ИУ5-65Б Погосян С.Л. РК1 Вар №16
Номер задачи - 2, номер набора данных - 8

Рубежный контроль №1 Задание 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?

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

Дополнительное требование:
Для студентов групп ИУ5-65Б — для набора данных построить «парные диаграммы»

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

0.0.1. Загружаем данные:

[2]:
```
data = pd.read_csv('FIFA_2018.csv', sep=",")
```

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

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

[4]:
```
#
data.dtypes
```

[4]:
```
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
```

[5]:

```
#
data.isnull().sum()
```

[5]:

```
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
```

[6]:

```
#      10
data.head(10)
```

```
[6]:          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
     5  15-06-2018          Iran       Morocco            1               ␣
   ↪36
     6  15-06-2018      Portugal         Spain            3               ␣
   ↪39
     7  15-06-2018         Spain      Portugal            3               ␣
   ↪61
     8  16-06-2018        France     Australia            2               ␣
   ↪51
     9  16-06-2018     Australia        France            1               ␣
   ↪49

        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
     5         8          2           5        1        2  …            3
     6         8          3           2        3        4  …            1
     7        12          5           5        2        5  …            1
     8        12          5           4        3        5  …            1
     9         4          1           2        1        1  …            3

        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
     5             0    0               Yes      90.0  Group Stage   No
     6             0    0                No       4.0  Group Stage   No
     7             0    0               Yes      24.0  Group Stage   No
     8             0    0               Yes      58.0  Group Stage   No
     9             0    0                No      62.0  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              0        NaN            NaN
6              0        NaN            NaN
7              0        NaN            NaN
8              0        NaN            NaN
9              0        1.0           81.0

[10 rows x 27 columns]
```

[7]:

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

```
      : 128
```

0.0.2. Обработка пропуков в числовых данных

[8]:

```
#
#
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))
```

```
     1st Goal.        float64.                 34, 26.56%.
     Own goals.        float64.                 116, 90.62%.
     Own goal Time.        float64.                 116,
90.62%.
```

В колонках Own goals и Own goals содержится информация о наличии забитых голов в свои ворота и времени, когда это было сделано. Будем считать, что отсутствие информации говорит о том, что не было мячей забитых в свои ворота. Поэтому заполним пропуски в этих колонках нулями.

[9]:

```
#                         "Own goals"
data_new_3 = data[['Own goals']].fillna(0)
```

```
[10]: data[['Own goals']]=data_new_3
```

```
[11]: #                         "Own goal Time"
      data_new_4 = data[['Own goal Time']].fillna(0)
```

```
[12]: data[['Own goal Time']]=data_new_4
```

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

```
[13]:          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            0.0
      1             0        0.0            0.0
      2             0        0.0            0.0
      3             0        0.0            0.0
      4             0        1.0           90.0

      [5 rows x 27 columns]
```

В колонке 1st goal содержится информация о времени от начала игры, в которое был забит первый гол. Поэтому заполним пропуски нулями в тех строках, где значение

“Goal Scored” ранво нулю.

[14]:
```
data.loc[(data['Goal Scored'] == 0), '1st Goal'] = 0
```

Заполним оставшиеся пропуски в столбце “1st Goal” средним значением по столбцу.

[15]:
```
data[['1st Goal']].describe()
```

[15]:
```
         1st Goal
count  127.000000
mean    29.204724
std     27.289552
min      0.000000
25%      0.000000
50%     24.000000
75%     51.000000
max     90.000000
```

[16]:
```
res = np.where(np.isnan(data['1st Goal']), np.ma.array(data['1st Goal'],

               mask = np.isnan(data['1st Goal'])).mean(axis = 0),␣
 ↪data['1st Goal'])
```

[17]:
```
data['1st Goal']=res
```

[18]:
```
data.head()
```

[18]:
```
         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       0.0  Group Stage   No
2             0    0                No       0.0  Group Stage   No
3             0    0               Yes      89.0  Group Stage   No
4             0    0                No       0.0  Group Stage   No

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

[5 rows x 27 columns]
```

