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.

Решение

Загрузка набора данных (Пассажиры Титаника)

```
import pandas as pd

# Загрузка данных
df = pd.read_csv('Titanic Dataset.csv')

[10] ✓ 0.0s

Python
```

Данные датасета:



Информация по столбцам набора данных

```
# оценим информацию о столбцах датасета
         df.info()
[12] \( \square 0.0s
... <class 'pandas.core.frame.DataFrame'>
      RangeIndex: 1309 entries, 0 to 1308
      Data columns (total 14 columns):
      # Column Non-Null Count Dtype
                       1309 non-null int64
       0 pclass
      1 survived 1309 non-null int64
2 name 1309 non-null object
3 sex 1309 non-null object
       4 age 1046 non-null float64
5 sibsp 1309 non-null int64
       6 parch 1309 non-null int64
7 ticket 1309 non-null object
                     1308 non-null float64
295 non-null object
       8 fare
       9 cabin
       10 embarked 1307 non-null object
11 boat 486 non-null object
       11 boat
       12 body 121 non-null float64
13 home.dest 745 non-null object
      dtypes: float64(3), int64(4), object(7)
      memory usage: 143.3+ KB
```

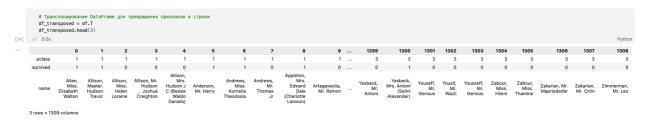
Количество пропусков по столбцам:

```
p v na_mask = df.isna()
       na_counts = na_mask.sum()
      na_counts
[16] ✓ 0.0s
··· pclass
    name
    sex
                 a
              263
    age
    sibsp
    parch
    ticket
    fare
                 1
    cabin
               1014
    embarked
               823
    boat
    body
    home.dest 564
    dtype: int64
```

Задача 10 (выявляем самое частое значение и заполняем пропуски)

Проверка:

Задача 30



Удаление повторений

```
# Удаление дублирующихся строк (признаков)
df_transposed.drop_duplicates(inplace=True)

✓ 0.0s

# Обратное транспонирование для возвращения к исходному виду
df = df_transposed.T

[26] ✓ 0.0s
```

Проверка

```
df.info()
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1309 entries, 0 to 1308
Data columns (total 14 columns):
# Column Non-Null Count Dtype
               survived 1309 non-null
name 1309 non-null
                                                             object
                                  1309 non-null
1309 non-null
             sex
age
sibsp
                                                             object
                                  1046 non-null
1309 non-null
                                                             object
               parch
ticket
                                  1309 non-null
                                                             object
object
                                  1309 non-null
        9 cabin 295 non-null
10 embarked 1307 non-null
11 boat 486 non-null
12 body 121 a
                                                             object
                                                             object
        13 home.dest 1309 non-null object dtypes: object(14)
        memory usage: 143.3+ KB
```

		df.head	()												
[27]	~	0.0s													
		pclass	survived	name	sex	age	sibsp	parch	ticket	fare	cabin	embarked	boat	body	home.dest
	0	1	1	Allen, Miss. Elisabeth Walton	female	29.0	0	0	24160	211.3375	B5	s	2	NaN	St Louis, MO
	1	1	1	Allison, Master. Hudson Trevor	male	0.92	1	2	113781	151.55	C22 C26	S	11	NaN	Montreal, PQ / Chesterville, ON
	2	1	0	Allison, Miss. Helen Loraine	female	2.0	1	2	113781	151.55	C22 C26	S	NaN	NaN	Montreal, PQ / Chesterville, ON
	3	1	0	Allison, Mr. Hudson Joshua Creighton	male	30.0	1	2	113781	151.55	C22 C26	S	NaN	135.0	Montreal, PQ / Chesterville, ON
	4	1	0	Allison, Mrs. Hudson J C (Bessie Waldo Daniels)	female	25.0	1	2	113781	151.55	C22 C26	s	NaN	NaN	Montreal, PQ / Chesterville, ON