```
import numpy as np
```

```
import pandas as pd
```

```
import os
```

```
for dirname, _, filenames in os.walk('/kaggle/input'):
    for filename in filenames:
        print(os.path.join(dirname, filename))
```

```
import matplotlib.pyplot as plt
```

```
import seaborn as sns
```

```
import plotly.express as px
```

```
import numpy as np
```

```
from sklearn.preprocessing import LabelEncoder
```

```
df_train = pd.read_csv("/content/titanic.csv")
```

```
print("Training Data is: \n",df_train.head())
```

```
Training Data is:
    PassengerId  Survived  Pclass                                Name     Sex  \
0          892         0       3                    Kelly, Mr. James    male
1          893         1       3  Wilkes, Mrs. James (Ellen Needs)  female
2          894         0       2         Myles, Mr. Thomas Francis    male
3          895         0       3                  Wirz, Mr. Albert    male
4          896         1       3                  Wirz, Mr. Albert  female

    Age  SibSp  Parch   Ticket     Fare Cabin Embarked
0  34.5      0      0   330911   7.8292   NaN        Q
1  47.0      1      0   363272   7.0000   NaN        S
2  62.0      0      0   240276   9.6875   NaN        Q
3  27.0      0      0   315154   8.6625   NaN        S
4  22.0      1      1  3101298  12.2875   NaN        S
```

```
print("Missing Values: ")
```

```
Missing Values:
```

```
df_train.isnull().sum()
```

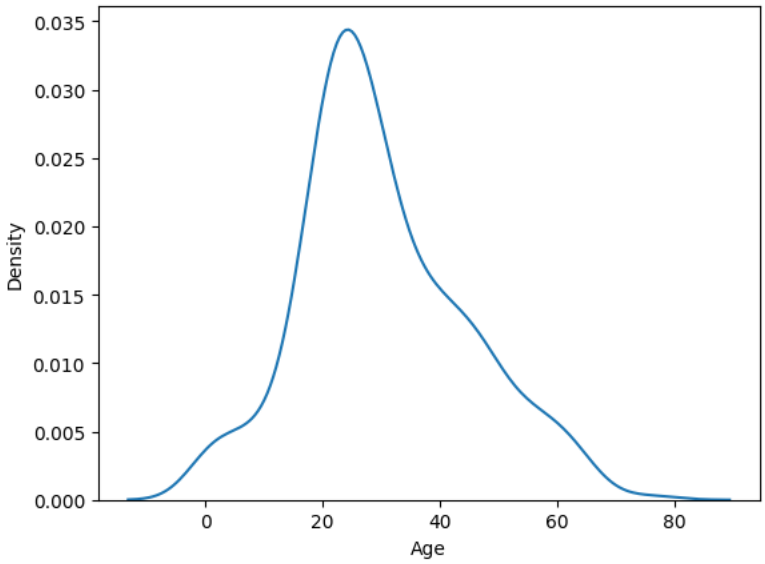
```
PassengerId      0
Survived         0
Pclass           0
Name             0
Sex              0
Age             86
SibSp            0
Parch            0
Ticket           0
Fare             1
Cabin          327
Embarked         0
dtype: int64
```

```
print("Let's see age distribution")
```

```
Let's see age distribution
```

```
sns.kdeplot(df_train['Age'])
```

```
<Axes: xlabel='Age', ylabel='Density'>
```




```
print("Let's see cabin distribution")
```

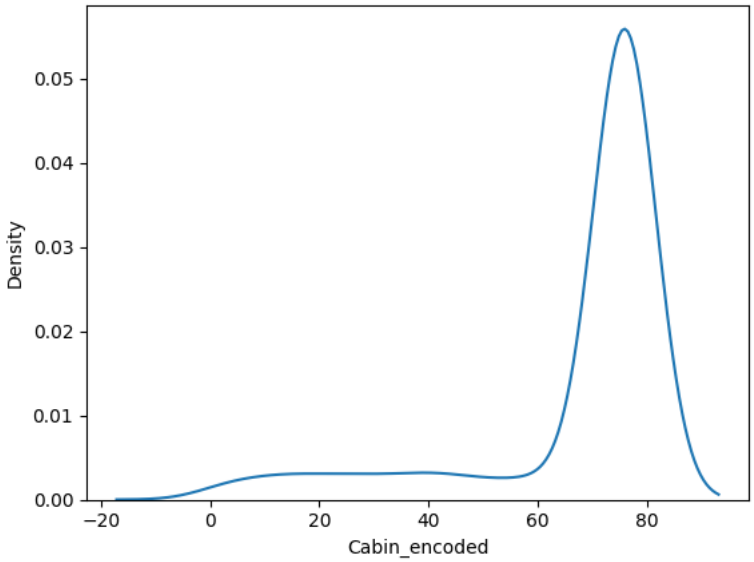
```
Let's see cabin distribution
```

```
label_encoder = LabelEncoder()
```

```
df_train['Cabin_encoded'] = label_encoder.fit_transform(df_train['Cabin'])
```

```
sns.kdeplot(df_train['Cabin_encoded'])
```

```
<Axes: xlabel='Cabin_encoded', ylabel='Density'>
```




```
print("Let's see embarked distribution")
```

