```
In [ ]:
import numpy as np
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
%matplotlib inline
sns.set(style = "ticks")
In [ ]:
data = pd.read csv('./states all extended.csv', sep=",")
In [ ]:
data.shape
(1715, 266)
In [ ]:
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
In [ ]:
data.dtypes
In [ ]:
data.head()
     PRIMARY_KEY
                      STATE YEAR ENROLL TOTAL_REVENUE FEDERAL_REVENUE STATE_REVENUE LOCAL_REVEN
                                                                                                71568
0
    1992_ALABAMA
                    ALABAMA
                              1992
                                      NaN
                                                2678885.0
                                                                  304177.0
                                                                                1659028.0
      1992 ALASKA
                     ALASKA
                              1992
                                      NaN
                                                 1049591.0
                                                                  106780.0
                                                                                 720711.0
                                                                                                22210
1
     1992_ARIZONA
                    ARIZONA
                              1992
                                                 3258079.0
                                                                  297888.0
                                                                                1369815.0
                                                                                               159037
2
                                      NaN
   1992_ARKANSAS
                   ARKANSAS
                                      NaN
                                                 1711959.0
                                                                  178571.0
                                                                                 958785.0
                                                                                                57460
3
                              1992
  1992_CALIFORNIA CALIFORNIA
                              1992
                                      NaN
                                                26260025.0
                                                                 2072470.0
                                                                               16546514.0
                                                                                               764104
5 rows × 266 columns
In [ ]:
data.isnull().sum()
In [ ]:
totalCount = data.shape[0]
print("Bcero ctpok: {}" .format(totalCount))
Всего строк: 1715
In [ ]:
catCols = []
for col in data.columns:
```

```
tempNullCount = data[data[col].isnull()].shape[0]
    dt = str(data[col].dtype)
    if tempNullCount>0 and (dt=='float64'):
        catCols.append(col)
         temp perc = round((tempNullCount / totalCount) * 100.0, 2)
        print('Колонка {}. Тип данных {}. Количество пустых значений {}, {}%.'.format(co
1, dt, tempNullCount, temp perc))
Колонка ENROLL. Тип данных float64. Количество пустых значений 491, 28.63%.
Колонка TOTAL REVENUE. Тип данных float64. Количество пустых значений 440, 25.66%.
Колонка FEDERAL REVENUE. Тип данных float64. Количество пустых значений 440, 25.66%.
Колонка STATE_REVENUE. Тип данных float64. Количество пустых значений 440, 25.66%.
Колонка LOCAL REVENUE. Тип данных float64. Количество пустых значений 440, 25.66%.
Колонка TOTAL EXPENDITURE. Тип данных float64. Количество пустых значений 440, 25.66%.
Колонка INSTRUCTION EXPENDITURE. Тип данных float64. Количество пустых значений 440, 25.6
Колонка SUPPORT SERVICES EXPENDITURE. Тип данных float64. Количество пустых значений 440,
Колонка OTHER EXPENDITURE. Тип данных float64. Количество пустых значений 491, 28.63%.
Колонка CAPITAL OUTLAY EXPENDITURE. Тип данных float64. Количество пустых значений 440, 2
Колонка А А А. Тип данных float64. Количество пустых значений 83, 4.84%.
Колонка G01 A A. Тип данных float64. Количество пустых значений 83, 4.84%.
Колонка G02 A A. Тип данных float64. Количество пустых значений 83, 4.84%.
Колонка G03\_A\_A. Тип данных float64. Количество пустых значений 83, 4.84%. Колонка G04\_A\_A. Тип данных float64. Количество пустых значений 83, 4.84%.
Колонка G05\_A\_A. Тип данных float64. Количество пустых значений 83, 4.84%.
Колонка G06\_A\_A. Тип данных float64. Количество пустых значений 83, 4.84%.
Колонка G07\_A\_A. Тип данных float64. Количество пустых значений 83, 4.84%.
Колонка G08_A_A. Тип данных float64. Количество пустых значений 83, 4.84%.
