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)

ФАКУЛЬТЕТ Информатика и системы управления

КАФЕДРА Системы обработки информации и управления

Отчёт по рубежному контролю №1

По дисциплине:
«Технологии машинного обучения»

Выполнил:

Студент группы ИУ5 _____ **Забелина В.А.**

(Подпись, дата) (Фамилия И.О.)

Проверил:

_____ **Гапанюк Ю. Е.**

(Подпись, дата) (Фамилия И.О.)

Москва, 2021

Вариант №9

Задача №2.

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? Для пары произвольных колонок данных построить график "Диаграмма рассеяния".

Набор данных

https://www.kaggle.com/karangadiya/fifa19

Решение

Ввод [1]:
```
import numpy as np
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
from pandas.plotting import scatter_matrix
import warnings
warnings.filterwarnings('ignore')
sns.set(style="ticks")
%matplotlib inline
```

Ввод [2]:
```
data = pd.read_csv('data.csv')
```

Ввод [3]:
```
data.head()
```

Out[3]:

	Unnamed: 0	ID	Name	Age	Photo	Nationality	Flag	Overall	Potential	Club	..
0	0	158023	L. Messi	31	https://cdn.sofifa.org/players/4/19/158023.png	Argentina	https://cdn.sofifa.org/flags/52.png	94	94	FC Barcelona	..
1	1	20801	Cristiano Ronaldo	33	https://cdn.sofifa.org/players/4/19/20801.png	Portugal	https://cdn.sofifa.org/flags/38.png	94	94	Juventus	..
2	2	190871	Neymar Jr	26	https://cdn.sofifa.org/players/4/19/190871.png	Brazil	https://cdn.sofifa.org/flags/54.png	92	93	Paris Saint-Germain	..
3	3	193080	De Gea	27	https://cdn.sofifa.org/players/4/19/193080.png	Spain	https://cdn.sofifa.org/flags/45.png	91	93	Manchester United	..
4	4	192985	K. De Bruyne	27	https://cdn.sofifa.org/players/4/19/192985.png	Belgium	https://cdn.sofifa.org/flags/7.png	91	92	Manchester City	..

5 rows × 89 columns

Ввод [4]:
```
data.dtypes
```

Out[4]:
```
Unnamed: 0          int64
ID                  int64
Name               object
Age                 int64
Photo              object
                    ...
GKHandling        float64
GKKicking         float64
GKPositioning     float64
GKReflexes        float64
Release Clause     object
Length: 89, dtype: object
```

Ввод [5]:
```
data.isnull().sum()
# проверим есть ли пропущенные значения
```

Out[5]:
```
Unnamed: 0           0
ID                   0
Name                 0
Age                  0
Photo                0
                   ...
GKHandling          48
GKKicking           48
GKPositioning       48
GKReflexes          48
Release Clause    1564
Length: 89, dtype: int64
```

