Московский государственный технический университет им. Н.Э. Баумана

Факультет «Информатика и системы управления»
Кафедра ИУ5 «Системы обработки информации и управления»

Курс «Технологии машинного обучения»

Отчет по рубежному контролю №1
Вариант 12

Выполнил:
студент группы ИУ5-63Б
Кузнецов Г.И.

Проверил:
преподаватель каф. ИУ5
Гапанюк Ю.Е.

Москва, 2022 г.

Рубежный контроль №1 по ТМО

Вариант 12

Кузнецов Григорий ИУ5-63Б

Задача №2, набор №4 - https://www.kaggle.com/datasets/noriuk/us-education-datasets-unification-project (файл states_all.csv)

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?

Импорт библиотек

In [ ]:
```
import numpy as np
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
import sklearn.impute
%matplotlib inline
sns.set(style="ticks")
```

Первичный анализ данных

In [ ]:
```
data = pd.read_csv('states_all.csv')
data.head()
```

Out[ ]:

	PRIMARY_KEY	STATE	YEAR	ENROLL	TOTAL_REVENUE	FEDERAL_REVENUE	STATE_REVE
0	1992_ALABAMA	ALABAMA	1992	NaN	2678885.0	304177.0	16590
1	1992_ALASKA	ALASKA	1992	NaN	1049591.0	106780.0	7207
2	1992_ARIZONA	ARIZONA	1992	NaN	3258079.0	297888.0	13698
3	1992_ARKANSAS	ARKANSAS	1992	NaN	1711959.0	178571.0	9587
4	1992_CALIFORNIA	CALIFORNIA	1992	NaN	26260025.0	2072470.0	165465

5 rows × 25 columns

In [ ]:
```
data.shape
```

Out[ ]:
```
(1715, 25)
```

Заполнение пропусков данных

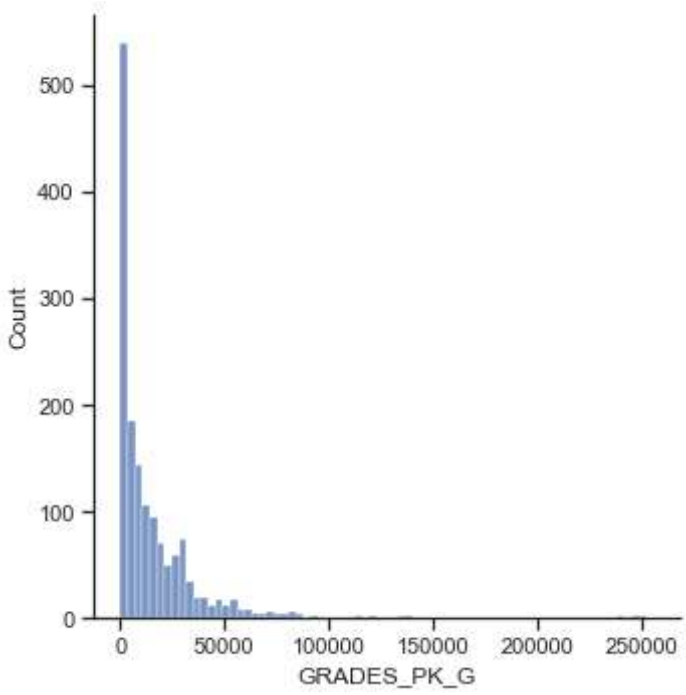
In [ ]:
```
data.isnull().sum()
```

Out[ ]:
```
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
```

Для обработки пропусков количественных данных выберем колонку 'GRADES_PK_G'

In [ ]:
```
sns.displot(data['GRADES_PK_G'])
```

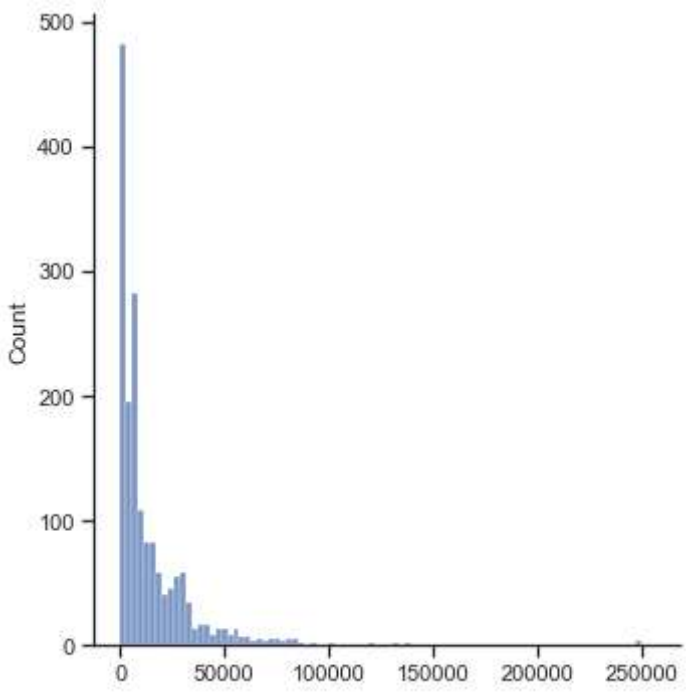
Out[ ]:
```
<seaborn.axisgrid.FacetGrid at 0x2bb2c9c5db0>
```



