In [2]:
```
import numpy as np
import pandas as pd
from sklearn import preprocessing
import matplotlib.pyplot as plt
import seaborn as sns
sns.set(style="white")
sns.set(style="whitegrid",color_codes=True)
import warnings
warnings.simplefilter(action='ignore')
```

In [6]:
```
df=pd.read_csv(r"C:\Users\pucha\Downloads\heart disease.csv")
df
```

Out[6]:

	male	age	education	currentSmoker	cigsPerDay	BPMeds	prevalentStroke	prevalentHyp	diabetes	totChol	sysBP	diaBP	BMI	heartRa
0	1	39	4.0	0	0.0	0.0	0	0	0	195.0	106.0	70.0	26.97	80
1	0	46	2.0	0	0.0	0.0	0	0	0	250.0	121.0	81.0	28.73	95
2	1	48	1.0	1	20.0	0.0	0	0	0	245.0	127.5	80.0	25.34	75
3	0	61	3.0	1	30.0	0.0	0	1	0	225.0	150.0	95.0	28.58	65
4	0	46	3.0	1	23.0	0.0	0	0	0	285.0	130.0	84.0	23.10	85
...	...	...	...	...	...	...	...	...	...	...	...	...	...	
4233	1	50	1.0	1	1.0	0.0	0	1	0	313.0	179.0	92.0	25.97	66
4234	1	51	3.0	1	43.0	0.0	0	0	0	207.0	126.5	80.0	19.71	65
4235	0	48	2.0	1	20.0	NaN	0	0	0	248.0	131.0	72.0	22.00	84
4236	0	44	1.0	1	15.0	0.0	0	0	0	210.0	126.5	87.0	19.16	86
4237	0	52	2.0	0	0.0	0.0	0	0	0	269.0	133.5	83.0	21.47	80

4238 rows × 16 columns

In [7]:
```
df.head(11)
```

Out[7]:

	male	age	education	currentSmoker	cigsPerDay	BPMeds	prevalentStroke	prevalentHyp	diabetes	totChol	sysBP	diaBP	BMI	heartRate
0	1	39	4.0	0	0.0	0.0	0	0	0	195.0	106.0	70.0	26.97	80.0
1	0	46	2.0	0	0.0	0.0	0	0	0	250.0	121.0	81.0	28.73	95.0
2	1	48	1.0	1	20.0	0.0	0	0	0	245.0	127.5	80.0	25.34	75.0
3	0	61	3.0	1	30.0	0.0	0	1	0	225.0	150.0	95.0	28.58	65.0
4	0	46	3.0	1	23.0	0.0	0	0	0	285.0	130.0	84.0	23.10	85.0
5	0	43	2.0	0	0.0	0.0	0	1	0	228.0	180.0	110.0	30.30	77.0
6	0	63	1.0	0	0.0	0.0	0	0	0	205.0	138.0	71.0	33.11	60.0
7	0	45	2.0	1	20.0	0.0	0	0	0	313.0	100.0	71.0	21.68	79.0
8	1	52	1.0	0	0.0	0.0	0	1	0	260.0	141.5	89.0	26.36	76.0
9	1	43	1.0	1	30.0	0.0	0	1	0	225.0	162.0	107.0	23.61	93.0
10	0	50	1.0	0	0.0	0.0	0	0	0	254.0	133.0	76.0	22.91	75.0

In [8]:
```
df.describe
```

Out[8]:
```
<bound method NDFrame.describe of       male  age  education  currentSmoker  cigsPerDay  BPMeds  \
0        1   39        4.0              0         0.0     0.0
1        0   46        2.0              0         0.0     0.0
2        1   48        1.0              1        20.0     0.0
3        0   61        3.0              1        30.0     0.0
4        0   46        3.0              1        23.0     0.0
...    ...  ...        ...            ...         ...     ...
4233     1   50        1.0              1         1.0     0.0
4234     1   51        3.0              1        43.0     0.0
4235     0   48        2.0              1        20.0     NaN
4236     0   44        1.0              1        15.0     0.0
4237     0   52        2.0              0         0.0     0.0

      prevalentStroke  prevalentHyp  diabetes  totChol  sysBP  diaBP    BMI
0                   0             0         0    195.0  106.0   70.0  26.97  \
1                   0             0         0    250.0  121.0   81.0  28.73
2                   0             0         0    245.0  127.5   80.0  25.34
3                   0             1         0    225.0  150.0   95.0  28.58
4                   0             0         0    285.0  130.0   84.0  23.10
...               ...           ...       ...      ...    ...    ...    ...
4233                0             1         0    313.0  179.0   92.0  25.97
4234                0             0         0    207.0  126.5   80.0  19.71
4235                0             0         0    248.0  131.0   72.0  22.00
4236                0             0         0    210.0  126.5   87.0  19.16
4237                0             0         0    269.0  133.5   83.0  21.47

      heartRate  glucose  TenYearCHD
0          80.0     77.0           0
1          95.0     76.0           0
2          75.0     70.0           0
3          65.0    103.0           1
4          85.0     85.0           0
...         ...      ...         ...
4233       66.0     86.0           1
4234       65.0     68.0           0
4235       84.0     86.0           0
4236       86.0      NaN           0
4237       80.0    107.0           0

[4238 rows x 16 columns]>
```