Колонка G09_A_A. Тип данных float64. Количество пустых значений 83, 4.84%.
Колонка G10 A A. Тип данных float64. Количество пустых значений 83, 4.84%.
Колонка G11 A A. Тип данных float64. Количество пустых значений 83, 4.84%.
Колонка G12 A A. Тип данных float64. Количество пустых значений 83, 4.84%.
Колонка КG A A. Тип данных float64. Количество пустых значений 83, 4.84%.
Колонка РК А А. Тип данных float64. Количество пустых значений 173, 10.09%.
Колонка G01-G08 A A. Тип данных float64. Количество пустых значений 695, 40.52%.
Колонка G09-G12 A A. Тип данных float64. Количество пустых значений 644, 37.55%.
Колонка G01 AM F. Тип данных float64. Количество пустых значений 1308, 76.27%.
Колонка G01 AM M. Тип данных float64. Количество пустых значений 1307, 76.21%.
Колонка G01_AS_F. Тип данных float64. Количество пустых значений 1307, 76.21%.
Колонка G01 AS M. Тип данных float64. Количество пустых значений 1307, 76.21%.
Колонка G01 BL F. Тип данных float64. Количество пустых значений 1307, 76.21%. Колонка G01 BL M. Тип данных float64. Количество пустых значений 1307, 76.21%.
Колонка G01_HI_F. Тип данных float64. Количество пустых значений 1308, 76.27%.
Колонка G01_{HI_M}. Тип данных float64. Количество пустых значений 1307, 76.21%.
Колонка G01_HP_F. Тип данных float64. Количество пустых значений 1351, 78.78%.
Колонка G01_HP_M. Тип данных float64. Количество пустых значений 1352, 78.83%.
Колонка G01_TR_F. Тип данных float64. Количество пустых значений 1344, 78.37%.
Колонка G01 TR M. Тип данных float64. Количество пустых значений 1344, 78.37%.
Колонка G01 WH F. Тип данных float64. Количество пустых значений 1307, 76.21%.
Колонка G01 WH M. Тип данных float64. Количество пустых значений 1307, 76.21%.
Колонка G02 AM F. Тип данных float64. Количество пустых значений 1309, 76.33%.
Колонка G02 AM M. Тип данных float64. Количество пустых значений 1308, 76.27%.
Колонка G02 AS F. Тип данных float64. Количество пустых значений 1307, 76.21%.
Колонка G02 AS M. Тип данных float64. Количество пустых значений 1307, 76.21%.
Колонка G02 BL F. Тип данных float64. Количество пустых значений 1307, 76.21%.
Колонка G02 BL M. Тип данных float64. Количество пустых значений 1307, 76.21%.
Колонка G02 HI F. Тип данных float64. Количество пустых значений 1307, 76.21%.
Колонка G02 HI M. Тип данных float64. Количество пустых значений 1307, 76.21%.
Колонка G02 HP F. Тип данных float64. Количество пустых значений 1351, 78.78%.
Колонка GO2_HP_M. Тип данных float64. Количество пустых значений 1351, 78.78%. Колонка GO2_TR_F. Тип данных float64. Количество пустых значений 1344, 78.37%. Колонка GO2_TR_M. Тип данных float64. Количество пустых значений 1344, 78.37%.
Колонка G02_WH_F. Тип данных float64. Количество пустых значений 1307, 76.21%.
Колонка G02_WH_M. Тип данных float64. Количество пустых значений 1307, 76.21%.
Колонка G03_AM_F. Тип данных float64. Количество пустых значений 1308, 76.27%.
Колонка G03_AM_M. Тип данных float64. Количество пустых значений 1309, 76.33%.
Колонка G03 AS F. Тип данных float64. Количество пустых значений 1307, 76.21%.
Колонка G03 AS M. Тип данных float64. Количество пустых значений 1307, 76.21%.
Колонка G03 BL F. Тип данных float64. Количество пустых значений 1307, 76.21%.
Колонка G03 BL M. Тип данных float64. Количество пустых значений 1307, 76.21%.
