dsbda-a1-a2

May 22, 2023

[11]:
```
import pandas as pd
import numpy as np
```

[12]:
```
df = pd.read_csv("test.csv")
df.head()
```

[12]:

	PassengerId	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
0	892	3	Kelly, Mr. James	male	34.5	0	0	330911	7.8292	NaN	Q
1	893	3	Wilkes, Mrs. James (Ellen Needs)	female	47.0	1	0	363272	7.0000	NaN	S
2	894	2	Myles, Mr. Thomas Francis	male	62.0	0	0	240276	9.6875	NaN	Q
3	895	3	Wirz, Mr. Albert	male	27.0	0	0	315154	8.6625	NaN	S
4	896	3	Hirvonen, Mrs. Alexander (Helga E Lindqvist)	female	22.0	1	1	3101298	12.2875	NaN	S

[13]:
```
column={}
for x in df.columns:
    column[x] = df[x].isnull().any()
print(column)
df.describe()
```

```
{'PassengerId': False, 'Pclass': False, 'Name': False, 'Sex': False, 'Age':
True, 'SibSp': False, 'Parch': False, 'Ticket': False, 'Fare': True, 'Cabin':
True, 'Embarked': False}
```

[13]:

	PassengerId	Pclass	Age	SibSp	Parch	Fare
count	418.000000	418.000000	332.000000	418.000000	418.000000	417.000000
mean	1100.500000	2.265550	30.272590	0.447368	0.392344	35.627188
std	120.810458	0.841838	14.181209	0.896760	0.981429	55.907576
min	892.000000	1.000000	0.170000	0.000000	0.000000	0.000000
25%	996.250000	1.000000	21.000000	0.000000	0.000000	7.895800
50%	1100.500000	3.000000	27.000000	0.000000	0.000000	14.454200
75%	1204.750000	3.000000	39.000000	1.000000	0.000000	31.500000

```
max    1309.000000    3.000000   76.000000    8.000000    9.000000  512.329200
```

[14]:
```
df.shape
```

[14]:
```
(418, 11)
```

[15]:
```
df.dtypes
```

[15]:
```
PassengerId      int64
Pclass           int64
Name            object
Sex             object
Age            float64
SibSp            int64
Parch            int64
Ticket          object
Fare           float64
Cabin           object
Embarked        object
dtype: object
```

[16]:
```
df = df.drop(['Cabin','Embarked'], axis=1)
df.head()
```

[16]:

	PassengerId	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare
0	892	3	Kelly, Mr. James	male	34.5	0	0	330911	7.8292
1	893	3	Wilkes, Mrs. James (Ellen Needs)	female	47.0	1	0	363272	7.0000
2	894	2	Myles, Mr. Thomas Francis	male	62.0	0	0	240276	9.6875
3	895	3	Wirz, Mr. Albert	male	27.0	0	0	315154	8.6625
4	896	3	Hirvonen, Mrs. Alexander (Helga E Lindqvist)	female	22.0	1	1	3101298	12.2875

[17]:
```
from sklearn.preprocessing import LabelEncoder
from sklearn.preprocessing import OneHotEncoder

le = LabelEncoder()

df["Name"] = le.fit_transform(df["Name"].values)
```

[18]:
```
df.dtypes
```

[18]:
```
PassengerId      int64
Pclass           int64
Name             int32
Sex             object
Age            float64
SibSp            int64
Parch            int64
Ticket          object
Fare           float64
dtype: object
```

[21]:
```
from sklearn.preprocessing import LabelEncoder
from sklearn.preprocessing import OneHotEncoder

le = LabelEncoder()


df["Age"]=le.fit_transform(df["Age"].values)
```

[22]:
```
df["Age"]
```

[22]:
```
0      44
1      60
2      74
3      34
4      27
       ..
413    79
414    51
415    50
416    79
417    79
Name: Age, Length: 418, dtype: int64
```

[25]:
```
df["Name"]
```

[25]:
```
0                                    Kelly, Mr. James
1                    Wilkes, Mrs. James (Ellen Needs)
2                           Myles, Mr. Thomas Francis
3                                    Wirz, Mr. Albert
4        Hirvonen, Mrs. Alexander (Helga E Lindqvist)
                              ...
413                             Spector, Mr. Woolf
414                   Oliva y Ocana, Dona. Fermina
415                   Saether, Mr. Simon Sivertsen
416                            Ware, Mr. Frederick
417                       Peter, Master. Michael J
```

```
Name: Name, Length: 418, dtype: object
```

[30]:
```
df["Name"]=df["Name"].astype("string")
```

[31]:
```
df.dtypes
```

[31]:
```
PassengerId      int64
Pclass           int64
Name            string
Sex             object
Age              int64
SibSp            int64
Parch            int64
Ticket          object
Fare           float64
Cabin           object
Embarked        object
dtype: object
```

[ ]:
```

```