**РК1** **ИУ5-65Б** **Нагдимаев** **Ильягу** **Номер** **варианта** **-** **12**

**Номер** **задачи** **-** **2**

**Номер** **набора** **данных,** **указанного** **в** **задаче** **–** **4**

**Условие** **задания:**

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?

**Дополнительное** **задание:**

Для пары произвольных колонок данных построить график "Парные диаграммы".

**Набор** **данных:**

<https://www.kaggle.com/noriuk/us-education-datasets-unification-project>(файл states\_all.csv)

**Импорт** **библиотек**

In [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

In [2]: data = pd.read\_csv('sample\_data/states\_all.csv')

In [3]: data**.**head()

Out[3]:

**0**

**1**

**2**

**3**

**4**

**PRIMARY\_KEY**

1992\_ALABAMA

1992\_ALASKA

1992\_ARIZONA

1992\_ARKANSAS

1992\_CALIFORNIA

**STATE**

ALABAMA

ALASKA

ARIZONA

ARKANSAS

CALIFORNIA

**YEAR** **ENROLL**

1992 NaN

1992 NaN

1992 NaN

1992 NaN

1992 NaN

**TOTAL\_REVENUE**

2678885.0

1049591.0

3258079.0

1711959.0

26260025.0

**FEDERAL\_REVENUE**

304177.0

106780.0

297888.0

178571.0

2072470.0

**STATE\_REVEN**

16590

7207

13698

9587

165465

5

In [4]:

Out[4]:

data**.**dtypes

PRIMARY\_KEY STATE

YEAR ENROLL

TOTAL\_REVENUE FEDERAL\_REVENUE STATE\_REVENUE LOCAL\_REVENUE TOTAL\_EXPENDITURE INSTRUCTION\_EXPENDITURE

SUPPORT\_SERVICES\_EXPENDITURE OTHER\_EXPENDITURE CAPITAL\_OUTLAY\_EXPENDITURE GRADES\_PK\_G

GRADES\_KG\_G GRADES\_4\_G GRADES\_8\_G GRADES\_12\_G GRADES\_1\_8\_G GRADES\_9\_12\_G GRADES\_ALL\_G AVG\_MATH\_4\_SCORE AVG\_MATH\_8\_SCORE AVG\_READING\_4\_SCORE AVG\_READING\_8\_SCORE dtype: object

object object int64 float64 float64 float64 float64 float64 float64 float64 float64 float64 float64 float64 float64 float64 float64 float64 float64 float64 float64 float64 float64 float64 float64

In [5]: data**.**isnull()**.**sum()

*#* *проверим* *есть* *ли* *пропущенные* *значения*

Out[5]:

In [6]:

PRIMARY\_KEY 0 STATE 0 YEAR 0 ENROLL 491 TOTAL\_REVENUE 440 FEDERAL\_REVENUE 440 STATE\_REVENUE 440 LOCAL\_REVENUE 440 TOTAL\_EXPENDITURE 440 INSTRUCTION\_EXPENDITURE 440 SUPPORT\_SERVICES\_EXPENDITURE 440 OTHER\_EXPENDITURE 491 CAPITAL\_OUTLAY\_EXPENDITURE 440 GRADES\_PK\_G 173 GRADES\_KG\_G 83 GRADES\_4\_G 83 GRADES\_8\_G 83 GRADES\_12\_G 83 GRADES\_1\_8\_G 695 GRADES\_9\_12\_G 644 GRADES\_ALL\_G 83 AVG\_MATH\_4\_SCORE 1150 AVG\_MATH\_8\_SCORE 1113 AVG\_READING\_4\_SCORE 1065 AVG\_READING\_8\_SCORE 1153 dtype: int64

data**.**info()

<class 'pandas.core.frame.DataFrame'> RangeIndex: 1715 entries, 0 to 1714 Data columns (total 25 columns):

