Kms)    0
```

```
Fouls Committed             0
Yellow Card                 0
Yellow & Red                0
Red                         0
Man of the Match            0
1st Goal                   34
Round                       0
PSO                         0
Goals in PSO                0
Own goals                 116
Own goal Time             116
dtype: int64
```

[5]:
```
data.head()
```

[5]:
```
         Date          Team      Opponent  Goal Scored  Ball
↪Possession %  \
0  14-06-2018        Russia  Saudi Arabia            5
↪    40
1  14-06-2018  Saudi Arabia        Russia            0
↪    60
2  15-06-2018         Egypt       Uruguay            0
↪    43
3  15-06-2018       Uruguay         Egypt            1
↪    57
4  15-06-2018       Morocco          Iran            0
↪    64

   Attempts  On-Target  Off-Target  Blocked  Corners  …  Yellow
↪Card  \
0        13          7           3        3        6  …
↪0
1         6          0           3        3        2  …
↪0
```

```
2          8           3           3         2         0  …           
↪2
3         14           4           6         4         5  …           
↪0
4         13           3           6         4         5  …           
↪1

   Yellow & Red  Red  Man of the Match  1st Goal        Round  PSO
↪ \
0             0    0              Yes      12.0  Group Stage   No
1             0    0               No       NaN  Group Stage   No
2             0    0               No       NaN  Group Stage   No
3             0    0              Yes      89.0  Group Stage   No
4             0    0               No       NaN  Group Stage   No

   Goals in PSO  Own goals  Own goal Time
0             0        NaN            NaN
1             0        NaN            NaN
2             0        NaN            NaN
3             0        NaN            NaN
4             0        1.0           90.0

[5 rows x 27 columns]
```

0.1.1 Обработка пропусков количественного столбца

[6]:
```
data['Own goals'] = data['Own goals'].fillna(0.0)
```

[7]:
```
data.head()
```

[7]:
```
         Date          Team      Opponent  Goal Scored  Ball
↪Possession %  \
0  14-06-2018        Russia  Saudi Arabia            5           
↪    40
```

```
1  14-06-2018  Saudi Arabia        Russia             0             
 ↪    60
2  15-06-2018         Egypt       Uruguay            0             
 ↪    43
3  15-06-2018       Uruguay         Egypt            1             
 ↪    57
4  15-06-2018       Morocco          Iran            0             
 ↪    64

   Attempts  On-Target  Off-Target  Blocked  Corners  …  Yellow
 ↪Card  \
0        13          7           3        3        6  …           
 ↪0
1         6          0           3        3        2  …           
 ↪0
2         8          3           3        2        0  …           
 ↪2
3        14          4           6        4        5  …           
 ↪0
4        13          3           6        4        5  …           
 ↪1

   Yellow & Red  Red  Man of the Match  1st Goal        Round  PSO
 ↪ \
0             0    0               Yes      12.0  Group Stage   No
1             0    0                No       NaN  Group Stage   No
2             0    0                No       NaN  Group Stage   No
3             0    0               Yes      89.0  Group Stage   No
4             0    0                No       NaN  Group Stage   No

   Goals in PSO  Own goals  Own goal Time
0             0        0.0            NaN
1             0        0.0            NaN
```

```
2              0        0.0            NaN
3              0        0.0            NaN
4              0        1.0           90.0

[5 rows x 27 columns]
```

[8]:
```
data.isnull().sum()
```

[8]:
```
Date                        0
Team                        0
Opponent                    0
Goal Scored                 0
Ball Possession %           0
Attempts                    0
On-Target                   0
Off-Target                  0
Blocked                     0
Corners                     0
Offsides                    0
Free Kicks                  0
Saves                       0
Pass Accuracy %             0
Passes                      0
Distance Covered (Kms)      0
Fouls Committed             0
Yellow Card                 0
Yellow & Red                0
Red                         0
Man of the Match            0
1st Goal                   34
Round                       0
PSO                         0
Goals in PSO                0
Own goals                   0
```

```
Own goal Time             116
dtype: int64
```

0.1.2 Обработка пропусков категориального столбца

Т.к. У нас нет катеригориальных признаков в датасете с пропусками, сделаем их сами. Здесь заполняем пустым значением случайным образом строки столбца ‘Round’, с коэффициентом 30%

[9]:
```
data['Round'].unique()
```

[9]:
```
array(['Group Stage', 'Round of 16', 'Quarter Finals', 'Semi-
↪Finals',
       '3rd Place', 'Final'], dtype=object)
```

[10]:
```
import random
data_col = data['Round'].copy()
for i, item in enumerate(data_col):
    if random.randint(0,100) > 70:
        data_col[i] = np.nan
data['Round'] = data_col
```

[11]:
```
data.isnull().sum()
```

[11]:
```
Date                     0
Team                     0
Opponent                 0
Goal Scored              0
Ball Possession %        0
Attempts                 0
On-Target                0
Off-Target               0
Blocked                  0
Corners                  0
Offsides                 0
Free Kicks               0
```

