1. Đọc dữ liệu từ nhiều nơi

Từ local **from** **google.colab** **import** files

uploaded=files.upload()

Từ web **import** **pandas** **as** **pd**

pd.read\_csv("url đến dataset")

1. Xử lý missing

Kiểm tra có missing dữ liệu hay không

train\_df.isnull().sum()

Tính % dữ liệu missing so với mẫu

(train\_df['Age'].isnull().sum()/train\_df.shape[0])\*100)

Nếu % missing quá lớn thì bỏ cột dữ liệu đấy

train\_data.drop('Cabin', axis=1, inplace=**True**)

Nếu % missing nhỏ, với dữ liệu định lượng ta tính mean/median của dữ liệu và thay vào null

train\_df["Age"].mean(skipna=**True**)

train\_df["Age"].median(skipna=**True**)

train\_data["Age"].fillna(train\_df["Age"].median(skipna=**True**), inplace=**True**)

Với dữ liệu định tính ta đếm giá trị xuất hiện nhiều nhất và thay vào null

train\_data["Embarked"].fillna(train\_df['Embarked'].value\_counts().idxmax(), inplace=**True**)

1. Summary dữ liệu dùng các đại lượng thống kê mô tả

describe()

count 9.000000 số dòng dữ liệu trong mẫu

mean 11.622222 giá trị trung bình

std 15.124548 độ lệch chuẩn

min -5.000000 giá trị nhỏ nhất

25% 0.100000 phân vị dưới

50% 8.000000 trung vị

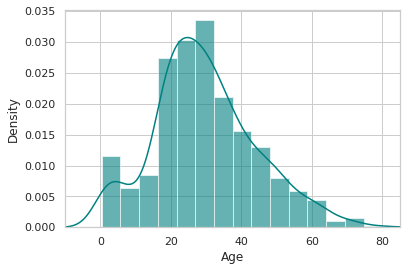
75% 21.000000 phân vị trên

max 41.000000 giá trị lớn nhất

dtype: float64 kiểu dữ liệu

1. Trực quan dữ liệu

Phân tích biến định lượng thường dùng histogram

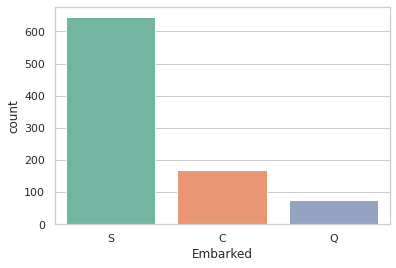
ax = train\_df["Age"].hist(bins=15, density=**True**, stacked=**True**, color='teal', alpha=0.6)

train\_df["Age"].plot(kind='density', color='teal')

ax.set(xlabel='Age')

plt.xlim(-10,85)

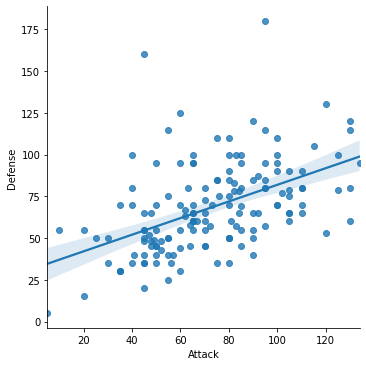
plt.show()



Phân tích biến phân loại thường dùng countplot và subplot

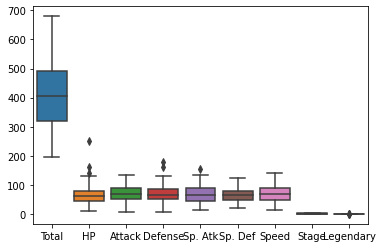
sns.countplot(x='Embarked', data=train\_df, palette='Set2')

plt.show()



Phân tích sự tương quan giữa 2 biến định lượng dùng scatterplot và jointplot

sns.lmplot(x='Attack', y='Defense', data=df)

Phân tích sự tương quan giữa biến định lượng và biến phân loại dùng boxplot

sns.boxplot(data=df)