**Рубежный контроль №1  
Студент: Попов М.Ю.  
Группа: ИУ5-25М**

**Вариант 10**

Задача №10

Для набора данных проведите устранение пропусков для одного (произвольного) категориального признака с использованием метода заполнения наиболее распространенным значением.

Задача №30

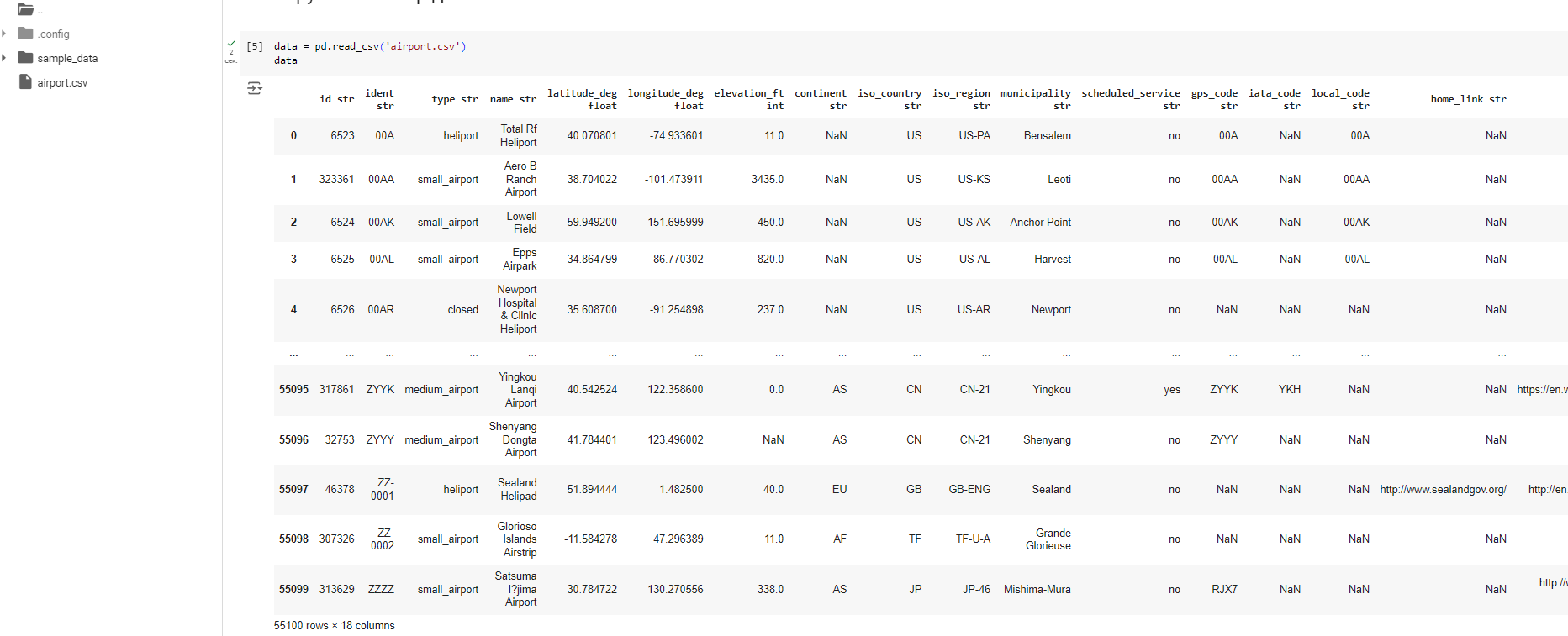
Для набора данных проведите удаление повторяющихся признаков.