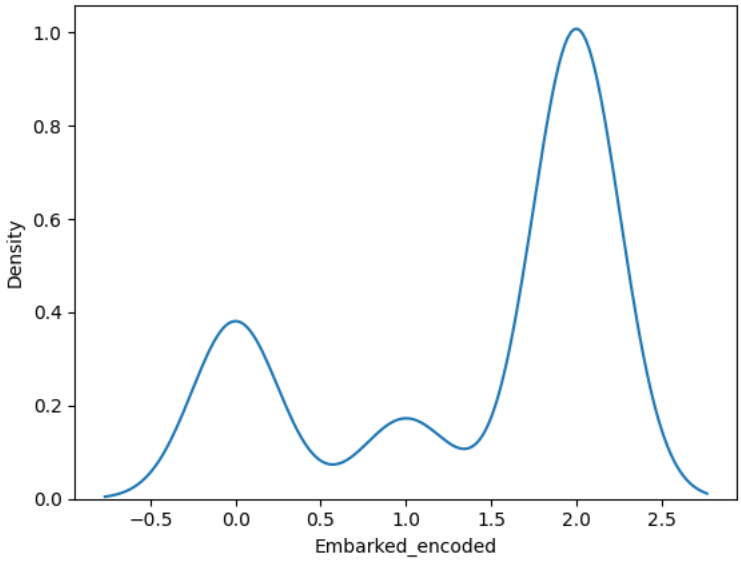
```
Let's see embarked distribution
```

```
label_encoder = LabelEncoder()
```

```
df_train["Embarked_encoded"] = label_encoder.fit_transform(df_train["Embarked"])
```

```
sns.kdeplot(df_train["Embarked_encoded"])
```

```
<Axes: xlabel='Embarked_encoded', ylabel='Density'>
```




```
print("Let's adjust age")
```

```
Let's adjust age
```

```
non_null_age = len(df_train['Age']) - df_train['Age'].isnull().sum()
```

```
mean_age = (df_train['Age'].sum())/non_null_age
```

```
median_age = df_train['Age'].median(skipna = True)
```

```
print("Mean value age :", mean_age)
print("Median value age :", median_age)
```

```
Mean value age : 30.272590361445783
Median value age : 27.0
```

```
train_data = df_train.copy()
```

```
train_data["Age"].fillna(df_train["Age"].median(skipna=True), inplace=True)
```

```
print("Let's adjust embarked")
```

```
Let's adjust embarked
```

```
train_data["Embarked"].fillna(df_train['Embarked'].value_counts().idxmax(), inplace = True)
```

```
train_data.head()
```

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked	Cabin_encoded	Embarked_encod
0	892	0	3	Kelly, Mr. James	male	34.5	0	0	330911	7.8292	NaN	Q	76	
1	893	1	3	Wilkes, Mrs. James (Ellen Needs)	female	47.0	1	0	363272	7.0000	NaN	S	76	

Next steps: Generate code with train_data View recommended plots

```
print("Let's adjust cabin")
```

```
Let's adjust cabin
```

```
train_data.drop('Cabin', axis = 1, inplace = True)
```

```
train_data.head()
```

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Embarked	Cabin_encoded	Embarked_encoded
0	892	0	3	Kelly, Mr. James	male	34.5	0	0	330911	7.8292	Q	76	1
1	893	1	3	Wilkes, Mrs. James (Ellen Needs)	female	47.0	1	0	363272	7.0000	S	76	2

Next steps: Generate code with `train_data` View recommended plots

```
train_data.isnull().sum()
```

```
PassengerId         0
Survived            0
Pclass              0
Name                0
Sex                 0
Age                 0
SibSp               0
Parch               0
Ticket              0
Fare                1
Embarked            0
Cabin_encoded       0
Embarked_encoded    0
dtype: int64
```

```
df_train.head()
```

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked	Cabin_encoded	Embarked_encod
0	892	0	3	Kelly, Mr. James	male	34.5	0	0	330911	7.8292	NaN	Q	76	
1	893	1	3	Wilkes, Mrs. James (Ellen Needs)	female	47.0	1	0	363272	7.0000	NaN	S	76	

Next steps: Generate code with `df_train` View recommended plots

Based on Age

```
unq_age = set(df_train['Age'])
surv = []
upd_unq_age = []

k = 0

for i in unq_age:
    sum_surv = 0
    upd_unq_age.append(i)
    for j,k in zip(df_train['Age'], df_train['Survived']):
        if(j == i):
            #print(j)
            sum_surv += k
        else:
            pass
            #print(j,k)
    surv.append(sum_surv)

print("surv is", surv)
```

```
surv is [0, 3, 1, 1, 0, 0, 0, 1, 1, 1, 0, 2, 0, 0, 1, 2, 2, 7, 1, 4, 3, 10, 5, 5, 0, 4, 4, 0, 1, 6, 2, 0, 3, 0, 3, 5, 3, 0, 2, 0, 0,
```