Ввод [6]:
```
data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 18207 entries, 0 to 18206
Data columns (total 89 columns):
 #   Column                    Non-Null Count  Dtype
---  ------                    --------------  -----
 0   Unnamed: 0                18207 non-null  int64
 1   ID                        18207 non-null  int64
 2   Name                      18207 non-null  object
 3   Age                       18207 non-null  int64
 4   Photo                     18207 non-null  object
 5   Nationality               18207 non-null  object
 6   Flag                      18207 non-null  object
 7   Overall                   18207 non-null  int64
 8   Potential                 18207 non-null  int64
 9   Club                      17966 non-null  object
 10  Club Logo                 18207 non-null  object
 11  Value                     18207 non-null  object
 12  Wage                      18207 non-null  object
 13  Special                   18207 non-null  int64
 14  Preferred Foot            18159 non-null  object
 15  International Reputation  18159 non-null  float64
 16  Weak Foot                 18159 non-null  float64
 17  Skill Moves               18159 non-null  float64
 18  Work Rate                 18159 non-null  object
 19  Body Type                 18159 non-null  object
 20  Real Face                 18159 non-null  object
 21  Position                  18147 non-null  object
 22  Jersey Number             18147 non-null  float64
 23  Joined                    16654 non-null  object
 24  Loaned From               1264 non-null   object
 25  Contract Valid Until      17918 non-null  object
 26  Height                    18159 non-null  object
 27  Weight                    18159 non-null  object
 28  LS                        16122 non-null  object
 29  ST                        16122 non-null  object
 30  RS                        16122 non-null  object
 31  LW                        16122 non-null  object
 32  LF                        16122 non-null  object
 33  CF                        16122 non-null  object
 34  RF                        16122 non-null  object
 35  RW                        16122 non-null  object
 36  LAM                       16122 non-null  object
 37  CAM                       16122 non-null  object
 38  RAM                       16122 non-null  object
 39  LM                        16122 non-null  object
 40  LCM                       16122 non-null  object
 41  CM                        16122 non-null  object
 42  RCM                       16122 non-null  object
 43  RM                        16122 non-null  object
 44  LWB                       16122 non-null  object
 45  LDM                       16122 non-null  object
 46  CDM                       16122 non-null  object
 47  RDM                       16122 non-null  object
 48  RWB                       16122 non-null  object
 49  LB                        16122 non-null  object
 50  LCB                       16122 non-null  object
 51  CB                        16122 non-null  object
 52  RCB                       16122 non-null  object
 53  RB                        16122 non-null  object
 54  Crossing                  18159 non-null  float64
 55  Finishing                 18159 non-null  float64
 56  HeadingAccuracy           18159 non-null  float64
 57  ShortPassing              18159 non-null  float64
 58  Volleys                   18159 non-null  float64
 59  Dribbling                 18159 non-null  float64
 60  Curve                     18159 non-null  float64
 61  FKAccuracy                18159 non-null  float64
 62  LongPassing               18159 non-null  float64
 63  BallControl               18159 non-null  float64
 64  Acceleration              18159 non-null  float64
 65  SprintSpeed               18159 non-null  float64
 66  Agility                   18159 non-null  float64
 67  Reactions                 18159 non-null  float64
 68  Balance                   18159 non-null  float64
 69  ShotPower                 18159 non-null  float64
 70  Jumping                   18159 non-null  float64
 71  Stamina                   18159 non-null  float64
 72  Strength                  18159 non-null  float64
 73  LongShots                 18159 non-null  float64
 74  Aggression                18159 non-null  float64
 75  Interceptions             18159 non-null  float64
 76  Positioning               18159 non-null  float64
 77  Vision                    18159 non-null  float64
```

```
 76  Positioning               18159 non-null  float64
 77  Vision                    18159 non-null  float64
 78  Penalties                 18159 non-null  float64
 79  Composure                 18159 non-null  float64
 80  Marking                   18159 non-null  float64
 81  StandingTackle            18159 non-null  float64
 82  SlidingTackle             18159 non-null  float64
 83  GKDiving                  18159 non-null  float64
 84  GKHandling                18159 non-null  float64
 85  GKKicking                 18159 non-null  float64
 86  GKPositioning             18159 non-null  float64
 87  GKReflexes                18159 non-null  float64
 88  Release Clause            16643 non-null  object
dtypes: float64(38), int64(6), object(45)
memory usage: 9.2+ MB
```

Ввод [7]:
```
# Удаляем столбцы, которые не несут значимой информации
data.drop(['Photo', 'Age'], axis = 1, inplace = True)
```

