File Tools Kernel View Run Help rk forum AA

```
РК №1 по "Технологиям машинного обучения"
Желанкина Анна
ИУ5-63
```

0.0s

РК №1 по "Технологиям машинного обучения"
Желанкина Анна
ИУ5-63

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
from datalore.plot import *
import sklearn
```

4.5s

```
url = 'https://archive.ics.uci.edu/ml/machine-learning-databases/meta-data/meta.data'
data = pd.read_csv(url, sep=',',header=None)
```

```
data
```

3.3s

	0	1	2	3	4	5	6	7	8	9
0	Aust_Credit	690	690	14	2	4	0	1.2623	0.1024	0.7713
1	Aust_Credit	690	690	14	2	4	0	1.2623	0.1024	0.7713
2	Aust_Credit	690	690	14	2	4	0	1.2623	0.1024	0.7713
3	Aust_Credit	690	690	14	2	4	0	1.2623	0.1024	0.7713
4	Aust_Credit	690	690	14	2	4	0	1.2623	0.1024	0.7713
5	Aust_Credit	690	690	14	2	4	0	1.2623	0.1024	0.7713
6	Aust_Credit	690	690	14	2	4	0	1.2623	0.1024	0.7713
7	Aust_Credit	690	690	14	2	4	0	1.2623	0.1024	0.7713
8	Aust_Credit	690	690	14	2	4	0	1.2623	0.1024	0.7713

```
for d in data:
    data[d].replace('?', np.NAN, inplace=True)

data
```

0.1s

	0	1	2	3	4	5	6	7	8	9
0	Aust_Credit	690	690	14	2	4	0	1.2623	0.1024	0.7713
1	Aust_Credit	690	690	14	2	4	0	1.2623	0.1024	0.7713
2	Aust_Credit	690	690	14	2	4	0	1.2623	0.1024	0.7713
3	Aust_Credit	690	690	14	2	4	0	1.2623	0.1024	0.7713
4	Aust_Credit	690	690	14	2	4	0	1.2623	0.1024	0.7713
5	Aust_Credit	690	690	14	2	4	0	1.2623	0.1024	0.7713
6	Aust_Credit	690	690	14	2	4	0	1.2623	0.1024	0.7713
7	Aust_Credit	690	690	14	2	4	0	1.2623	0.1024	0.7713
8	Aust_Credit	690	690	14	2	4	0	1.2623	0.1024	0.7713

```
data[12] = data[12].astype(float)
```

0.0s

```
categorical_columns = [c for c in data.columns if data[c].dtype.name == 'object']
numerical_columns = [c for c in data.columns if data[c].dtype.name != 'object']
print(categorical_columns)
print(numerical_columns)
```

0.1s

```
[0, 8, 10, 20]
[1, 2, 3, 4, 5, 6, 7, 9, 11, 12, 13, 14, 15, 16, 17, 18, 19, 21]
```

```
#%%
data.isnull().sum()
```

0.0s

+ Show All

```
0       0
1       0
2       0
3       0
4       0
5       0
6       0
7       0
8      24
9       0
10    240
11      0
12    240
13      0
14      0
15      0
16      0
17      0
18      0
```

```
num_cols = []
for col in data.columns:
    # Количество пустых значений
    temp_null_count = data[data[col].isnull()].shape[0]
    dt = str(data[col].dtype)
    if temp_null_count>0:
        num_cols.append(col)
        temp_perc = round((temp_null_count / data.shape[0]) * 100.0, 2)
        print('Колонка {}. Тип данных {}. Количество пустых значений {}, {}%.'.format(col, dt, temp_null_count, temp_perc))
```

0.3s

```
Колонка 8. Тип данных object. Количество пустых значений 24, 4.55%.
Колонка 10. Тип данных object. Количество пустых значений 240, 45.45%.
Колонка 12. Тип данных float64. Количество пустых значений 240, 45.45%.
```

```
data[8].describe()
```

0.0s

```
count        504
unique        21
top       0.2178
freq          24
Name: 8, dtype: object
```

```
from sklearn.preprocessing import Imputer

data[8] = Imputer(missing_values=np.nan, strategy='most_frequent').fit_transform(data[[8]])
```

0.1s

```
data
```

0.1s

	0	1	2	3	4	5	6	7	8	9
0	Aust_Credit	690	690	14	2	4	0	1.2623	0.1024	0.7713
1	Aust_Credit	690	690	14	2	4	0	1.2623	0.1024	0.7713
2	Aust_Credit	690	690	14	2	4	0	1.2623	0.1024	0.7713

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

0.0s

+ Show All

```
3       0
4       0
5       0
6       0
7       0
8       0
9       0
10    240
11      0
12    240
13      0
14      0
15      0
16      0
17      0
18      0
19      0
20      0
```

```
12    240
13      0
14      0
15      0
16      0
17      0
18      0
19      0
20      0
21      0
dtype: int64
```

```
data = data.drop((12), axis=1)
```

0.3s