ИУ5-61Б Головацкий А. Д. РК №1

Оглавление

Задание (к оглавлению)

> 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".

> Необходимо подготовить отчет по рубежному контролю и разместить его в Вашем репозитории. Вы можете использовать титульный лист, или в начале ноутбука в текстовой ячейке указать Ваши Ф.И.О. и группу.

Описание датасета (к оглавлению)

Датасет `FIFA 19 complete player dataset` создан для футбольной аналитики. Он содержит подробные атрибуты каждого игрока, зарегистрированного в базе данных FIFA 19.

Импорт библиотек (к оглавлению)

In [24]:

```
import numpy as np
import pandas as pd

import seaborn as sns
import matplotlib.pyplot as plt
from sklearn.impute import SimpleImputer
%matplotlib inline
```

Загрузка и первичный анализ данных (к оглавлению)

In [25]:

```
data = pd.read_csv("C:\\Users\\Andrew\\Desktop\\ML-2022\\datasets\\data.csv", sep=",")
data
```

Out[25]:

	Unnamed: 0	ID	Name	Age	Photo	Nation
0	0	158023	L. Messi	31	https://cdn.sofifa.org/players/4/19/158023.png	Arge
1	1	20801	Cristiano Ronaldo	33	https://cdn.sofifa.org/players/4/19/20801.png	Port
2	2	190871	Neymar Jr	26	https://cdn.sofifa.org/players/4/19/190871.png	B
3	3	193080	De Gea	27	https://cdn.sofifa.org/players/4/19/193080.png	S
4	4	192985	K. De Bruyne	27	https://cdn.sofifa.org/players/4/19/192985.png	Bel
...	...	...	...	...	...	
18202	18202	238813	J. Lundstram	19	https://cdn.sofifa.org/players/4/19/238813.png	Eng
18203	18203	243165	N. Christoffersson	19	https://cdn.sofifa.org/players/4/19/243165.png	Swe
18204	18204	241638	B. Worman	16	https://cdn.sofifa.org/players/4/19/241638.png	Eng
18205	18205	246268	D. Walker-Rice	17	https://cdn.sofifa.org/players/4/19/246268.png	Eng
18206	18206	246269	G. Nugent	16	https://cdn.sofifa.org/players/4/19/246269.png	Eng

18207 rows × 89 columns

In [26]:

```
# Переименуем столбцы, чтобы избавиться от пробелов в именах
data = data.rename(columns={
    "Club Logo": "ClubLogo",
    "Preferred Foot": "PreferredFoot",
    "International Reputation": "InternationalReputation",
    "Weak Foot": "WeakFoot",
    "Skill Moves": "SkillMoves",
    "Work Rate": "WorkRate",
    "Body Type": "BodyType",
    "Real Face": "RealFace",
    "Jersey Number": "JerseyNumber",
    "Loaned From": "LoanedFrom",
    "Contract Valid Until": "ContractValidUntil",
    "Release Clause": "ReleaseClause",
})
```

In [27]:

```
data.describe()
```

Out[27]:

	Unnamed: 0	ID	Age	Overall	Potential	Special	Ir
count	18207.000000	18207.000000	18207.000000	18207.000000	18207.000000	18207.000000	
mean	9103.000000	214298.338606	25.122206	66.238699	71.307299	1597.809908	
std	5256.052511	29965.244204	4.669943	6.908930	6.136496	272.586016	
min	0.000000	16.000000	16.000000	46.000000	48.000000	731.000000	
25%	4551.500000	200315.500000	21.000000	62.000000	67.000000	1457.000000	
50%	9103.000000	221759.000000	25.000000	66.000000	71.000000	1635.000000	
75%	13654.500000	236529.500000	28.000000	71.000000	75.000000	1787.000000	
max	18206.000000	246620.000000	45.000000	94.000000	95.000000	2346.000000	

8 rows × 44 columns

In [28]:

```
data.shape
```

Out[28]:

```
(18207, 89)
```

In [29]:

```
data.dtypes
```

Out[29]:

```
Unnamed: 0         int64
ID                 int64
Name              object
Age                int64
Photo             object
                  ...
GKHandling       float64
GKKicking        float64
GKPositioning    float64
GKReflexes       float64
ReleaseClause     object
Length: 89, dtype: object
```