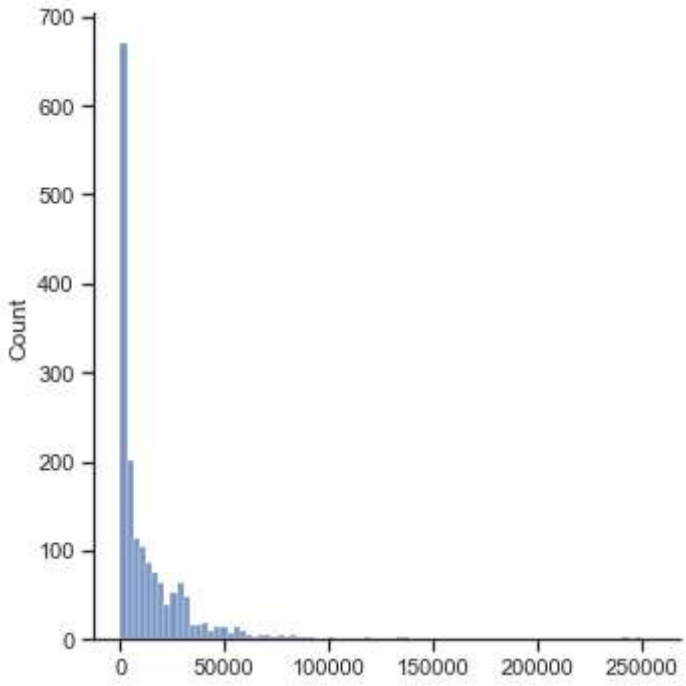
In [ ]:
```
from sklearn.impute import SimpleImputer

def fill_data(strategy):
    imputer = SimpleImputer(strategy=strategy)
```

```
    new_data = imputer.fit_transform(data[['GRADES_PK_G']])
    sns.displot(new_data.transpose()[0])
```

In [ ]:
```
fill_data('median')
```



In [ ]:
```
fill_data('most_frequent')
```



Не смотря на то, что при заполнении с помощью медианного способа появляется небольшой выброс, данная стратегия будет лучше, так как общая форма распределение сохраняется лучше. В отличе от модального заполнения, где график распределения сильно "вытягивается" в начале.

В данном датасете отсутствуют пропущенные значения категориальных признаков. Заполнение категориальных признаков происходит аналогично заполнению числовых