```
Saves                          0
Pass Accuracy %                0
Passes                         0
Distance Covered (Kms)         0
Fouls Committed                0
Yellow Card                    0
Yellow & Red                   0
Red                            0
Man of the Match               0
1st Goal                      34
Round                         40
PSO                            0
Goals in PSO                   0
Own goals                      0
Own goal Time                116
dtype: int64
```

[12]:
```
data.head()
```

[12]:
```
         Date          Team      Opponent  Goal Scored  Ball
↪Possession %  \
0  14-06-2018        Russia  Saudi Arabia            5
↪    40
1  14-06-2018  Saudi Arabia        Russia            0
↪    60
2  15-06-2018         Egypt       Uruguay            0
↪    43
3  15-06-2018       Uruguay         Egypt            1
↪    57
4  15-06-2018       Morocco          Iran            0
↪    64

  Attempts  On-Target  Off-Target  Blocked  Corners  …  Yellow
↪Card  \
```

```
0        13          7           3        3        6  …           
  ↪0
1         6          0           3        3        2  …           
  ↪0
2         8          3           3        2        0  …           
  ↪2
3        14          4           6        4        5  …           
  ↪0
4        13          3           6        4        5  …           
  ↪1

   Yellow & Red  Red  Man of the Match  1st Goal        Round  PSO
  ↪ \
0             0    0               Yes      12.0          NaN   No
1             0    0                No       NaN  Group Stage   No
2             0    0                No       NaN          NaN   No
3             0    0               Yes      89.0          NaN   No
4             0    0                No       NaN  Group Stage   No

   Goals in PSO  Own goals  Own goal Time
0             0        0.0            NaN
1             0        0.0            NaN
2             0        0.0            NaN
3             0        0.0            NaN
4             0        1.0           90.0

[5 rows x 27 columns]
```

[13]:
```
from sklearn.impute import SimpleImputer
```

[14]:
```
imp = SimpleImputer(missing_values=np.nan,
  ↪strategy='most_frequent')
data['Round'] = imp.fit_transform(data[['Round']])
```

[15]:
```
data.isnull().sum()
```

[15]:
```
Date                          0
Team                          0
Opponent                      0
Goal Scored                   0
Ball Possession %             0
Attempts                      0
On-Target                     0
Off-Target                    0
Blocked                       0
Corners                       0
Offsides                      0
Free Kicks                    0
Saves                         0
Pass Accuracy %               0
Passes                        0
Distance Covered (Kms)        0
Fouls Committed               0
Yellow Card                   0
Yellow & Red                  0
Red                           0
Man of the Match              0
1st Goal                     34
Round                         0
PSO                           0
Goals in PSO                  0
Own goals                     0
Own goal Time               116
dtype: int64
```

[16]:
```
corr = data.corr()
corr
```

[16]:

	Goal Scored	Ball Possession %	Attempts	On-Target
Goal Scored	1.000000	0.034759	0.144915	0.461702
Ball Possession %	0.034759	1.000000	0.541185	0.297234
Attempts	0.144915	0.541185	1.000000	0.731243
On-Target	0.461702	0.297234	0.731243	1.000000
Off-Target	-0.020374	0.361767	0.718972	0.324672
Blocked	-0.087072	0.521510	0.754307	0.331333
Corners	0.040446	0.542992	0.686892	0.407576
Offsides	0.045105	0.057706	-0.016508	0.073176
Free Kicks	0.046815	0.273831	0.140850	0.093090
Saves	-0.118893	-0.293658	-0.268217	-0.321557
Pass Accuracy %	0.135688	0.713872	0.397614	0.291659
Passes	0.043971	0.880611	0.582831	0.348099
Distance Covered (Kms)	0.014355	-0.059054	0.171381	0.065475
Fouls Committed	0.030331	-0.296477	-0.248773	-0.192242
Yellow Card	-0.048838	-0.205511	-0.185544	-0.115259

```
Yellow & Red          -0.035031      0.090924 -0.074594 
↪-0.051742
Red                   -0.089714      0.024316  0.009795 
↪-0.023439
1st Goal              -0.272170     -0.048316  0.072737 
↪-0.071730
Goals in PSO          -0.011204     -0.010086  0.149836 
↪0.040322
Own goals             -0.066164      0.059499 -0.041677 
↪-0.071880
Own goal Time         -0.228729      0.588196 -0.012727 
↪-0.328175

