РК1 ММО Савельев Алексей ИУ5-24М 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.

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:

```
# вывод данных датасета
df.head(5)
```

[11] ✓ 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	29.00	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	151.5500	C22 C26	S	11	NaN	Montreal, PQ / Chesterville, ON
2	1	0	Allison, Miss. Helen Loraine	female	2.00	1	2	113781	151.5500	C22 C26	S	NaN	NaN	Montreal, PQ / Chesterville, ON
3	1	0	Allison, Mr. Hudson Joshua Creighton	male	30.00	1	2	113781	151.5500	C22 C26	S	NaN	135.0	Montreal, PQ / Chesterville, ON
4	1	0	Allison, Mrs. Hudson J C (Bessie Waldo Daniels)	female	25.00	1	2	113781	151.5500	C22 C26	S	NaN	NaN	Montreal, PQ / Chesterville, ON

Информация по столбцам набора данных

```
# оценим информацию о столбцах датасета
df.info()
```

[12] ✓ 0.0s

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1309 entries, 0 to 1308
Data columns (total 14 columns):
 #   Column     Non-Null Count  Dtype
---  ------     --------------  -----
 0   pclass     1309 non-null   int64
 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
 8   fare       1308 non-null   float64
 9   cabin      295 non-null    object
 10  embarked   1307 non-null   object
 11  boat       486 non-null    object
 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:

```
na_mask = df.isna()
na_counts = na_mask.sum()
na_counts
```

[16] ✓ 0.0s

```
pclass          0
survived        0
name            0
sex             0
age           263
sibsp           0
parch           0
ticket          0
fare            1
cabin        1014
embarked        2
boat          823
body         1188
home.dest     564
dtype: int64
```

Задача 10 (выявляем самое частое значение и заполняем пропуски)

```
# получаем самое частоповторяемое значение в столбце
most_common = df['home.dest'].mode()[0]
most_common
```

[17] ✓ 0.0s

```
'New York, NY'
```

```
# заполняем все пустые значения самым частым
df['home.dest'].fillna(most_common, inplace=True)
```

[18] ✓ 0.0s

Проверка:

```
df.info()
```

[19] ✓ 0.0s Python

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1309 entries, 0 to 1308
Data columns (total 14 columns):
 #   Column     Non-Null Count  Dtype
---  ------     --------------  -----
 0   pclass     1309 non-null   int64
 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
 8   fare       1308 non-null   float64
 9   cabin      295 non-null    object
 10  embarked   1307 non-null   object
 11  boat       486 non-null    object
 12  body       121 non-null    float64
 13  home.dest  1309 non-null   object
dtypes: float64(3), int64(4), object(7)
memory usage: 143.3+ KB
```

```
na_mask = df.isna()
na_counts = na_mask.sum()
na_counts
```

[20] ✓ 0.0s Python

```
pclass          0
survived        0
name            0
sex             0
age           263
sibsp           0
parch           0
ticket          0
fare            1
cabin        1014
embarked        2
boat          823
body         1188
home.dest       0
dtype: int64
```

Задача 30

```
# Транспонирование DataFrame для превращения признаков в строки
df_transposed = df.T
df_transposed.head(3)
```

[24] ✓ 0.0s Python

	0	1	2	3	4	5	6	7	8	9	...	1299	1300	1301	1302	1303	1304	1305	1306	1307	1308
pclass	1	1	1	1	1	1	1	1	1	1	...	3	3	3	3	3	3	3	3	3	3
survived	1	1	0	0	0	1	1	0	1	0	...	0	1	0	0	0	0	0	0	0	0
name	Allen, Miss. Elisabeth Walton	Allison, Master. Hudson Trevor	Allison, Miss. Helen Loraine	Allison, Mr. Hudson Joshua Creighton	Allison, Mrs. Hudson J C (Bessie Waldo Daniels)	Anderson, Mr. Harry	Andrews, Miss. Kornelia Theodosia	Andrews, Mr. Thomas Jr	Appleton, Mrs. Edward Dale (Charlotte Lamson)	Artagaveytia, Mr. Ramon	...	Yasbeck, Mr. Antoni	Yasbeck, Mrs. Antoni (Selini Alexander)	Youseff, Mr. Gerious	Yousif, Mr. Wazli	Yousseff, Mr. Gerious	Zabour, Miss. Hileni	Zabour, Miss. Thamine	Zakarian, Mr. Mapriededer	Zakarian, Mr. Ortin	Zimmerman, Mr. Leo

3 rows × 1309 columns

Удаление повторений

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

[25] ✓ 0.0s

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

[26] ✓ 0.0s

Проверка

```
df.info()
```

[28] ✓ 0.0s

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1309 entries, 0 to 1308
Data columns (total 14 columns):
 #   Column     Non-Null Count  Dtype
---  ------     --------------  -----
 0   pclass     1309 non-null   object
 1   survived   1309 non-null   object
 2   name       1309 non-null   object
 3   sex        1309 non-null   object
 4   age        1046 non-null   object
 5   sibsp      1309 non-null   object
 6   parch      1309 non-null   object
 7   ticket     1309 non-null   object
 8   fare       1308 non-null   object
 9   cabin      295 non-null    object
 10  embarked   1307 non-null   object
 11  boat       486 non-null    object
 12  body       121 non-null    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