Ввод [8]:
```
data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 18207 entries, 0 to 18206
Data columns (total 87 columns):
 #   Column                    Non-Null Count  Dtype
---  ------                    --------------  -----
 0   Unnamed: 0                18207 non-null  int64
 1   ID                        18207 non-null  int64
 2   Name                      18207 non-null  object
 3   Nationality               18207 non-null  object
 4   Flag                      18207 non-null  object
 5   Overall                   18207 non-null  int64
 6   Potential                 18207 non-null  int64
 7   Club                      17966 non-null  object
 8   Club Logo                 18207 non-null  object
 9   Value                     18207 non-null  object
 10  Wage                      18207 non-null  object
 11  Special                   18207 non-null  int64
 12  Preferred Foot            18159 non-null  object
 13  International Reputation  18159 non-null  float64
 14  Weak Foot                 18159 non-null  float64
 15  Skill Moves               18159 non-null  float64
 16  Work Rate                 18159 non-null  object
 17  Body Type                 18159 non-null  object
 18  Real Face                 18159 non-null  object
```

```
 47  LB                        16122 non-null  object
 48  LCB                       16122 non-null  object
 49  CB                        16122 non-null  object
 50  RCB                       16122 non-null  object
 51  RB                        16122 non-null  object
 52  Crossing                  18159 non-null  float64
 53  Finishing                 18159 non-null  float64
 54  HeadingAccuracy           18159 non-null  float64
 55  ShortPassing              18159 non-null  float64
 56  Volleys                   18159 non-null  float64
 57  Dribbling                 18159 non-null  float64
 58  Curve                     18159 non-null  float64
 59  FKAccuracy                18159 non-null  float64
 60  LongPassing               18159 non-null  float64
 61  BallControl               18159 non-null  float64
 62  Acceleration              18159 non-null  float64
 63  SprintSpeed               18159 non-null  float64
 64  Agility                   18159 non-null  float64
 65  Reactions                 18159 non-null  float64
 66  Balance                   18159 non-null  float64
 67  ShotPower                 18159 non-null  float64
 68  Jumping                   18159 non-null  float64
 69  Stamina                   18159 non-null  float64
 70  Strength                  18159 non-null  float64
 71  LongShots                 18159 non-null  float64
 72  Aggression                18159 non-null  float64
 73  Interceptions             18159 non-null  float64
 74  Positioning               18159 non-null  float64
 75  Vision                    18159 non-null  float64
 76  Penalties                 18159 non-null  float64
 77  Composure                 18159 non-null  float64
 78  Marking                   18159 non-null  float64
 79  StandingTackle            18159 non-null  float64
 80  SlidingTackle             18159 non-null  float64
 81  GKDiving                  18159 non-null  float64
 82  GKHandling                18159 non-null  float64
 83  GKKicking                 18159 non-null  float64
 84  GKPositioning             18159 non-null  float64
 85  GKReflexes                18159 non-null  float64
 86  Release Clause            16643 non-null  object
dtypes: float64(38), int64(5), object(44)
memory usage: 9.0+ MB
```

```
memory usage: 9.0+ MB
```

Ввод [9]:
```
# Заполняем отсутствующие значения
data['GKReflexes'] = data['GKReflexes'].replace(0,np.nan)
data['GKReflexes'] = data['GKReflexes'].fillna(data['GKReflexes'].mean())
```

Ввод [10]:
```
data.head()
```

Out[10]:

	Unnamed: 0	ID	Name	Nationality	Flag	Overall	Potential	Club	Club Logo	Value
0	0	158023	L. Messi	Argentina	https://cdn.sofifa.org/flags/52.png	94	94	FC Barcelona	https://cdn.sofifa.org/teams/2/light/241.png	€110.5M
1	1	20801	Cristiano Ronaldo	Portugal	https://cdn.sofifa.org/flags/38.png	94	94	Juventus	https://cdn.sofifa.org/teams/2/light/45.png	€77M
2	2	190871	Neymar Jr	Brazil	https://cdn.sofifa.org/flags/54.png	92	93	Paris Saint-Germain	https://cdn.sofifa.org/teams/2/light/73.png	€118.5M
3	3	193080	De Gea	Spain	https://cdn.sofifa.org/flags/45.png	91	93	Manchester United	https://cdn.sofifa.org/teams/2/light/11.png	€72M
4	4	192985	K. De Bruyne	Belgium	https://cdn.sofifa.org/flags/7.png	91	92	Manchester City	https://cdn.sofifa.org/teams/2/light/10.png	€102M

5 rows × 87 columns

Ввод [11]:
```
data.isnull().sum()
# проверим есть ли пропущенные значения в столбце business_latitude
```