                      Off-Target   Blocked   Corners  Offsides 
↪Free Kicks  \
Goal Scored            -0.020374 -0.087072  0.040446  0.045105 
↪ 0.046815
Ball Possession %      0.361767  0.521510  0.542992  0.057706 
↪ 0.273831
Attempts               0.718972  0.754307  0.686892 -0.016508 
↪ 0.140850
On-Target              0.324672  0.331333  0.407576  0.073176 
↪ 0.093090
Off-Target             1.000000  0.299712  0.440633 -0.095919 
↪ 0.142367
Blocked                0.299712  1.000000  0.636172 -0.002231 
↪ 0.086021
Corners                0.440633  0.636172  1.000000 -0.034054 
↪ 0.085216
Offsides              -0.095919 -0.002231 -0.034054  1.000000 
↪ 0.089121
Free Kicks             0.142367  0.086021  0.085216  0.089121 
↪ 1.000000
```

```
Saves                    -0.126644 -0.142960 -0.233787  0.006539 
 ↪-0.231637
Pass Accuracy %           0.189760  0.401699  0.330363  0.127421 
 ↪ 0.131951
Passes                    0.398949  0.532913  0.524661  0.034715 
 ↪ 0.175695
Distance Covered (Kms)    0.229930  0.082111  0.099700  0.031324 
 ↪ 0.076927
Fouls Committed          -0.186858 -0.174606 -0.165382 -0.040084 
 ↪ 0.080341
Yellow Card              -0.124094 -0.163111 -0.169929 -0.047757 
 ↪-0.026741
Yellow & Red             -0.145577  0.033717  0.014543  0.228531 
 ↪ 0.029701
Red                      -0.119332  0.165295  0.117960 -0.036432 
 ↪-0.104164
1st Goal                  0.109203  0.091321  0.163760 -0.112602 
 ↪-0.007801
Goals in PSO              0.185807  0.108220  0.088997 -0.064892 
 ↪ 0.138046
Own goals                -0.092474  0.074879 -0.028876 -0.070461 
 ↪ 0.018867
Own goal Time             0.300204  0.014851  0.023942 -0.312680 
 ↪-0.057103

                           Saves  …    Passes  Distance Covered
 ↪(Kms)  \
Goal Scored           -0.118893  …  0.043971                0.
 ↪014355
Ball Possession %     -0.293658  …  0.880611               -0.
 ↪059054
Attempts              -0.268217  …  0.582831                0.
 ↪171381
```

```
On-Target                -0.321557  …  0.348099                 0.
↪065475
Off-Target               -0.126644  …  0.398949                 0.
↪229930
Blocked                  -0.142960  …  0.532913                 0.
↪082111
Corners                  -0.233787  …  0.524661                 0.
↪099700
Offsides                  0.006539  …  0.034715                 0.
↪031324
Free Kicks               -0.231637  …  0.175695                 0.
↪076927
Saves                     1.000000  … -0.264425                 0.
↪125645
Pass Accuracy %          -0.190740  …  0.693113                -0.
↪210874
Passes                   -0.264425  …  1.000000                 0.
↪184601
Distance Covered (Kms)    0.125645  …  0.184601                 1.
↪000000
Fouls Committed           0.074976  … -0.352782                 0.
↪149800
Yellow Card               0.009670  … -0.223882                 0.
↪001971
Yellow & Red              0.140307  …  0.053837                -0.
↪023296
Red                       0.016875  …  0.010752                -0.
↪093268
1st Goal                 -0.127567  … -0.093453                -0.
↪079939
Goals in PSO              0.110014  …  0.226844                 0.
↪720547
```

```
Own goals                0.069340  … -0.031464                0.
↪011522
Own goal Time           -0.063865  …  0.296754               -0.
↪313181

                      Fouls Committed  Yellow Card  Yellow & Red
↪      Red  \
Goal Scored                  0.030331    -0.048838     -0.035031
↪-0.089714
Ball Possession %           -0.296477    -0.205511      0.090924
↪ 0.024316
Attempts                    -0.248773    -0.185544     -0.074594
↪ 0.009795
On-Target                   -0.192242    -0.115259     -0.051742
↪-0.023439
Off-Target                  -0.186858    -0.124094     -0.145577
↪-0.119332
Blocked                     -0.174606    -0.163111      0.033717
↪ 0.165295
Corners                     -0.165382    -0.169929      0.014543
↪ 0.117960
Offsides                    -0.040084    -0.047757      0.228531
↪-0.036432
Free Kicks                   0.080341    -0.026741      0.029701
↪-0.104164
Saves                        0.074976     0.009670      0.140307
↪ 0.016875
Pass Accuracy %             -0.327737    -0.113498      0.094756
↪ 0.009492
Passes                      -0.352782    -0.223882      0.053837
↪ 0.010752
Distance Covered (Kms)       0.149800     0.001971     -0.023296
↪-0.093268
```