In [30]:

```
# Количество пустых значений
total_count = data.shape[0]
for col in data.columns:
    temp_null_count = data[data[col].isnull()].shape[0]
    temp_perc = round((temp_null_count / total_count) * 100.0, 2)
    print('Колонка {} – {}, {}%'.format(col, temp_null_count, temp_perc))
```

```
Колонка Unnamed: 0 – 0, 0.0%
Колонка ID – 0, 0.0%
Колонка Name – 0, 0.0%
Колонка Age – 0, 0.0%
Колонка Photo – 0, 0.0%
Колонка Nationality – 0, 0.0%
Колонка Flag – 0, 0.0%
Колонка Overall – 0, 0.0%
Колонка Potential – 0, 0.0%
Колонка Club – 241, 1.32%
Колонка ClubLogo – 0, 0.0%
Колонка Value – 0, 0.0%
Колонка Wage – 0, 0.0%
Колонка Special – 0, 0.0%
Колонка PreferredFoot – 48, 0.26%
Колонка InternationalReputation – 48, 0.26%
Колонка WeakFoot – 48, 0.26%
Колонка SkillMoves – 48, 0.26%
Колонка WorkRate – 48, 0.26%
Колонка BodyType – 48, 0.26%
Колонка RealFace – 48, 0.26%
Колонка Position – 60, 0.33%
Колонка JerseyNumber – 60, 0.33%
Колонка Joined – 1553, 8.53%
Колонка LoanedFrom – 16943, 93.06%
Колонка ContractValidUntil – 289, 1.59%
Колонка Height – 48, 0.26%
Колонка Weight – 48, 0.26%
Колонка LS – 2085, 11.45%
Колонка ST – 2085, 11.45%
Колонка RS – 2085, 11.45%
Колонка LW – 2085, 11.45%
Колонка LF – 2085, 11.45%
Колонка CF – 2085, 11.45%
Колонка RF – 2085, 11.45%
Колонка RW – 2085, 11.45%
Колонка LAM – 2085, 11.45%
Колонка CAM – 2085, 11.45%
Колонка RAM – 2085, 11.45%
Колонка LM – 2085, 11.45%
Колонка LCM – 2085, 11.45%
Колонка CM – 2085, 11.45%
Колонка RCM – 2085, 11.45%
Колонка RM – 2085, 11.45%
Колонка LWB – 2085, 11.45%
Колонка LDM – 2085, 11.45%
Колонка CDM – 2085, 11.45%
Колонка RDM – 2085, 11.45%
Колонка RWB – 2085, 11.45%
Колонка LB – 2085, 11.45%
Колонка LCB – 2085, 11.45%
Колонка CB – 2085, 11.45%
```

```
Колонка RCB – 2085, 11.45%
Колонка RB – 2085, 11.45%
Колонка Crossing – 48, 0.26%
Колонка Finishing – 48, 0.26%
Колонка HeadingAccuracy – 48, 0.26%
Колонка ShortPassing – 48, 0.26%
Колонка Volleys – 48, 0.26%
Колонка Dribbling – 48, 0.26%
Колонка Curve – 48, 0.26%
Колонка FKAccuracy – 48, 0.26%
Колонка LongPassing – 48, 0.26%
Колонка BallControl – 48, 0.26%
Колонка Acceleration – 48, 0.26%
Колонка SprintSpeed – 48, 0.26%
Колонка Agility – 48, 0.26%
Колонка Reactions – 48, 0.26%
Колонка Balance – 48, 0.26%
Колонка ShotPower – 48, 0.26%
Колонка Jumping – 48, 0.26%
Колонка Stamina – 48, 0.26%
Колонка Strength – 48, 0.26%
Колонка LongShots – 48, 0.26%
Колонка Aggression – 48, 0.26%
Колонка Interceptions – 48, 0.26%
Колонка Positioning – 48, 0.26%
Колонка Vision – 48, 0.26%
Колонка Penalties – 48, 0.26%
Колонка Composure – 48, 0.26%
Колонка Marking – 48, 0.26%
Колонка StandingTackle – 48, 0.26%
Колонка SlidingTackle – 48, 0.26%
Колонка GKDiving – 48, 0.26%
Колонка GKHandling – 48, 0.26%
Колонка GKKicking – 48, 0.26%
Колонка GKPositioning – 48, 0.26%
Колонка GKReflexes – 48, 0.26%
Колонка ReleaseClause – 1564, 8.59%
```

Проводя визуализацию данных, я бы удалил столбец LoanedFrom, который имеет более 93% пустых значений и к тому же является не очень информативным, а также столбец Unnamed, по сути реализующий лишнюю индексацию элементов.