```

```
Колонка G03 HI F. Тип данных float64. Количество пустых значений 1307, 76.21%.
Колонка G03 HI M. Тип данных float64. Количество пустых значений 1307, 76.21%.
Колонка G03 HP F. Тип данных float64. Количество пустых значений 1352, 78.83%.
Колонка G03 HP M. Тип данных float64. Количество пустых значений 1349, 78.66%.
Колонка G03_TR_F. Тип данных float64. Количество пустых значений 1344, 78.37%.
Колонка G03_TR_M. Тип данных float64. Количество пустых значений 1344, 78.37%. Колонка G03_WH_F. Тип данных float64. Количество пустых значений 1307, 76.21%.
Колонка G03\_WH\_M. Тип данных float64. Количество пустых значений 1307, 76.21%. Колонка G04\_AM\_F. Тип данных float64. Количество пустых значений 1308, 76.27%.
Колонка G04\_AM\_M. Тип данных float64. Количество пустых значений 1308, 76.27%.
Колонка G04\_AS\_F. Тип данных float64. Количество пустых значений 1307, 76.21%.
Колонка G04\_AS\_M. Тип данных float64. Количество пустых значений 1307, 76.21%.
Колонка G04 BL F. Тип данных float64. Количество пустых значений 1307, 76.21%.
Колонка G04 BL M. Тип данных float64. Количество пустых значений 1307, 76.21%.
Колонка G04 HI F. Тип данных float64. Количество пустых значений 1307, 76.21%.
Колонка G04 HI M. Тип данных float64. Количество пустых значений 1307, 76.21%.
Колонка G04 HP F. Тип данных float64. Количество пустых значений 1351, 78.78%.
Колонка G04 HP M. Тип данных float64. Количество пустых значений 1351, 78.78%.
Колонка G04 TR F. Тип данных float64. Количество пустых значений 1344, 78.37%.
Колонка G04 TR M. Тип данных float64. Количество пустых значений 1344, 78.37%.
Колонка G04 WH F. Тип данных float64. Количество пустых значений 1307, 76.21%.
Колонка G04 WH M. Тип данных float64. Количество пустых значений 1307, 76.21%.
Колонка G05 AM F. Тип данных float64. Количество пустых значений 1308, 76.27%.
Колонка G05 AM M. Тип данных float64. Количество пустых значений 1308, 76.27%.
Колонка G05\_AS\_F. Тип данных float64. Количество пустых значений 1307, 76.21%. Колонка G05\_AS\_M. Тип данных float64. Количество пустых значений 1307, 76.21%.
Колонка G05\_BL\_F. Тип данных float64. Количество пустых значений 1307, 76.21%.
Колонка G05\_BL\_M. Тип данных float64. Количество пустых значений 1307, 76.21%.
Колонка G05_{HI_F}. Тип данных float64. Количество пустых значений 1307, 76.21%.
Колонка G05_{HI_M}. Тип данных float64. Количество пустых значений 1307, 76.21%.
Колонка G05_HP_F. Тип данных float64. Количество пустых значений 1352, 78.83%.
Колонка G05 HP M. Тип данных float64. Количество пустых значений 1349, 78.66%.
Колонка G05 TR F. Тип данных float64. Количество пустых значений 1344, 78.37%.
Колонка G05 TR M. Тип данных float64. Количество пустых значений 1344, 78.37%.
Колонка G05 WH F. Тип данных float64. Количество пустых значений 1307, 76.21%.
Колонка G05 WH M. Тип данных float64. Количество пустых значений 1307, 76.21%.
Колонка G06 AM F. Тип данных float64. Количество пустых значений 1307, 76.21%.
Колонка G06 AM M. Тип данных float64. Количество пустых значений 1308, 76.27%.
Колонка G06_AS_F. Тип данных float64. Количество пустых значений 1307, 76.21%.
Колонка G06 AS M. Тип данных float64. Количество пустых значений 1307, 76.21%.
Колонка G06_BL_F. Тип данных float64. Количество пустых значений 1307, 76.21%.
