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 ИУ5-65Б **Петренко С.С**

(Подпись, дата) (Фамилия И.О.)

Проверил:

Гапанюк Ю. Е.

(Подпись, дата) (Фамилия И.О.)

Москва, 2021

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?

Для студентов группы ИУ5-65Б - для набора данных построить "парные диаграммы".

Набор данных:

https://www.kaggle.com/noriuk/us-education-datasets-unification-project

РК1

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('states_all_extended.csv')
```

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

Out[3]:

	PRIMARY_KEY	STATE	YEAR	ENROLL	TOTAL_REVENUE	FEDERAL_REVENUE	S
0	1992_ALABAMA	ALABAMA	1992	NaN	2678885.0	304177.0	
1	1992_ALASKA	ALASKA	1992	NaN	1049591.0	106780.0	
2	1992_ARIZONA	ARIZONA	1992	NaN	3258079.0	297888.0	
3	1992_ARKANSAS	ARKANSAS	1992	NaN	1711959.0	178571.0	
4	1992_CALIFORNIA	CALIFORNIA	1992	NaN	26260025.0	2072470.0	

5 rows × 266 columns

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

Out[4]:
```
PRIMARY_KEY              object
STATE                    object
YEAR                      int64
ENROLL                  float64
TOTAL_REVENUE           float64
                          ...
G08_AM_A_MATHEMATICS    float64
G08_HP_A_READING        float64
G08_HP_A_MATHEMATICS    float64
G08_TR_A_READING        float64
G08_TR_A_MATHEMATICS    float64
Length: 266, dtype: object
```

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

Out[5]:
```
PRIMARY_KEY                0
STATE                      0
YEAR                       0
ENROLL                   491
TOTAL_REVENUE            440
                        ...
G08_AM_A_MATHEMATICS    1655
G08_HP_A_READING        1701
G08_HP_A_MATHEMATICS    1702
G08_TR_A_READING        1574
G08_TR_A_MATHEMATICS    1570
Length: 266, dtype: int64
```

In [6]:
```
data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1715 entries, 0 to 1714
Columns: 266 entries, PRIMARY_KEY to G08_TR_A_MATHEMATICS
dtypes: float64(263), int64(1), object(2)
memory usage: 3.5+ MB
```

Обработка пропусков

In [7]:
```
# Удаляем столбцы, которые не несут значимой информации
data.drop(['G08_TR_A_MATHEMATICS','G08_TR_A_MATHEMATICS'], axis = 1, inpla
```

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

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1715 entries, 0 to 1714
Columns: 265 entries, PRIMARY_KEY to G08_TR_A_READING
dtypes: float64(262), int64(1), object(2)
memory usage: 3.5+ MB
```

Обработка пропусков в числовых данных

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

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

Out[10]:

	PRIMARY_KEY	STATE	YEAR	ENROLL	TOTAL_REVENUE	FEDERAL_REVEN
0	1992_ALABAMA	ALABAMA	1992	917541.566176	2678885.0	30417
1	1992_ALASKA	ALASKA	1992	917541.566176	1049591.0	10678
2	1992_ARIZONA	ARIZONA	1992	917541.566176	3258079.0	29788
3	1992_ARKANSAS	ARKANSAS	1992	917541.566176	1711959.0	17857
4	1992_CALIFORNIA	CALIFORNIA	1992	917541.566176	26260025.0	207247

5 rows × 265 columns

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

Out[11]:
```
PRIMARY_KEY                0
STATE                      0
YEAR                       0
ENROLL                     0
TOTAL_REVENUE            440
                        ...
G08_AM_A_READING        1654
G08_AM_A_MATHEMATICS    1655
G08_HP_A_READING        1701
G08_HP_A_MATHEMATICS    1702
G08_TR_A_READING        1574
Length: 265, dtype: int64
```