In [31]:

```
# df = data.drop('LoanedFrom', axis=1)
# df = data.drop('Unnamed', axis=1)
```

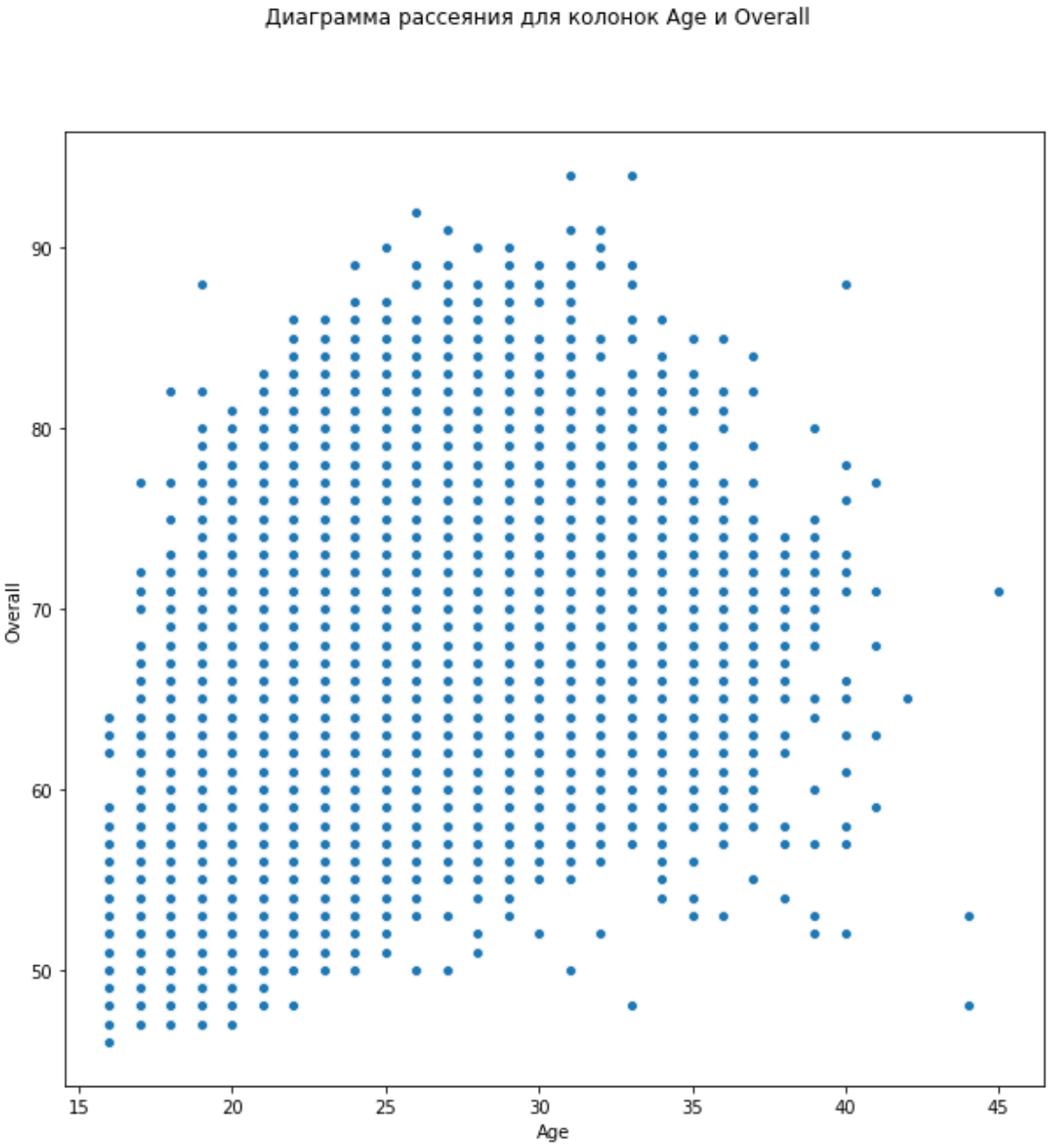
Диаграмма рассеяния для двух стобцов (к оглавлению)

In [32]:

```
fig, ax = plt.subplots(figsize=(10, 10))
fig.suptitle("Диаграмма рассеяния для колонок Age и Overall")
sns.scatterplot(ax=ax, x='Age', y='Overall', data=data)
```

Out[32]:

```
<matplotlib.axes._subplots.AxesSubplot at 0x2ba47fa7848>
```



Обработка пропусков в данных (к оглавлению)

Количественный признак

In [33]:

```
# Импьютация столбца JerseyNumber с помощью медианы

temp_col = SimpleImputer(strategy='median').fit_transform(data[['JerseyNumber']])
data[['JerseyNumber']] = temp_col
```