[19]:

```
#
data.isnull().sum()
```

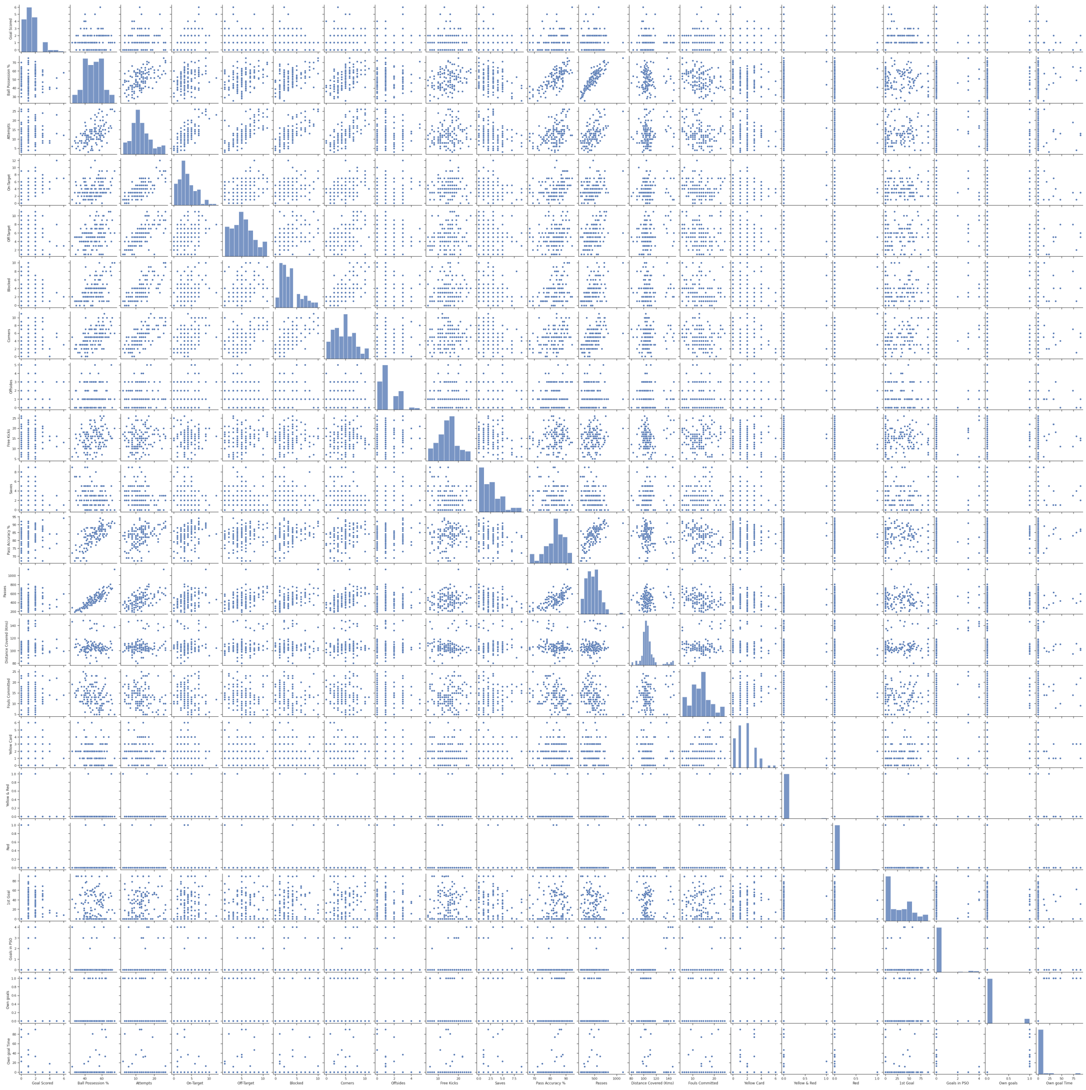
[19]:

```
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                  0
Round                     0
PSO                       0
Goals in PSO              0
Own goals                 0
Own goal Time             0
dtype: int64
```

1. Парные диаграммы

[20]:
```
sns.pairplot(data)
```

[20]:
```
<seaborn.axisgrid.PairGrid at 0x7f658236d0a0>
```



[ ]:
```

```