# Column Non-Null Count Dtype --- ------ -------------- -----

0 PRIMARY\_KEY 1715 non-null object 1 STATE 1715 non-null object

2 YEAR 1715 non-null 3 ENROLL 1224 non-null 4 TOTAL\_REVENUE 1275 non-null 5 FEDERAL\_REVENUE 1275 non-null 6 STATE\_REVENUE 1275 non-null 7 LOCAL\_REVENUE 1275 non-null 8 TOTAL\_EXPENDITURE 1275 non-null 9 INSTRUCTION\_EXPENDITURE 1275 non-null 10 SUPPORT\_SERVICES\_EXPENDITURE 1275 non-null 11 OTHER\_EXPENDITURE 1224 non-null 12 CAPITAL\_OUTLAY\_EXPENDITURE 1275 non-null 13 GRADES\_PK\_G 1542 non-null 14 GRADES\_KG\_G 1632 non-null 15 GRADES\_4\_G 1632 non-null 16 GRADES\_8\_G 1632 non-null 17 GRADES\_12\_G 1632 non-null 18 GRADES\_1\_8\_G 1020 non-null 19 GRADES\_9\_12\_G 1071 non-null 20 GRADES\_ALL\_G 1632 non-null 21 AVG\_MATH\_4\_SCORE 565 non-null 22 AVG\_MATH\_8\_SCORE 602 non-null 23 AVG\_READING\_4\_SCORE 650 non-null 24 AVG\_READING\_8\_SCORE 562 non-null

dtypes: float64(22), int64(1), object(2) memory usage: 335.1+ KB

**Обработка** **ненужных данных**

int64 float64 float64 float64 float64 float64 float64 float64 float64 float64 float64 float64 float64 float64 float64 float64 float64 float64 float64 float64 float64 float64 float64

In [7]: *#* *Удаляем* *столбцы,* *которые* *не* *несут* *значимой* *информации* data**.**drop(['INSTRUCTION\_EXPENDITURE','YEAR'], axis **=** 1, inplace **=** **True**)

In [8]: data**.**info()

<class 'pandas.core.frame.DataFrame'> RangeIndex: 1715 entries, 0 to 1714 Data columns (total 23 columns):

# Column Non-Null Count --- ------ --------------

0 PRIMARY\_KEY 1715 non-null 1 STATE 1715 non-null 2 ENROLL 1224 non-null 3 TOTAL\_REVENUE 1275 non-null 4 FEDERAL\_REVENUE 1275 non-null 5 STATE\_REVENUE 1275 non-null 6 LOCAL\_REVENUE 1275 non-null 7 TOTAL\_EXPENDITURE 1275 non-null 8 SUPPORT\_SERVICES\_EXPENDITURE 1275 non-null 9 OTHER\_EXPENDITURE 1224 non-null 10 CAPITAL\_OUTLAY\_EXPENDITURE 1275 non-null 11 GRADES\_PK\_G 1542 non-null 12 GRADES\_KG\_G 1632 non-null 13 GRADES\_4\_G 1632 non-null 14 GRADES\_8\_G 1632 non-null 15 GRADES\_12\_G 1632 non-null 16 GRADES\_1\_8\_G 1020 non-null 17 GRADES\_9\_12\_G 1071 non-null 18 GRADES\_ALL\_G 1632 non-null 19 AVG\_MATH\_4\_SCORE 565 non-null 20 AVG\_MATH\_8\_SCORE 602 non-null 21 AVG\_READING\_4\_SCORE 650 non-null 22 AVG\_READING\_8\_SCORE 562 non-null

dtypes: float64(21), object(2) memory usage: 308.3+ KB

Dtype -----object object

float64 float64 float64 float64 float64 float64 float64 float64 float64 float64 float64 float64 float64 float64 float64 float64 float64 float64 float64 float64 float64

In [9]: *#* *Заполняем* *отсутствующие* *значения*

|  |
| --- |
|  |
| data['TOTAL\_REVENUE'] **=** data['TOTAL\_REVENUE']**.**replace(0,np**.**nan) data['TOTAL\_REVENUE'] **=** data['TOTAL\_REVENUE']**.**fillna(data['TOTAL\_REVENUE']**.**mean()) |

In [10]: data**.**head()

Out[10]:

