```
import pandas as pd
import numpy as np

df = pd.read_csv('train.csv')
df.head()
```

```
   PassengerId  Survived  Pclass  \
0            1         0       3
1            2         1       1
2            3         1       3
3            4         1       1
4            5         0       3

                                                Name     Sex   Age
SibSp  \
0                            Braund, Mr. Owen Harris    male  22.0
1
1  Cumings, Mrs. John Bradley (Florence Briggs Th...  female  38.0
1
2                             Heikkinen, Miss. Laina  female  26.0
0
3       Futrelle, Mrs. Jacques Heath (Lily May Peel)  female  35.0
1
4                           Allen, Mr. William Henry    male  35.0
0

   Parch            Ticket     Fare Cabin Embarked
0      0         A/5 21171   7.2500   NaN        S
1      0          PC 17599  71.2833   C85        C
2      0  STON/O2. 3101282   7.9250   NaN        S
3      0            113803  53.1000  C123        S
4      0            373450   8.0500   NaN        S
```

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 891 entries, 0 to 890
Data columns (total 12 columns):
 #   Column       Non-Null Count  Dtype
---  ------       --------------  -----
 0   PassengerId  891 non-null    int64
 1   Survived     891 non-null    int64
 2   Pclass       891 non-null    int64
 3   Name         891 non-null    object
 4   Sex          891 non-null    object
 5   Age          714 non-null    float64
 6   SibSp        891 non-null    int64
 7   Parch        891 non-null    int64
 8   Ticket       891 non-null    object
 9   Fare         891 non-null    float64
 10  Cabin        204 non-null    object
 11  Embarked     889 non-null    object
dtypes: float64(2), int64(5), object(5)
memory usage: 66.2+ KB
```

```
df.describe()
```

```
       PassengerId    Survived      Pclass         Age       SibSp  \
count   891.000000  891.000000  891.000000  714.000000  891.000000
mean    446.000000    0.383838    2.308642   29.699118    0.523008
std     257.353842    0.486592    0.836071   14.526497    1.102743
min       1.000000    0.000000    1.000000    0.420000    0.000000
25%     223.500000    0.000000    2.000000   20.125000    0.000000
50%     446.000000    0.000000    3.000000   28.000000    0.000000
75%     668.500000    1.000000    3.000000   38.000000    1.000000
max     891.000000    1.000000    3.000000   80.000000    8.000000

            Parch        Fare
count  891.000000  891.000000
mean     0.381594   32.204208
std      0.806057   49.693429
min      0.000000    0.000000
25%      0.000000    7.910400
50%      0.000000   14.454200
75%      0.000000   31.000000
max      6.000000  512.329200
```

```
df.isnull().sum()
```

```
PassengerId      0
Survived         0
Pclass           0
Name             0
Sex              0
Age            177
SibSp            0
Parch            0
Ticket           0
Fare             0
Cabin          687
Embarked         2
dtype: int64
```

```
df.dropna(subset=['Embarked'],inplace=True)
df.shape
```

```
(889, 12)
```

```
def fill_age(age , sibsp):
    if pd.isna(age):
        return np.random.randint(25,76) if sibsp > 0 else
```

```
np.random.randint(0,25)
    else :
        return age
df['Age'] = df.apply(lambda row :
fill_age(row['Age'],row['SibSp']),axis = 1)
```

```
df.isnull().sum()
```

```
PassengerId      0
Survived         0
Pclass           0
Name             0
Sex              0
Age              0
SibSp            0
Parch            0
Ticket           0
Fare             0
Cabin          687
Embarked         0
dtype: int64
```

```
df['Age'].isnull().sum()
```

```
0
```

```
df['Cabin'].isna().sum()/889*100
```

```
77.27784026996626
```

```
df.dropna(axis = 1, inplace = True)
```

```
df.head()
```

```
   PassengerId  Survived  Pclass  \
0            1         0       3
1            2         1       1
2            3         1       3
3            4         1       1
4            5         0       3