Обработка пропусков в категориальных данных

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

```
Всего строк: 1715
```

In [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 {}, {
```

In [14]:
```
# Заполняем отсутствующие значения
data['PRIMARY_KEY'] = data.fillna("Nane")
data.head()
```

Out[14]:

	PRIMARY_KEY	STATE	YEAR	ENROLL	TOTAL_REVENUE	FEDERAL_REVEN
0	1992_ALABAMA	ALABAMA	1992	917541.566176	2678885.0	30417
1	1992_ALASKA	ALASKA	1992	917541.566176	1049591.0	10678
2	1992_ARIZONA	ARIZONA	1992	917541.566176	3258079.0	29788
3	1992_ARKANSAS	ARKANSAS	1992	917541.566176	1711959.0	17857
4	1992_CALIFORNIA	CALIFORNIA	1992	917541.566176	26260025.0	207247

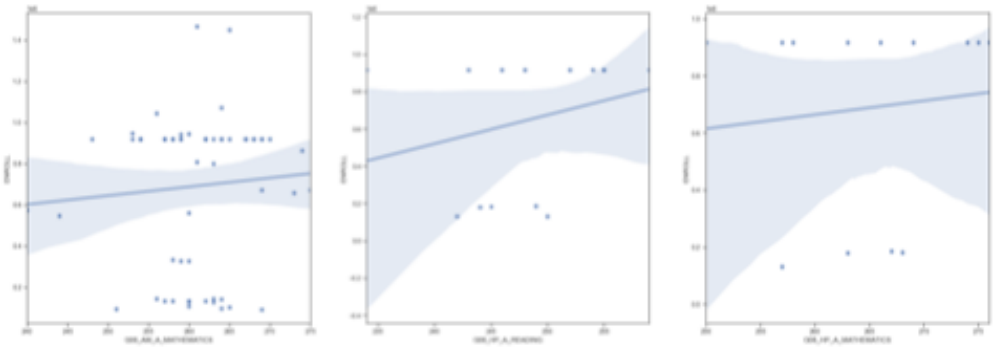
5 rows × 265 columns

In [15]:
```
data.isnull().sum()
# проверим есть ли пропущенные значения в столбце
```

Out[15]:
```
PRIMARY_KEY                0
STATE                      0
YEAR                       0
ENROLL                     0
TOTAL_REVENUE            440
                        ...
G08_AM_A_READING        1654
G08_AM_A_MATHEMATICS    1655
G08_HP_A_READING        1701
G08_HP_A_MATHEMATICS    1702
G08_TR_A_READING        1574
Length: 265, dtype: int64
```

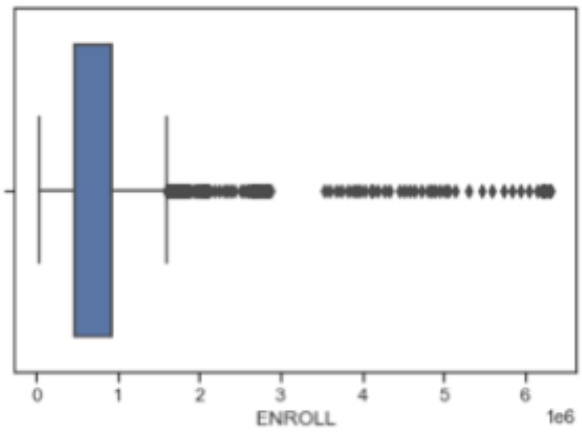
In [16]:
```
## Парные диаграммы
fig, axs = plt.subplots(ncols=3, figsize=(30,10))
sns.regplot(data['G08_AM_A_MATHEMATICS'], data['ENROLL'], ax = axs[0])
sns.regplot(data['G08_HP_A_READING'], data['ENROLL'], ax = axs[1])
sns.regplot(data['G08_HP_A_MATHEMATICS'], data['ENROLL'], ax = axs[2])
```

Out[16]:
```
<AxesSubplot:xlabel='G08_HP_A_MATHEMATICS', ylabel='ENROLL'>
```



In [17]:
```
sns.boxplot(data['ENROLL'])
```

Out[17]:
```
<AxesSubplot:xlabel='ENROLL'>
```



In [18]:
```
fig, ax = plt.subplots(figsize=(10,10))
sns.scatterplot(ax=ax, x='G08_AM_A_MATHEMATICS', y='ENROLL', data=data, hu
```

Out[18]: <AxesSubplot:xlabel='G08_AM_A_MATHEMATICS', ylabel='ENROLL'>