Колонка G06 BL M. Тип данных float64. Количество пустых значений 1307, 76.21%.
Колонка G06_HI_F. Тип данных float64. Количество пустых значений 1307, 76.21%. Колонка G06_HI_M. Тип данных float64. Количество пустых значений 1307, 76.21%.
Колонка G06_{\rm HP} F. Тип данных float64. Количество пустых значений 1351, 78.78%. Колонка G06_{\rm HP} M. Тип данных float64. Количество пустых значений 1349, 78.66%.
Колонка G06\_TR\_F. Тип данных float64. Количество пустых значений 1344, 78.37%.
Колонка G06_{TR}М. Тип данных float64. Количество пустых значений 1344, 78.37%.
Колонка G06_WH_F. Тип данных float64. Количество пустых значений 1307, 76.21%.
Колонка G06 WH M. Тип данных float64. Количество пустых значений 1307, 76.21%.
Колонка G07 AM F. Тип данных float64. Количество пустых значений 1307, 76.21%.
Колонка G07 AM M. Тип данных float64. Количество пустых значений 1308, 76.27%.
Колонка G07 AS F. Тип данных float64. Количество пустых значений 1307, 76.21%.
Колонка G07 AS M. Тип данных float64. Количество пустых значений 1307, 76.21%.
Колонка G07 BL F. Тип данных float64. Количество пустых значений 1307, 76.21%.
Колонка G07 BL M. Тип данных float64. Количество пустых значений 1307, 76.21%.
Колонка G07 HI F. Тип данных float64. Количество пустых значений 1307, 76.21%.
Колонка G07 HI M. Тип данных float64. Количество пустых значений 1307, 76.21%.
Колонка G07 HP F. Тип данных float64. Количество пустых значений 1350, 78.72%.
Колонка G07 HP M. Тип данных float64. Количество пустых значений 1352, 78.83%.
Колонка G07 TR F. Тип данных float64. Количество пустых значений 1344, 78.37%.
Колонка G07_TR_M. Тип данных float64. Количество пустых значений 1344, 78.37%.
Колонка G07\_WH\_F. Тип данных float64. Количество пустых значений 1307, 76.21%. Колонка G07\_WH\_M. Тип данных float64. Количество пустых значений 1307, 76.21%.
Колонка G08\_AM\_F. Тип данных float64. Количество пустых значений 1308, 76.27%.
Колонка G08_AM_M. Тип данных float64. Количество пустых значений 1307, 76.21%.
Колонка G08_AS_F. Тип данных float64. Количество пустых значений 1307, 76.21%.
Колонка G08_AS_M. Тип данных float64. Количество пустых значений 1307, 76.21%.
Колонка G08 BL F. Тип данных float64. Количество пустых значений 1307, 76.21%.
Колонка G08 BL M. Тип данных float64. Количество пустых значений 1307, 76.21%.
Колонка G08 HI F. Тип данных float64. Количество пустых значений 1307, 76.21%.
Колонка G08 HI M. Тип данных float64. Количество пустых значений 1307, 76.21%.
```

```
Колонка G08 HP F. Тип данных float64. Количество пустых значений 1350, 78.72%.
Колонка G08 HP M. Тип данных float64. Количество пустых значений 1349, 78.66%.
Колонка G08 TR F. Тип данных float64. Количество пустых значений 1344, 78.37%.
Колонка G08 TR M. Тип данных float64. Количество пустых значений 1344, 78.37%.
Колонка G08 WH F. Тип данных float64. Количество пустых значений 1307, 76.21%.
Колонка G08 WH M. Тип данных float64. Количество пустых значений 1307, 76.21%.
Колонка G09_AM_F. Тип данных float64. Количество пустых значений 1307, 76.21%.
Колонка G09\_AM\_M. Тип данных float64. Количество пустых значений 1308, 76.27%. Колонка G09\_AS\_F. Тип данных float64. Количество пустых значений 1307, 76.21%.
Колонка G09_AS_M. Тип данных float64. Количество пустых значений 1307, 76.21%.
Колонка G09_BL_F. Тип данных float64. Количество пустых значений 1307, 76.21%.