                                                Name     Sex   Age
SibSp  \
0                            Braund, Mr. Owen Harris    male  22.0
1
1  Cumings, Mrs. John Bradley (Florence Briggs Th...  female  38.0
1
2                             Heikkinen, Miss. Laina  female  26.0
0
3       Futrelle, Mrs. Jacques Heath (Lily May Peel)  female  35.0
1
```

```
4                           Allen, Mr. William Henry    male  35.0
0

   Parch            Ticket     Fare Embarked
0      0         A/5 21171   7.2500        S
1      0          PC 17599  71.2833        C
2      0  STON/O2. 3101282   7.9250        S
3      0            113803  53.1000        S
4      0            373450   8.0500        S

df.dtypes

PassengerId      int64
Survived         int64
Pclass           int64
Name            object
Sex             object
Age            float64
SibSp            int64
Parch            int64
Ticket          object
Fare           float64
Embarked        object
dtype: object

df['Sex'].value_counts()

male      577
female    312
Name: Sex, dtype: int64

new = df.copy()
new['Sex'] = new['Sex'].map({'male' : 1, 'female' : 0 })

from sklearn.preprocessing import LabelEncoder

le = LabelEncoder()
new['Embarked'] = le.fit_transform(df['Embarked'])
new.head(5)

   PassengerId  Survived  Pclass  \
0            1         0       3
1            2         1       1
2            3         1       3
3            4         1       1
4            5         0       3

                                                Name  Sex   Age  SibSp
Parch  \
0                            Braund, Mr. Owen Harris    1  22.0      1
0
```

```
1  Cumings, Mrs. John Bradley (Florence Briggs Th...    0  38.0      1
0
2                              Heikkinen, Miss. Laina    0  26.0      0
0
3       Futrelle, Mrs. Jacques Heath (Lily May Peel)    0  35.0      1
0
4                            Allen, Mr. William Henry    1  35.0      0
0

             Ticket     Fare  Embarked
0         A/5 21171   7.2500         2
1          PC 17599  71.2833         0
2  STON/O2. 3101282   7.9250         2
3            113803  53.1000         2
4            373450   8.0500         2
```

```
new.describe()
```

```
       PassengerId    Survived      Pclass         Sex         Age  \
count   889.000000  889.000000  889.000000  889.000000  889.000000
mean    446.000000    0.382452    2.311586    0.649044   27.704353
std     256.998173    0.486260    0.834700    0.477538   15.832678
min       1.000000    0.000000    1.000000    0.000000    0.000000
25%     224.000000    0.000000    2.000000    0.000000   18.000000
50%     446.000000    0.000000    3.000000    1.000000   26.000000
75%     668.000000    1.000000    3.000000    1.000000   37.000000
max     891.000000    1.000000    3.000000    1.000000   80.000000

            SibSp       Parch        Fare    Embarked
count  889.000000  889.000000  889.000000  889.000000
mean     0.524184    0.382452   32.096681    1.535433
std      1.103705    0.806761   49.697504    0.792088
min      0.000000    0.000000    0.000000    0.000000
25%      0.000000    0.000000    7.895800    1.000000
50%      0.000000    0.000000   14.454200    2.000000
75%      1.000000    0.000000   31.000000    2.000000
max      8.000000    6.000000  512.329200    2.000000
```

```
new.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 889 entries, 0 to 890
Data columns (total 11 columns):
 #   Column       Non-Null Count  Dtype
---  ------       --------------  -----
 0   PassengerId  889 non-null    int64
 1   Survived     889 non-null    int64
 2   Pclass       889 non-null    int64
 3   Name         889 non-null    object
 4   Sex          889 non-null    int64
 5   Age           889 non-null    float64
 6   SibSp         889 non-null    int64
 7   Parch         889 non-null    int64
 8   Ticket        889 non-null    object
 9   Fare          889 non-null    float64
 10  Embarked      889 non-null    int32
dtypes: float64(2), int32(1), int64(6), object(2)
memory usage: 72.9+ KB
```