Импортируем библиотеки

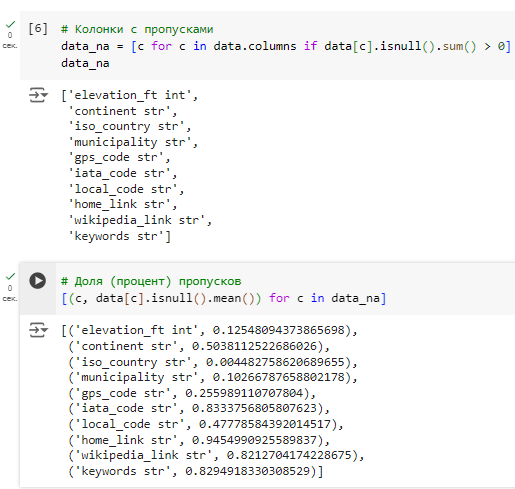
import pandas as pd  
import matplotlib.pyplot as plt  
import numpy as np  
from sklearn.impute import MissingIndicator, SimpleImputer  
import seaborn as sns

def plot\_hist\_diff(old\_ds, new\_ds, cols):  
    """  
    Разница между распределениями до и после устранения пропусков  
    """  
    for c in cols:     
        fig = plt.figure()  
        ax = fig.add\_subplot(111)  
        ax.title.set\_text('Поле - ' + str(c))  
        old\_ds[c].hist(bins=50, ax=ax, density=True, color='green')  
        new\_ds[c].hist(bins=50, ax=ax, color='blue', density=True, alpha=0.5)  
        plt.show()

Загружаем набор данных



В качестве исходных данных был выбран набор данных 'airport.csv'. В нём некоторые признаки содержат пропуски:



Задание 1

Для устранения пропусков наиболее распространенным значением выберем категориальный признак аббривеатуры континента "continent str". Данные этого признака не являются главной информацией о полете и контенентов не так много, поэтому большиство стран находятся на одном континенте.

column\_for\_fill=['continent str']

changed\_data\_1 = data[column\_for\_fill].copy()

def impute\_column(dataset, column, strategy\_param, fill\_value\_param=None):

    """

    Заполнение пропусков в одном признаке

    """

    temp\_data = dataset[[column]].values

    size = temp\_data.shape[0]

    indicator = MissingIndicator()

    mask\_missing\_values\_only = indicator.fit\_transform(temp\_data)

    imputer = SimpleImputer(strategy=strategy\_param,

                            fill\_value=fill\_value\_param)

    all\_data = imputer.fit\_transform(temp\_data)

    missed\_data = temp\_data[mask\_missing\_values\_only]

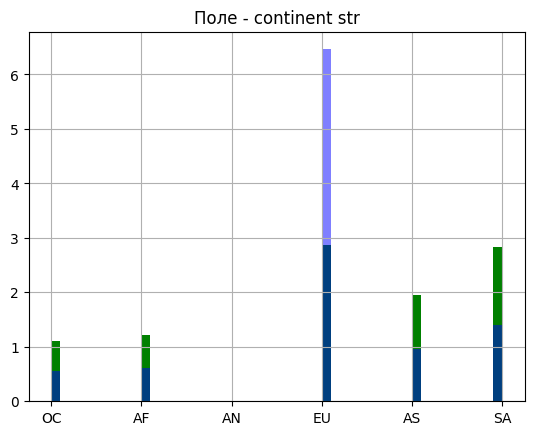
    filled\_data = all\_data[mask\_missing\_values\_only]

    return all\_data.reshape((size,)), filled\_data, missed\_data

elevation\_new, \_, \_ = impute\_column(changed\_data\_1, 'continent str', 'most\_frequent')

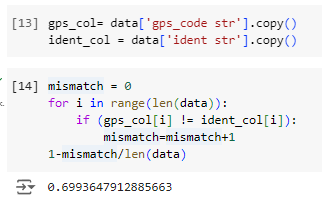
changed\_data\_1['continent str'] = elevation\_new

plot\_hist\_diff(data, changed\_data\_1, column\_for\_fill)



Задание 2

При визуализации части данных набора в начале первого задания была заметна схожесть значений признаков "gps\_code str" и "ident str". Проверим, насколько они отличаются.



Мы видим, что значения столбцов совпадают на 70 процентов. Причем 26% всех записей признака "gps\_code str" просто пропущены. Можно предположить что они и составляют большую долю несовпадений. Эти наблюдения позволяют принять решение об удалении столбца "indent str"

changed\_data\_2 = data.drop(['ident str'], axis=1)