```
Fouls Committed                1.000000     0.433834      0.039790
 ↪ 0.012408
Yellow Card                    0.433834     1.000000     -0.114064
 ↪ 0.029075
Yellow & Red                   0.039790    -0.114064      1.000000
 ↪-0.015873
Red                            0.012408     0.029075     -0.015873
 ↪ 1.000000
1st Goal                       0.131749     0.268910      0.036355
 ↪-0.001947
Goals in PSO                   0.126270     0.102478     -0.031834
 ↪-0.031834
Own goals                      0.113231     0.053925      0.175596
 ↪-0.040522
Own goal Time                  0.002941     0.319416     -0.239862
 ↪      NaN

                      1st Goal  Goals in PSO  Own goals  Own
 ↪goal Time
Goal Scored          -0.272170     -0.011204  -0.066164      -0.
 ↪228729
Ball Possession %    -0.048316     -0.010086   0.059499       0.
 ↪588196
Attempts              0.072737      0.149836  -0.041677      -0.
 ↪012727
On-Target            -0.071730      0.040322  -0.071880      -0.
 ↪328175
Off-Target            0.109203      0.185807  -0.092474       0.
 ↪300204
Blocked               0.091321      0.108220   0.074879       0.
 ↪014851
Corners               0.163760      0.088997  -0.028876       0.
 ↪023942
```

```
Offsides                -0.112602     -0.064892  -0.070461      -0.
↪312680
Free Kicks              -0.007801      0.138046   0.018867      -0.
↪057103
Saves                   -0.127567      0.110014   0.069340      -0.
↪063865
Pass Accuracy %         -0.084045     -0.143743   0.042372       0.
↪390021
Passes                  -0.093453      0.226844  -0.031464       0.
↪296754
Distance Covered (Kms) -0.079939      0.720547   0.011522      -0.
↪313181
Fouls Committed          0.131749      0.126270   0.113231       0.
↪002941
Yellow Card              0.268910      0.102478   0.053925       0.
↪319416
Yellow & Red             0.036355     -0.031834   0.175596      -0.
↪239862
Red                     -0.001947     -0.031834  -0.040522
↪   NaN
1st Goal                 1.000000     -0.030558   0.017958       0.
↪332835
Goals in PSO            -0.030558      1.000000   0.052096      -0.
↪355415
Own goals                0.017958      0.052096   1.000000
↪   NaN
Own goal Time            0.332835     -0.355415        NaN       1.
↪000000

[21 rows x 21 columns]
```

0.1.3 Столбцы, которые можно использовать для машинного обучения (для разных целевых признаков)

[17]:
```
from IPython.display import Markdown

def corr_values(series, feature):
    for k, v in series[feature].iteritems():
        if k != feature:
            yield (k, v)

for target in corr:
    good = [
        k
        for k, v in corr_values(corr, target)
        if v > 0.5
    ]

    if len(good) != 0:
        display(Markdown(f"- Хорошо коррелируют с целевым
 ↪признаком ({target})): {', '.join(good)}"))
```

- Хорошо коррелируют с целевым признаком (Ball Possession %)): Attempts, Blocked, Corners, Pass Accuracy %, Passes, Own goal Time
- Хорошо коррелируют с целевым признаком (Attempts)): Ball Possession %, On-Target, Off-Target, Blocked, Corners, Passes
- Хорошо коррелируют с целевым признаком (On-Target)): Attempts
- Хорошо коррелируют с целевым признаком (Off-Target)): Attempts
- Хорошо коррелируют с целевым признаком (Blocked)): Ball Possession %, Attempts, Corners, Passes
- Хорошо коррелируют с целевым признаком (Corners)): Ball Possession %, Attempts, Blocked, Passes
- Хорошо коррелируют с целевым признаком (Pass Accuracy %)): Ball Possession %, Passes

- Хорошо коррелируют с целевым признаком (Passes)): Ball Possession %, Attempts, Blocked, Corners, Pass Accuracy %
- Хорошо коррелируют с целевым признаком (Distance Covered (Kms))): Goals in PSO
- Хорошо коррелируют с целевым признаком (Goals in PSO)): Distance Covered (Kms)
- Хорошо коррелируют с целевым признаком (Own goal Time)): Ball Possession %

0.1.4 Violin plot

[18]:
```
import seaborn as sb
sb.violinplot(x=data["Own goals"]);
```