Колонка G09_BL_M. Тип данных float64. Количество пустых значений 1307, 76.21%.
Колонка G09 HI F. Тип данных float64. Количество пустых значений 1307, 76.21%.
Колонка G09 HI M. Тип данных float64. Количество пустых значений 1307, 76.21%.
Колонка G09 HP F. Тип данных float64. Количество пустых значений 1352, 78.83%.
Колонка G09 HP M. Тип данных float64. Количество пустых значений 1351, 78.78%.
Колонка G09 TR F. Тип данных float64. Количество пустых значений 1344, 78.37%.
Колонка G09 TR M. Тип данных float64. Количество пустых значений 1344, 78.37%.
Колонка G09 WH F. Тип данных float64. Количество пустых значений 1307, 76.21%.
Колонка G09 WH M. Тип данных float64. Количество пустых значений 1307, 76.21%.
Колонка G10 AM F. Тип данных float64. Количество пустых значений 1307, 76.21%.
Колонка G10 AM M. Тип данных float64. Количество пустых значений 1308, 76.27%.
Колонка G10 AS F. Тип данных float64. Количество пустых значений 1307, 76.21%.
Колонка G10 AS M. Тип данных float64. Количество пустых значений 1307, 76.21%.
Колонка G10\_BL\_F. Тип данных float64. Количество пустых значений 1307, 76.21%. Колонка G10\_BL\_M. Тип данных float64. Количество пустых значений 1307, 76.21%.
Колонка G10_{HI_F}. Тип данных float64. Количество пустых значений 1307, 76.21%.
Колонка G10_HI_M. Тип данных float64. Количество пустых значений 1307, 76.21%.
Колонка G10_HP_F. Тип данных float64. Количество пустых значений 1352, 78.83%.
Колонка G10 HP M. Тип данных float64. Количество пустых значений 1351, 78.78%.
Колонка G10_TR_F. Тип данных float64. Количество пустых значений 1344, 78.37%.
Колонка G10 TR M. Тип данных float64. Количество пустых значений 1344, 78.37%.
Колонка G10 WH F. Тип данных float64. Количество пустых значений 1307, 76.21%.
Колонка G10 WH M. Тип данных float64. Количество пустых значений 1307, 76.21%.
Колонка G11 AM F. Тип данных float64. Количество пустых значений 1307, 76.21%.
Колонка G11 AM M. Тип данных float64. Количество пустых значений 1307, 76.21%.
Колонка G11 AS F. Тип данных float64. Количество пустых значений 1307, 76.21%.
Колонка G11 AS M. Тип данных float64. Количество пустых значений 1307, 76.21%.
Колонка G11_BL_F. Тип данных float64. Количество пустых значений 1307, 76.21%.
Колонка G11 BL M. Тип данных float64. Количество пустых значений 1307, 76.21%.
Колонка G11_HI_F. Тип данных float64. Количество пустых значений 1308, 76.27%.
Колонка G11 HI M. Тип данных float64. Количество пустых значений 1307, 76.21%.
Колонка G11_HP_F. Тип данных float64. Количество пустых значений 1352, 78.83%. Колонка G11_HP_M. Тип данных float64. Количество пустых значений 1354, 78.95%. Колонка G11_TR_F. Тип данных float64. Количество пустых значений 1344, 78.37%.
Колонка G11_TR_M. Тип данных float64. Количество пустых значений 1344, 78.37%.
Колонка G11_WH_F. Тип данных float64. Количество пустых значений 1307, 76.21%.
Колонка G11_WH_M. Тип данных float64. Количество пустых значений 1307, 76.21%.
Колонка G12_AM_F. Тип данных float64. Количество пустых значений 1309, 76.33%.
Колонка G12 AM M. Тип данных float64. Количество пустых значений 1310, 76.38%.
Колонка G12 AS F. Тип данных float64. Количество пустых значений 1307, 76.21%.
Колонка G12 AS M. Тип данных float64. Количество пустых значений 1307, 76.21%.