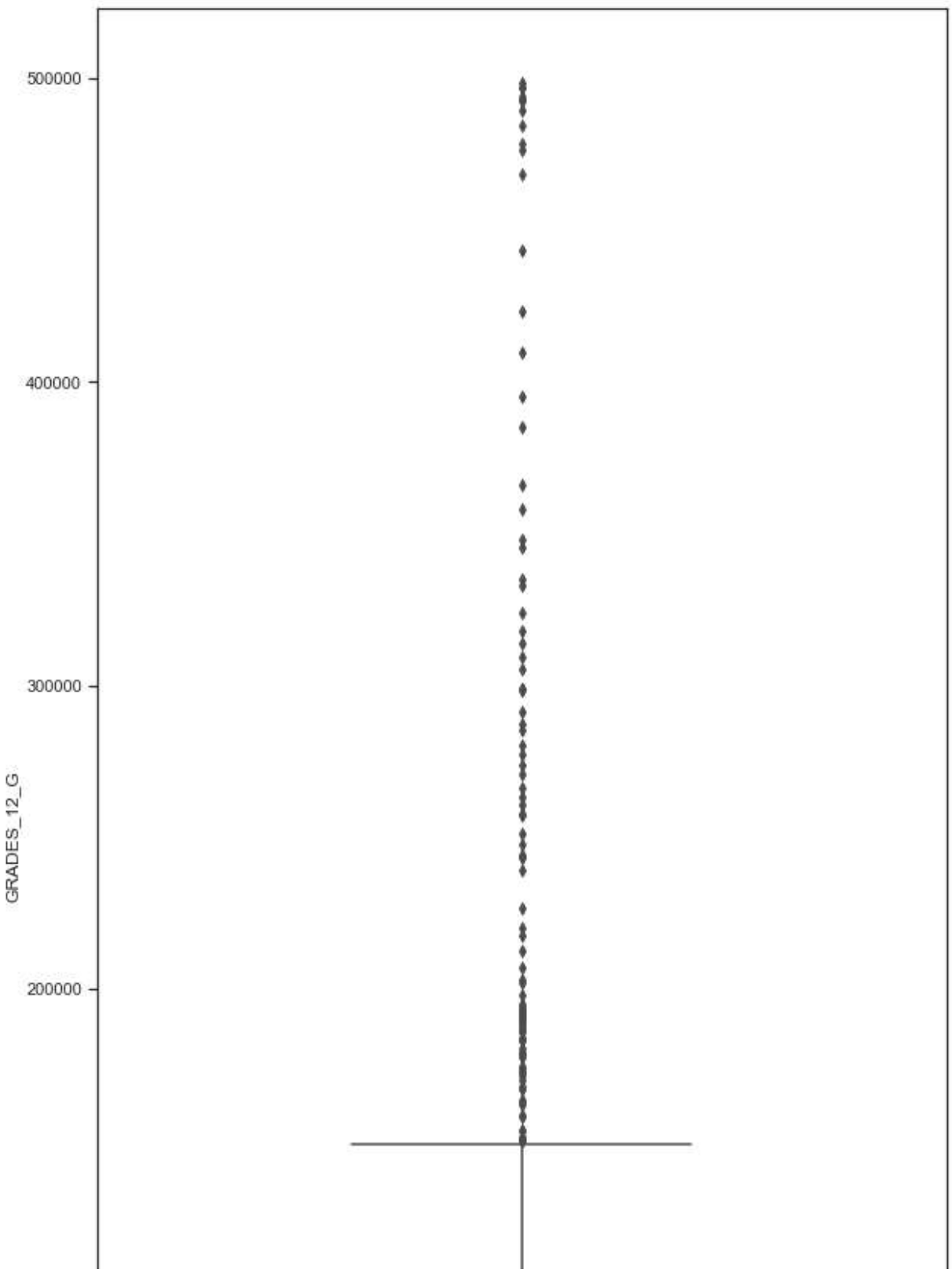
признаков, с возможными стартегиями: заполнение константой или заполнение модальным (наиболее встречающимся значением).

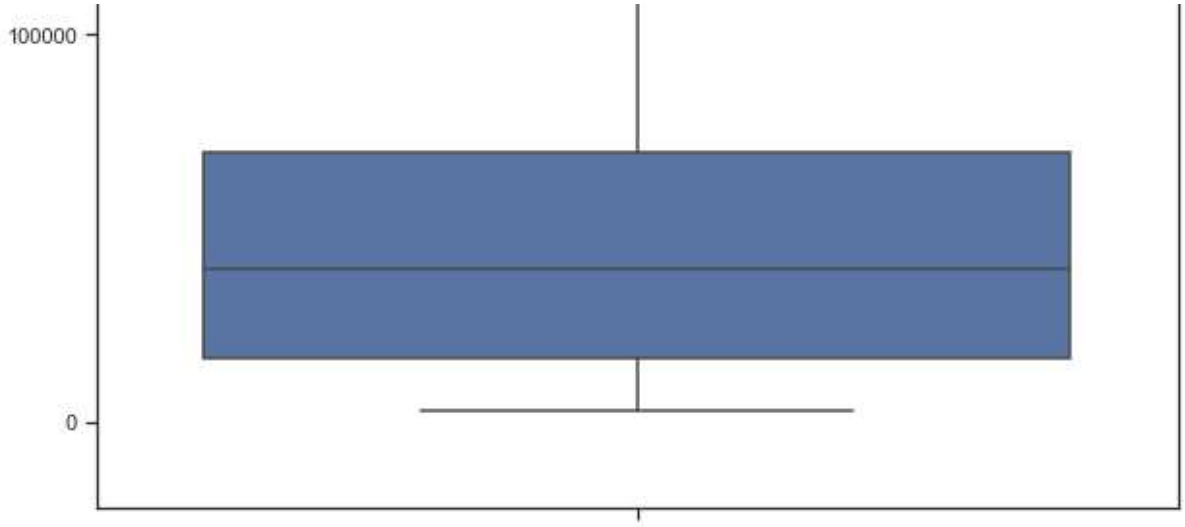
Дополнительное задание

Для произвольной колонки данных построить график "Ящик с усами (boxplot)"

In [ ]:
```
plt.subplots(figsize = (10,20))
sns.boxplot(data = data, y = 'GRADES_12_G')
```

Out[ ]:
```
<AxesSubplot:ylabel='GRADES_12_G'>
```





Видно, что наибольшее количество значений попадает в предел ~2000 - ~7000. Не смотря на это, данная колонка содержит много выбросов.