**0**

**1**

**2**

**3**

**4**

**PRIMARY\_KEY**

1992\_ALABAMA

1992\_ALASKA

1992\_ARIZONA

1992\_ARKANSAS

1992\_CALIFORNIA

**STATE** **ENROLL**

ALABAMA NaN

ALASKA NaN

ARIZONA NaN

ARKANSAS NaN

CALIFORNIA NaN

**TOTAL\_REVENUE**

2678885.0

1049591.0

3258079.0

1711959.0

26260025.0

**FEDERAL\_REVENUE**

304177.0

106780.0

297888.0

178571.0

2072470.0

**STATE\_REVENUE** **LO**

1659028.0

720711.0

1369815.0

958785.0

16546514.0

5 rows × 23 columns

In [11]: data**.**isnull()**.**sum()

*#* *проверим* *есть* *ли* *пропущенные* *значения* *в* *столбце* *business\_latitude*

Out[11]: PRIMARY\_KEY 0 STATE 0 ENROLL 491 TOTAL\_REVENUE 0 FEDERAL\_REVENUE 440 STATE\_REVENUE 440 LOCAL\_REVENUE 440 TOTAL\_EXPENDITURE 440 SUPPORT\_SERVICES\_EXPENDITURE 440 OTHER\_EXPENDITURE 491 CAPITAL\_OUTLAY\_EXPENDITURE 440 GRADES\_PK\_G 173 GRADES\_KG\_G 83 GRADES\_4\_G 83 GRADES\_8\_G 83 GRADES\_12\_G 83 GRADES\_1\_8\_G 695 GRADES\_9\_12\_G 644 GRADES\_ALL\_G 83 AVG\_MATH\_4\_SCORE 1150 AVG\_MATH\_8\_SCORE 1113 AVG\_READING\_4\_SCORE 1065 AVG\_READING\_8\_SCORE 1153 dtype: int64

In [12]:

In [13]:

total\_count **=** data**.**shape[0]

print('Всего строк: {}'**.**format(total\_count))

Всего строк: 1715

**Обработка** **пропусков категориальных данных**

*#* *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 {}, {}%.'**.**forma

In [14]:

РК1

*#* *Заполняем* *отсутствующие* *значения* data['STATE'] **=** data**.**fillna("Nane") data**.**head()

Out[14]:

**0**

**1**

**2**

**3**

**4**

**PRIMARY\_KEY**

1992\_ALABAMA

1992\_ALASKA

1992\_ARIZONA

1992\_ARKANSAS

1992\_CALIFORNIA

**STATE** **ENROLL**

1992\_ALABAMA NaN

1992\_ALASKA NaN

1992\_ARIZONA NaN

1992\_ARKANSAS NaN

1992\_CALIFORNIA NaN

**TOTAL\_REVENUE**

2678885.0

1049591.0

3258079.0

1711959.0

26260025.0

**FEDERAL\_REVENUE**

304177.0

106780.0

297888.0

178571.0

2072470.0

**STATE\_REVENU**

1659028

720711

1369815

958785

16546514

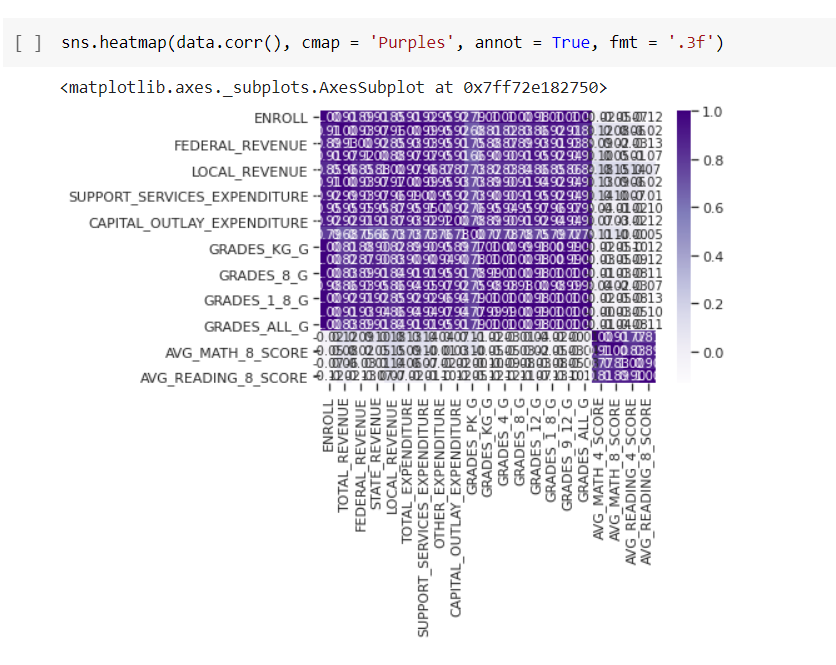
5 rows × 23 columns

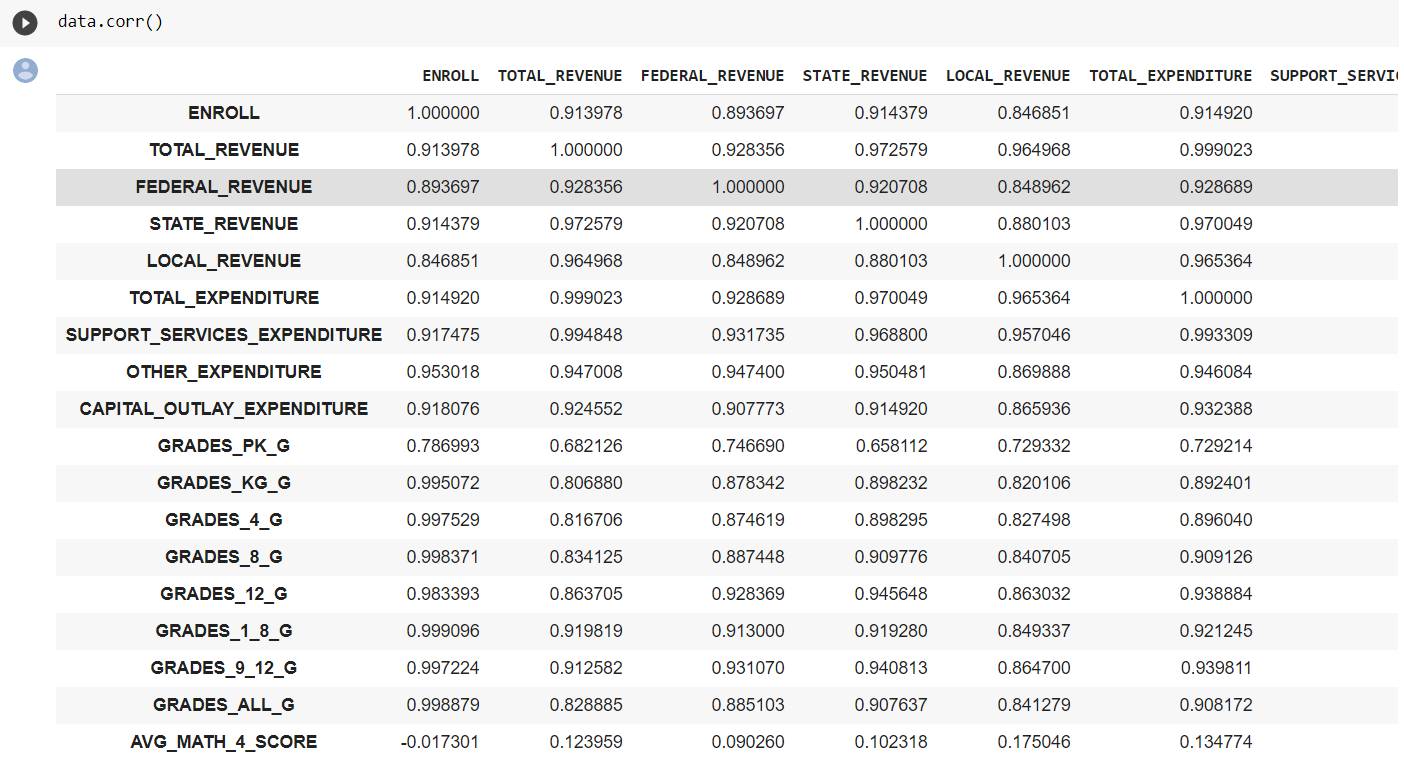
In [15]: data**.**isnull()**.**sum()