Колонка G12 BL F. Тип данных float64. Количество пустых значений 1307, 76.21%.
Колонка G12 BL M. Тип данных float64. Количество пустых значений 1307, 76.21%.
Колонка G12 HI F. Тип данных float64. Количество пустых значений 1307, 76.21%.
Колонка G12 HI M. Тип данных float64. Количество пустых значений 1307, 76.21%.
Колонка G12 HP F. Тип данных float64. Количество пустых значений 1352, 78.83%.
Колонка G12 HP M. Тип данных float64. Количество пустых значений 1352, 78.83%.
Колонка G12 TR F. Тип данных float64. Количество пустых значений 1344, 78.37%.
Колонка G12 TR M. Тип данных float64. Количество пустых значений 1344, 78.37%.
Колонка G12 WH F. Тип данных float64. Количество пустых значений 1307, 76.21%.
Колонка G12 WH M. Тип данных float64. Количество пустых значений 1307, 76.21%.
Колонка KG\_AM\_F. Тип данных float64. Количество пустых значений 1307, 76.21%.
Колонка KG_AM_M. Тип данных float64. Количество пустых значений 1308, 76.27%.
Колонка KG_AS_F. Тип данных float64. Количество пустых значений 1307, 76.21%.
Колонка КG AS M. Тип данных float64. Количество пустых значений 1307, 76.21%.
Колонка KG_BL_F. Тип данных float64. Количество пустых значений 1307, 76.21%.
Колонка KG_BL_M. Тип данных float64. Количество пустых значений 1307, 76.21%.
Колонка КG HI F. Тип данных float64. Количество пустых значений 1308, 76.27%.
Колонка КG HI M. Тип данных float64. Количество пустых значений 1307, 76.21%.
Колонка КG HP F. Тип данных float64. Количество пустых значений 1349, 78.66%.
Колонка КG HP M. Тип данных float64. Количество пустых значений 1350, 78.72%.
```

```
Колонка КG TR F. Тип данных float64. Количество пустых значений 1344, 78.37%.
Колонка КG TR M. Тип данных float64. Количество пустых значений 1344, 78.37%.
Колонка КG WH F. Тип данных float64. Количество пустых значений 1307, 76.21%.
Колонка КG WH M. Тип данных float64. Количество пустых значений 1307, 76.21%.
Колонка РК АМ F. Тип данных float64. Количество пустых значений 1332, 77.67%.
Колонка РК АМ М. Тип данных float64. Количество пустых значений 1321, 77.03%.
Колонка РК_AS_F. Тип данных float64. Количество пустых значений 1321, 77.03%. Колонка РК_AS_M. Тип данных float64. Количество пустых значений 1323, 77.14%. Колонка РК_BL_F. Тип данных float64. Количество пустых значений 1321, 77.03%.
Колонка РК_ВL_М. Тип данных float64. Количество пустых значений 1321, 77.03%.
Колонка РК_НІ_F. Тип данных float64. Количество пустых значений 1321, 77.03%.
Колонка РК_НІ_М. Тип данных float64. Количество пустых значений 1321, 77.03%.
Колонка РК НР F. Тип данных float64. Количество пустых значений 1387, 80.87%.
Колонка РК НР М. Тип данных float64. Количество пустых значений 1384, 80.7%.
Колонка РК TR F. Тип данных float64. Количество пустых значений 1357, 79.13%.
Колонка РК TR M. Тип данных float64. Количество пустых значений 1357, 79.13%.
Колонка РК WH F. Тип данных float64. Количество пустых значений 1321, 77.03%.
Колонка РК WH M. Тип данных float64. Количество пустых значений 1321, 77.03%.
Колонка G04 A A READING. Тип данных float64. Количество пустых значений 1065, 62.1%.
Колонка G04 A A MATHEMATICS. Тип данных float64. Количество пустых значений 1150, 67.06%.
Колонка G04 A M READING. Тип данных float64. Количество пустых значений 1065, 62.1%.
Колонка G04 A M MATHEMATICS. Тип данных float64. Количество пустых значений 1150, 67.06%.