Категориальный признак

In [34]:

```
# Импьютация константой NA столбца LS

temp_col = SimpleImputer(missing_values=np.nan, strategy='constant', fill_value='NA').fit_t
data[['LS']] = temp_col

# Количество пустых значений
total_count = data.shape[0]
for col in data.columns:
    temp_null_count = data[data[col].isnull()].shape[0]
    temp_perc = round((temp_null_count / total_count) * 100.0, 2)
    print('Колонка {} – {}, {}%'.format(col, temp_null_count, temp_perc))
```

```
Колонка Unnamed: 0 – 0, 0.0%
Колонка ID – 0, 0.0%
Колонка Name – 0, 0.0%
Колонка Age – 0, 0.0%
Колонка Photo – 0, 0.0%
Колонка Nationality – 0, 0.0%
Колонка Flag – 0, 0.0%
Колонка Overall – 0, 0.0%
Колонка Potential – 0, 0.0%
Колонка Club – 241, 1.32%
Колонка ClubLogo – 0, 0.0%
Колонка Value – 0, 0.0%
Колонка Wage – 0, 0.0%
Колонка Special – 0, 0.0%
Колонка PreferredFoot – 48, 0.26%
Колонка InternationalReputation – 48, 0.26%
Колонка WeakFoot – 48, 0.26%
Колонка SkillMoves – 48, 0.26%
Колонка WorkRate – 48, 0.26%
Колонка BodyType – 48, 0.26%
Колонка RealFace – 48, 0.26%
Колонка Position – 60, 0.33%
Колонка JerseyNumber – 0, 0.0%
Колонка Joined – 1553, 8.53%
Колонка LoanedFrom – 16943, 93.06%
Колонка ContractValidUntil – 289, 1.59%
Колонка Height – 48, 0.26%
Колонка Weight – 48, 0.26%
Колонка LS – 0, 0.0%
Колонка ST – 2085, 11.45%
Колонка RS – 2085, 11.45%
Колонка LW – 2085, 11.45%
Колонка LF – 2085, 11.45%
Колонка CF – 2085, 11.45%
Колонка RF – 2085, 11.45%
Колонка RW – 2085, 11.45%
Колонка LAM – 2085, 11.45%
Колонка CAM – 2085, 11.45%
Колонка RAM – 2085, 11.45%
Колонка LM – 2085, 11.45%
Колонка LCM – 2085, 11.45%
Колонка CM – 2085, 11.45%
Колонка RCM – 2085, 11.45%
Колонка RM – 2085, 11.45%
Колонка LWB – 2085, 11.45%
Колонка LDM – 2085, 11.45%
Колонка CDM – 2085, 11.45%
```

```
Колонка RDM – 2085, 11.45%
Колонка RWB – 2085, 11.45%
Колонка LB – 2085, 11.45%
Колонка LCB – 2085, 11.45%
Колонка CB – 2085, 11.45%
Колонка RCB – 2085, 11.45%
Колонка RB – 2085, 11.45%
Колонка Crossing – 48, 0.26%
Колонка Finishing – 48, 0.26%
Колонка HeadingAccuracy – 48, 0.26%
Колонка ShortPassing – 48, 0.26%
Колонка Volleys – 48, 0.26%
Колонка Dribbling – 48, 0.26%
Колонка Curve – 48, 0.26%
Колонка FKAccuracy – 48, 0.26%
Колонка LongPassing – 48, 0.26%
Колонка BallControl – 48, 0.26%
Колонка Acceleration – 48, 0.26%
Колонка SprintSpeed – 48, 0.26%
Колонка Agility – 48, 0.26%
Колонка Reactions – 48, 0.26%
Колонка Balance – 48, 0.26%
Колонка ShotPower – 48, 0.26%
Колонка Jumping – 48, 0.26%
Колонка Stamina – 48, 0.26%
Колонка Strength – 48, 0.26%
Колонка LongShots – 48, 0.26%
Колонка Aggression – 48, 0.26%
Колонка Interceptions – 48, 0.26%
Колонка Positioning – 48, 0.26%
Колонка Vision – 48, 0.26%
Колонка Penalties – 48, 0.26%
Колонка Composure – 48, 0.26%
Колонка Marking – 48, 0.26%
Колонка StandingTackle – 48, 0.26%
Колонка SlidingTackle – 48, 0.26%
Колонка GKDiving – 48, 0.26%
Колонка GKHandling – 48, 0.26%
Колонка GKKicking – 48, 0.26%
Колонка GKPositioning – 48, 0.26%
Колонка GKReflexes – 48, 0.26%
Колонка ReleaseClause – 1564, 8.59%
```

Ответы

Для обработки пропусков категориальных признаков использовалась замена константой. Для количественных признаков использовалась замена медианой.

В дальнейшем для построения моделей следует исключить из датасета признаки Unnamed и LonedFrom, т.к. столбец Unnamed реализует ненужную индексацию, а столбец LonedFrom в большинстве своем представляет пустые значения, что делает его неинформативным. Остальные признаки можно учитывать, они могут оказать большое влияние, большинство из них являются строковыми.