*#* *проверим* *есть* *ли* *пропущенные* *значения* *в* *столбце*

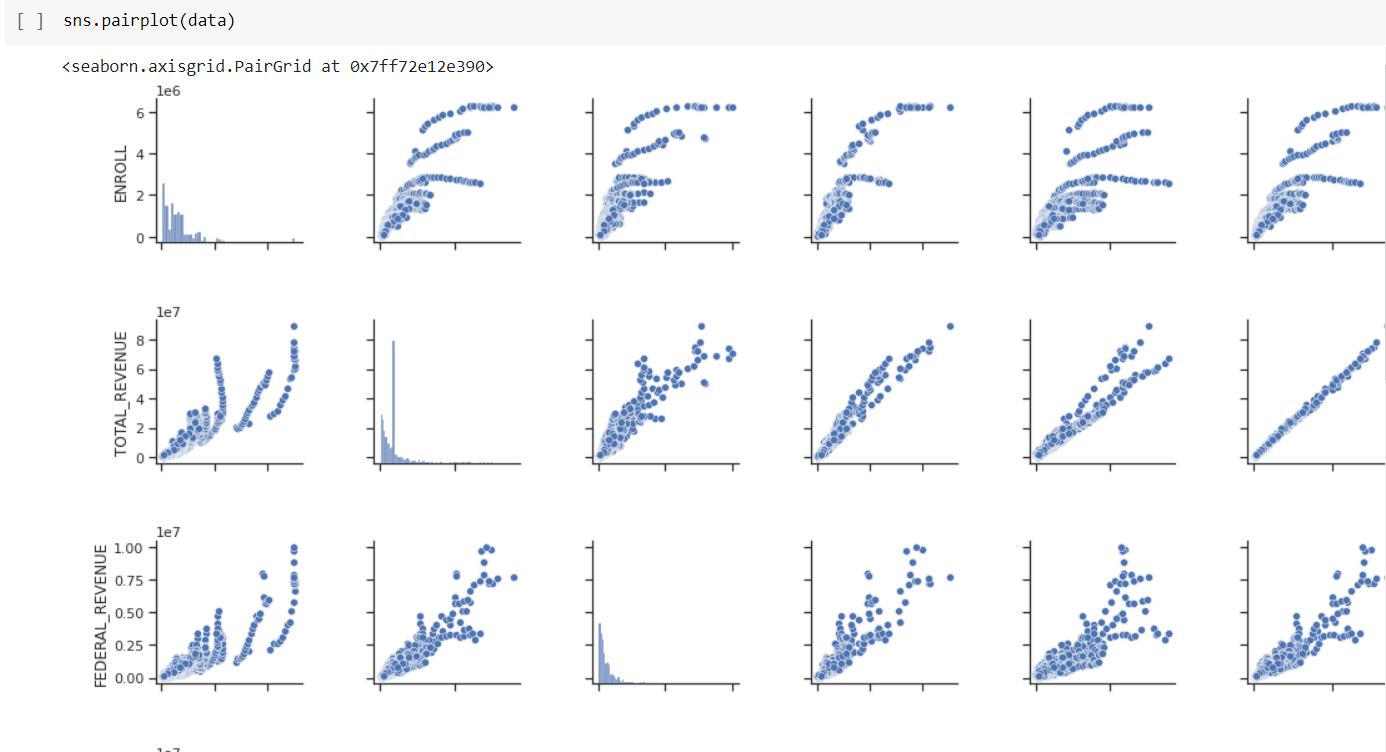
Out[15]: PRIMARY\_KEY 0 STATE 0 ENROLL 491 TOTAL\_REVENUE 0 FEDERAL\_REVENUE 440 STATE\_REVENUE 440 LOCAL\_REVENUE 440 TOTAL\_EXPENDITURE 440 SUPPORT\_SERVICES\_EXPENDITURE 440 OTHER\_EXPENDITURE 491 CAPITAL\_OUTLAY\_EXPENDITURE 440 GRADES\_PK\_G 173 GRADES\_KG\_G 83 GRADES\_4\_G 83 GRADES\_8\_G 83 GRADES\_12\_G 83 GRADES\_1\_8\_G 695 GRADES\_9\_12\_G 644 GRADES\_ALL\_G 83 AVG\_MATH\_4\_SCORE 1150 AVG\_MATH\_8\_SCORE 1113 AVG\_READING\_4\_SCORE 1065 AVG\_READING\_8\_SCORE 1153 dtype: int64

**Корреляционный анализ данных**

****

****

**Парные диаграммы**

****

****