Колонка G04 A F READING. Тип данных float64. Количество пустых значений 1065, 62.1%.
Колонка G04 A F MATHEMATICS. Тип данных float64. Количество пустых значений 1150, 67.06%.
Колонка G04 WH A READING. Тип данных float64. Количество пустых значений 1450, 84.55%.
Колонка G04 WH A MATHEMATICS. Тип данных float64. Количество пустых значений 1450, 84.55%
Колонка G04_BL_A_READING. Тип данных float64. Количество пустых значений 1489, 86.82%.
Колонка G04 BL A MATHEMATICS. Тип данных float64. Количество пустых значений 1486, 86.65%
Колонка G04 HI A READING. Тип данных float64. Количество пустых значений 1465, 85.42%.
Колонка G04 HI A MATHEMATICS. Тип данных float64. Количество пустых значений 1465, 85.42%
Колонка G04 AS A READING. Тип данных float64. Количество пустых значений 1551, 90.44%.
Колонка G04 AS A MATHEMATICS. Тип данных float64. Количество пустых значений 1547, 90.2%.
Колонка G04 AM A READING. Тип данных float64. Количество пустых значений 1651, 96.27%.
Колонка G04 AM A MATHEMATICS. Тип данных float64. Количество пустых значений 1652, 96.33%
Колонка G04 HP A READING. Тип данных float64. Количество пустых значений 1699, 99.07%.
Колонка G04 HP A MATHEMATICS. Тип данных float64. Количество пустых значений 1700, 99.13%
Колонка G04 TR A READING. Тип данных float64. Количество пустых значений 1532, 89.33%.
Колонка G04 TR A MATHEMATICS. Тип данных float64. Количество пустых значений 1532, 89.33%
Колонка G08 A A READING. Тип данных float64. Количество пустых значений 1153, 67.23%.
Колонка G08_A_A_MATHEMATICS. Тип данных float64. Количество пустых значений 1113, 64.9%.
Колонка G08_A_M_READING. Тип данных float64. Количество пустых значений 1153, 67.23%.
Колонка G08 A M MATHEMATICS. Тип данных float64. Количество пустых значений 1113, 64.9%.
Колонка G08_A_F_READING. Тип данных float64. Количество пустых значений 1153, 67.23%.
Колонка G08 A F MATHEMATICS. Тип данных float64. Количество пустых значений 1113, 64.9%.
Колонка G08 WH A READING. Тип данных float64. Количество пустых значений 1450, 84.55%.
Колонка G08 WH A MATHEMATICS. Тип данных float64. Количество пустых значений 1450, 84.55%
Колонка G08 BL A READING. Тип данных float64. Количество пустых значений 1493, 87.06%.
Колонка G08 BL A MATHEMATICS. Тип данных float64. Количество пустых значений 1494, 87.11%
Колонка G08 HI A READING. Тип данных float64. Количество пустых значений 1469, 85.66%.
Колонка G08 HI A MATHEMATICS. Тип данных float64. Количество пустых значений 1467, 85.54%
Колонка G08 AS A READING. Тип данных float64. Количество пустых значений 1562, 91.08%.
Колонка G08 AS A MATHEMATICS. Тип данных float64. Количество пустых значений 1558, 90.85%
Колонка G08 AM A READING. Тип данных float64. Количество пустых значений 1654, 96.44%.
Колонка G08 AM A MATHEMATICS. Тип данных float64. Количество пустых значений 1655, 96.5%.
Колонка G08_HP_A_READING. Тип данных float64. Количество пустых значений 1701, 99.18%.
Колонка G08 HP A MATHEMATICS. Тип данных float64. Количество пустых значений 1702, 99.24%
Колонка G08_TR_A_READING. Тип данных float64. Количество пустых значений 1574, 91.78%.