Out[11]:
```
Unnamed: 0            0
ID                    0
Name                  0
Nationality           0
Flag                  0
                   ...
GKHandling           48
GKKicking            48
GKPositioning        48
GKReflexes            0
Release Clause     1564
Length: 87, dtype: int64
```

Ввод [12]:
```
total_count = data.shape[0]
print('Всего строк: {}'.format(total_count))
```

```
Всего строк: 18207
```

Ввод [13]:
```
# Выберем категориальные колонки с пропущенными значениями
# Цикл по колонкам датасета
cat_cols = []
for col in data.columns:
    # Количество пустых значений
    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('Колонка {}. Тип данных {}. Количество пустых значений {}, {}%.'.format(col, dt, temp_null_count, temp_perc))
```

```
Колонка Club. Тип данных object. Количество пустых значений 241, 1.32%.
Колонка Preferred Foot. Тип данных object. Количество пустых значений 48, 0.26%.
Колонка Work Rate. Тип данных object. Количество пустых значений 48, 0.26%.
Колонка Body Type. Тип данных object. Количество пустых значений 48, 0.26%.
Колонка Real Face. Тип данных object. Количество пустых значений 48, 0.26%.
Колонка Position. Тип данных object. Количество пустых значений 60, 0.33%.
Колонка Joined. Тип данных object. Количество пустых значений 1553, 8.53%.
Колонка Loaned From. Тип данных object. Количество пустых значений 16943, 93.06%.
Колонка Contract Valid Until. Тип данных object. Количество пустых значений 289, 1.59%.
Колонка Height. Тип данных object. Количество пустых значений 48, 0.26%.
Колонка Weight. Тип данных object. Количество пустых значений 48, 0.26%.
Колонка LS. Тип данных object. Количество пустых значений 2085, 11.45%.
Колонка ST. Тип данных object. Количество пустых значений 2085, 11.45%.
Колонка RS. Тип данных object. Количество пустых значений 2085, 11.45%.
Колонка LW. Тип данных object. Количество пустых значений 2085, 11.45%.
Колонка LF. Тип данных object. Количество пустых значений 2085, 11.45%.
Колонка CF. Тип данных object. Количество пустых значений 2085, 11.45%.
Колонка RF. Тип данных object. Количество пустых значений 2085, 11.45%.
Колонка RW. Тип данных object. Количество пустых значений 2085, 11.45%.
Колонка LAM. Тип данных object. Количество пустых значений 2085, 11.45%.
Колонка CAM. Тип данных object. Количество пустых значений 2085, 11.45%.
Колонка RAM. Тип данных object. Количество пустых значений 2085, 11.45%.
Колонка LM. Тип данных object. Количество пустых значений 2085, 11.45%.
Колонка LCM. Тип данных object. Количество пустых значений 2085, 11.45%.
Колонка CM. Тип данных object. Количество пустых значений 2085, 11.45%.
Колонка RCM. Тип данных object. Количество пустых значений 2085, 11.45%.
Колонка RM. Тип данных object. Количество пустых значений 2085, 11.45%.
Колонка LWB. Тип данных object. Количество пустых значений 2085, 11.45%.
```

```
Колонка LB. Тип данных object. Количество пустых значений 2085, 11.45%.
Колонка LCB. Тип данных object. Количество пустых значений 2085, 11.45%.
Колонка CB. Тип данных object. Количество пустых значений 2085, 11.45%.
Колонка RCB. Тип данных object. Количество пустых значений 2085, 11.45%.
Колонка RB. Тип данных object. Количество пустых значений 2085, 11.45%.
Колонка Release Clause. Тип данных object. Количество пустых значений 1564, 8.59%.
```

Ввод [14]:

```
# Заполняем отсутствующие значения
data['Release Clause'] = data.fillna("Nane")
data.head()
```

Out[14]:

	Unnamed: 0	ID	Name	Nationality	Flag	Overall	Potential	Club	Club Logo	Value
0	0	158023	L. Messi	Argentina	https://cdn.sofifa.org/flags/52.png	94	94	FC Barcelona	https://cdn.sofifa.org/teams/2/light/241.png	€110.5M
1	1	20801	Cristiano Ronaldo	Portugal	https://cdn.sofifa.org/flags/38.png	94	94	Juventus	https://cdn.sofifa.org/teams/2/light/45.png	€77M
2	2	190871	Neymar Jr	Brazil	https://cdn.sofifa.org/flags/54.png	92	93	Paris Saint-Germain	https://cdn.sofifa.org/teams/2/light/73.png	€118.5M
3	3	193080	De Gea	Spain	https://cdn.sofifa.org/flags/45.png	91	93	Manchester United	https://cdn.sofifa.org/teams/2/light/11.png	€72M
4	4	192985	K. De Bruyne	Belgium	https://cdn.sofifa.org/flags/7.png	91	92	Manchester City	https://cdn.sofifa.org/teams/2/light/10.png	€102M

5 rows × 87 columns

Ввод [15]:

```
data.isnull().sum()
# проверим есть ли пропущенные значения в столбце violation_id
```

Out[15]:

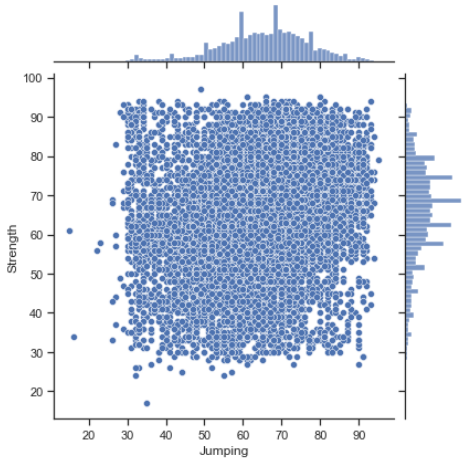
```
Unnamed: 0          0
ID                  0
Name                0
Nationality         0
Flag                0
                   ..
GKHandling         48
GKKicking          48
GKPositioning      48
GKReflexes          0
Release Clause      0
Length: 87, dtype: int64
```

Ввод [19]:

```
# Увеличенные диаграммы рассеяния
sns.jointplot(x = "Jumping", y = "Strength", kind="scatter", data = data)
```

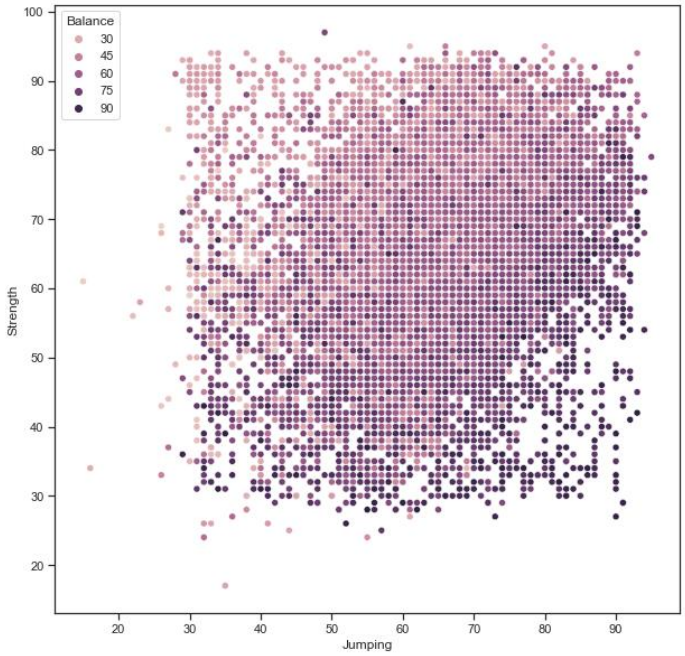
Out[19]:

```
<seaborn.axisgrid.JointGrid at 0x51a3cc70>
```



Ввод [20]:
```
fig, ax = plt.subplots(figsize=(10,10))
sns.scatterplot(ax=ax, x = "Jumping", y = "Strength", data=data, hue='Balance')
```

Out[20]: <AxesSubplot:xlabel='Jumping', ylabel='Strength'>



Ввод [ ]:
```
# The end.
```