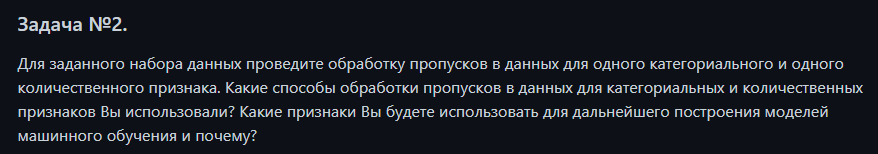
Лялин Александр, ИУ5-64Б, РК1

Вариант 13 – задача 2 – набор данных 5



**Условие:**





**Импортируем библиотеки:**

import pandas as pd

import matplotlib.pyplot as plt

import seaborn as sns

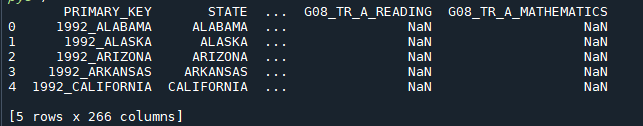
import numpy as np

**Смотрим общую информацию:**

file = 'states\_all\_extended.csv'

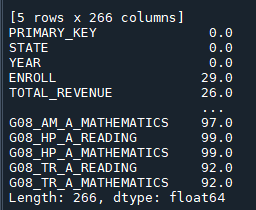
data = pd.read\_csv(file, sep=",")

print(data.head())



**Процент пропусков в каждом столбце:**

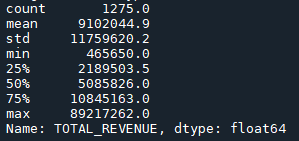
print(round(data.isna().sum()/(data.count()+data.isna().sum()), 2) \* 100)



**Заполним пропуски в TOTAL\_REVENUE**

pd.set\_option('display.float\_format', lambda x: '%.1f' % x)

print(data.TOTAL\_REVENUE.describe())



*Заполнение тремя разными способами:*

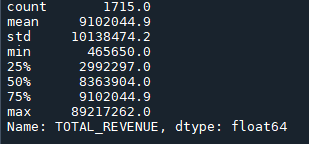
Матожидание - подходит для нормального распеделения, но плохо реагирует на выбросы и аномалии.

from sklearn.impute import SimpleImputer

imp = SimpleImputer(missing\_values=np.nan, strategy='mean')

data['TOTAL\_REVENUE'] = imp.fit\_transform(data[['TOTAL\_REVENUE']])

print(data.TOTAL\_REVENUE.describe())

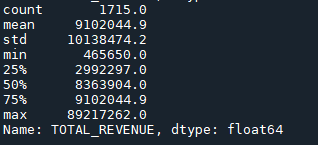


Медиана - оптимальная для набора данных с простыми связями между признаками. Не реагирует на выбросы.

imp = SimpleImputer(missing\_values=np.nan, strategy='median')

data['TOTAL\_REVENUE'] = imp.fit\_transform(data[['TOTAL\_REVENUE']])

print(data.TOTAL\_REVENUE.describe())

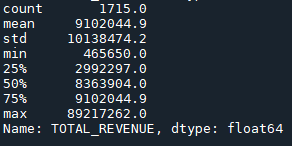


Мода - самое частое значение, для некатегориального признака не очень подходит.

imp = SimpleImputer(missing\_values=np.nan, strategy='most\_frequent')

data['TOTAL\_REVENUE'] = imp.fit\_transform(data[['TOTAL\_REVENUE']])

print(data.TOTAL\_REVENUE.describe())



**Вывод**: для дальнейшего построения моделей машинного обучения придется убрать столбцы с пропусками более 90%. Возможно, придется убрать столбцы TOTAL, чтобы избежать перевешивания этих признаков, так как TOTAL - сумма нескольких соответствующих признаков.

**Построим скрипичную диаграмму по столбцу LOCAL\_REVENUE**

print(sns.violinplot(x=data['LOCAL\_REVENUE']))

