

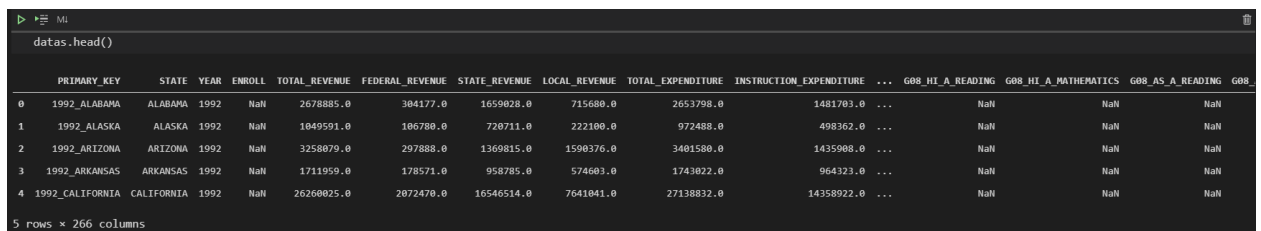
# Кобяк Андрей Вячеславович Рубежный контроль №1

## ИУ5-62Б

### Задание 2

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? + Гистограмма

Первые 5 строк датасета:

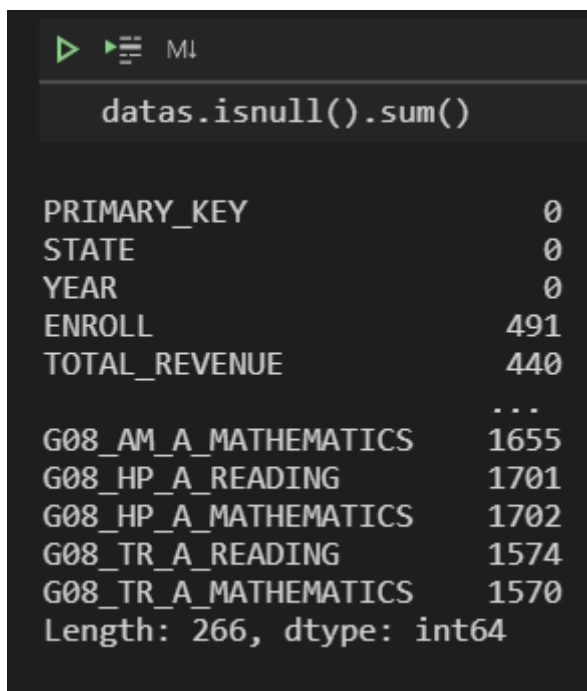


```
data.head()
```

	PRIMARY_KEY	STATE	YEAR	ENROLL	TOTAL_REVENUE	FEDERAL_REVENUE	STATE_REVENUE	LOCAL_REVENUE	TOTAL_EXPENDITURE	INSTRUCTION_EXPENDITURE	...	G08_HI_A_READING	G08_HI_A_MATHEMATICS	G08_AS_A_READING	G08
0	1992_ALABAMA	ALABAMA	1992	NaN	2678885.0	384177.0	1659028.0	715680.0	2653798.0	1481703.0	...	NaN	NaN	NaN	
1	1992_ALASKA	ALASKA	1992	NaN	1049591.0	106780.0	720711.0	222100.0	972488.0	498362.0	...	NaN	NaN	NaN	
2	1992_ARIZONA	ARIZONA	1992	NaN	3258079.0	297888.0	1369815.0	1590376.0	3401580.0	1435908.0	...	NaN	NaN	NaN	
3	1992_ARKANSAS	ARKANSAS	1992	NaN	1711959.0	178571.0	958785.0	574603.0	1743022.0	964323.0	...	NaN	NaN	NaN	
4	1992_CALIFORNIA	CALIFORNIA	1992	NaN	26260025.0	2072470.0	16546514.0	7041041.0	27138832.0	14358922.0	...	NaN	NaN	NaN	