Колонка G08 TR A MATHEMATICS. Тип данных float64. Количество пустых значений 1570, 91.55%
```

In []:

```
from sklearn.impute import SimpleImputer
from sklearn.impute import MissingIndicator
In [ ]:
state data = data[['STATE']]
state data.head()
       STATE
0
    ALABAMA
     ALASKA
1
     ARIZONA
   ARKANSAS
4 CALIFORNIA
In [ ]:
state data['STATE'].unique()
In [ ]:
state_data[state_data['STATE'].isnull()].shape
(0, 1)
In [ ]:
imp = SimpleImputer(missing values=np.nan, strategy="most frequent")
data_imp = imp.fit_transform(state_data)
data_imp
In [ ]:
np.unique(data_imp)
In [ ]:
data_imp.shape
(1715, 1)
In [ ]:
state enc = pd.DataFrame({"STATE": data_imp.T[0]})
state enc
           STATE
  0
         ALABAMA
          ALASKA
   1
         ARIZONA
        ARKANSAS
  3
       CALIFORNIA
1710
          VIRGINIA
1711
      WASHINGTON
1712 WEST_VIRGINIA
1713
       WISCONSIN
```

```
WYSMATE
1714
1715 rows x 1 columns
In [ ]:
from sklearn.preprocessing import OneHotEncoder
ohe = OneHotEncoder()
state_ohe = ohe.fit_transform(state_enc[['STATE']])
In [ ]:
state ohe.shape
(1715, 53)
In [ ]:
state ohe.todense()[0:10]
In [ ]:
pd.get_dummies(state_enc).head()
   STATE_ALABAMA STATE_ALASKA STATE_ARIZONA STATE_ARKANSAS STATE_CALIFORNIA STATE_COLORADO STATE
0
               1
                             0
                                            0
                                                            0
                                                                             0
                                                                                             0
1
               0
                             1
                                            0
                                                            0
                                                                             0
                                                                                             0
2
               0
                             0
                                            1
                                                            0
                                                                             0
                                                                                             0
3
               0
                             0
                                            0
                                                            1
                                                                             0
                                                                                             0
               0
                             0
                                            0
                                                            0
                                                                             1
5 rows × 53 columns
In [ ]:
pd.get dummies(state data, dummy na=True).head()
   STATE_ALABAMA STATE_ALASKA STATE_ARIZONA STATE_ARKANSAS STATE_CALIFORNIA STATE_COLORADO STATE
                             0
0
               1
                                            0
                                                            0
                                                                             0
                                                                                             0
1
               0
                             1
                                            0
                                                            0
                                                                             0
                                                                                             0
2
               0
                             0
                                                            0
                                                                             0
                                                                                             0
                                            1
3
               0
                             0
                                            0
                                                            1
                                                                             0
                                                                                             0
                0
                                            0
                                                            O
5 rows × 54 columns
4
In [ ]:
from sklearn.preprocessing import LabelEncoder
In [ ]:
state enc["STATE"].unique()
In [ ]:
```

le = LabelEncoder()

state le = le.fit transform(state enc["STATE"])

```
In [ ]:
state_le
In [ ]:
np.unique(state_le)
In [ ]:
le.inverse transform([n for n in range(17)])
In [ ]:
data digit = state enc.copy()
data_digit["STATE"] = state_le
In [ ]:
from sklearn.preprocessing import MinMaxScaler, StandardScaler, Normalizer
In [ ]:
sc1 = MinMaxScaler()
sc1 data = sc1.fit transform(data digit[["STATE"]])
In [ ]:
plt.hist(data_digit["STATE"], 50)
plt.show()
60
 50
 40
 30
 20
10
In [ ]:
plt.hist(sc1 data, 50)
plt.show()
 60
50
 40
 30
 20
10
```

0.4

0.6

0.8

```
In [ ]:
sc2 = StandardScaler()
sc2_data = sc2.fit_transform(data_digit[["STATE"]])
In [ ]:
plt.hist(sc2_data, 50)
plt.show()
60
50
40
30
20
10
 0
            -1.0
                 -0.5
                        0.0
                              0.5
                                   1.0
In [ ]:
sns.pairplot(data)
```

Здесь должны были быть графики, но они не вывелись у меня даже через 2 часа загрузки...