5 rows x 266 columns

Пропуски в наборе:



```
data.isnull().sum()
```

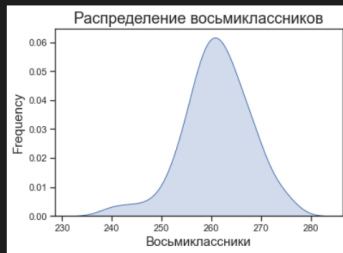
PRIMARY_KEY	0
STATE	0
YEAR	0
ENROLL	491
TOTAL_REVENUE	440
...	
G08_AM_A_MATHEMATICS	1655
G08_HP_A_READING	1701
G08_HP_A_MATHEMATICS	1702
G08_TR_A_READING	1574
G08_TR_A_MATHEMATICS	1570
Length: 266, dtype: int64	

В категориальных пропусков нет. Обработаем два количественных.

Категориальные признаки - это PRIMARY KEY, STATE, YEAR, однако в них пропусков нет. Обработаем количественный признак, например оценки по математике восьмиклассников, определенных как американские индейцы или коренные жители Аляски.

```
g = sns.kdeplot(data=datas, x="G08_AM_A_MATHEMATICS", shade=True)
g.set_xlabel("Восьмиклассники", size = 15)
g.set_ylabel("Frequency", size = 15)
plt.title("Распределение восьмиклассников", size = 18)
```

Text(0.5, 1.0, 'Распределение восьмиклассников')



## Используем моду

```
# Используем моду
indicator = MissingIndicator()
mask_missing_values_only = indicator.fit_transform(datas[['G08_AM_A_MATHEMATICS']])
imp_num = SimpleImputer(strategy='most_frequent')
data_num_imp = imp_num.fit_transform(datas[['G08_AM_A_MATHEMATICS']])
datas['G08_AM_A_MATHEMATICS'] = data_num_imp
filled_data = data_num_imp[mask_missing_values_only]
print('G08_AM_A_MATHEMATICS', 'most_frequent', filled_data.size, filled_data[0], filled_data[filled_data.size-1], sep='; ')
```

G08\_AM\_A\_MATHEMATICS; most\_frequent; 1655; 260.0; 260.0

## Для признака с процентом пропусков 99% удалим

Также, так как категориального мы не нашли, обработаем ещё один численный. Пусть это будет признак - средние оценки по математике по восьмиклассникам, определенным как жители острова Гавайи или других островов Тихого океана.

```
# Так как процент пропусков у этого признака целых 99%, то просто удалим его
datas.drop(['G08_HP_A_MATHEMATICS'], axis=1, inplace=True)
```

## В итоге:

```
# Просто проверим что замена на моду для предыдущего признака прошла успешно
datas['G08_AM_A_MATHEMATICS'].isnull().sum()

0
```

## Выводы

В данной работе для обработки пропусков данных мы воспользовались двумя стратегиями: **1)** удаление признака, содержащего большое количество пропусков (**99%**); **2)** импутация данных в признаке путем заполнения наиболее часто встречаемым значением

Из представленных выше признаков также стоит отбросить признаки с процентами пропусков от **75%**, а таковых очень много: удаление строк привело бы к серьезной потере размера датасета, а заполнение пропусков привело бы к возможному нарушению набора данных (неправильные данные). Так как из **266** признаков набора данных абсолютное большинство - это признаки с процентом пропусков больше **50** и выше, а сами признаки - это средние оценки каких либо групп учеников с очень высокой детализацией, то имеет смысл вообще удалить все ненужные признаки и оставить лишь те, которые нужны для модели, поскольку тогда размер набора уменьшится значительно.

Окончательное решение по выбору признаков, поступающих на вход модели, может приниматься после проведения корреляционного анализа. Также после проведения кросс-

валидации и подбора оптимальных параметров модели возможен пересмотр набора призна

аков: либо их удаление, либо их добавление в зависимости от результатов работы алгоритма машинного обучения