In [9]:
```
df.tail()
```

Out[9]:

	male	age	education	currentSmoker	cigsPerDay	BPMeds	prevalentStroke	prevalentHyp	diabetes	totChol	sysBP	diaBP	BMI	heartRa
4233	1	50	1.0	1	1.0	0.0	0	1	0	313.0	179.0	92.0	25.97	66
4234	1	51	3.0	1	43.0	0.0	0	0	0	207.0	126.5	80.0	19.71	65
4235	0	48	2.0	1	20.0	NaN	0	0	0	248.0	131.0	72.0	22.00	84
4236	0	44	1.0	1	15.0	0.0	0	0	0	210.0	126.5	87.0	19.16	86
4237	0	52	2.0	0	0.0	0.0	0	0	0	269.0	133.5	83.0	21.47	80

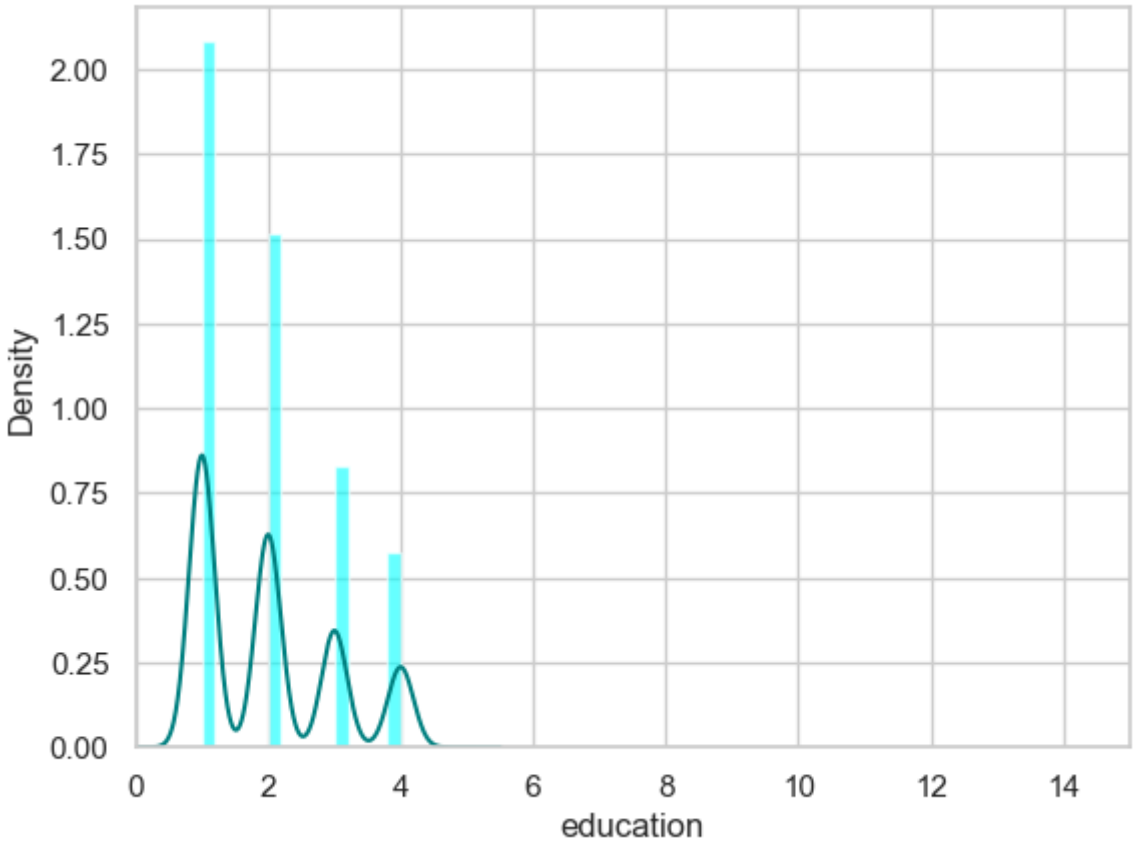
In [10]:
```
df.shape
```

Out[10]:
```
(4238, 16)
```

In [11]:
```
df.isnull().sum()
```

Out[11]:
```
male               0
age                0
education        105
currentSmoker      0
cigsPerDay        29
BPMeds            53
prevalentStroke    0
prevalentHyp       0
diabetes           0
totChol           50
sysBP              0
diaBP              0
BMI               19
heartRate          1
glucose          388
TenYearCHD         0
dtype: int64
```

In [12]:
```
ax=df["education"].hist(bins=15,density=True,stacked=True,color='cyan',alpha=0.6)
df["education"].plot(kind='density',color='teal')
ax.set(xlabel='education')
plt.xlim(-0,15)
plt.show()
```



In [13]:
```
print(df["education"].mean(skipna=True))
print(df["education"].median(skipna=True))
```

```
1.9789499153157513
2.0
```

In [14]:
```
print((df['glucose'].isnull().sum()/df.shape[0]*100))
```

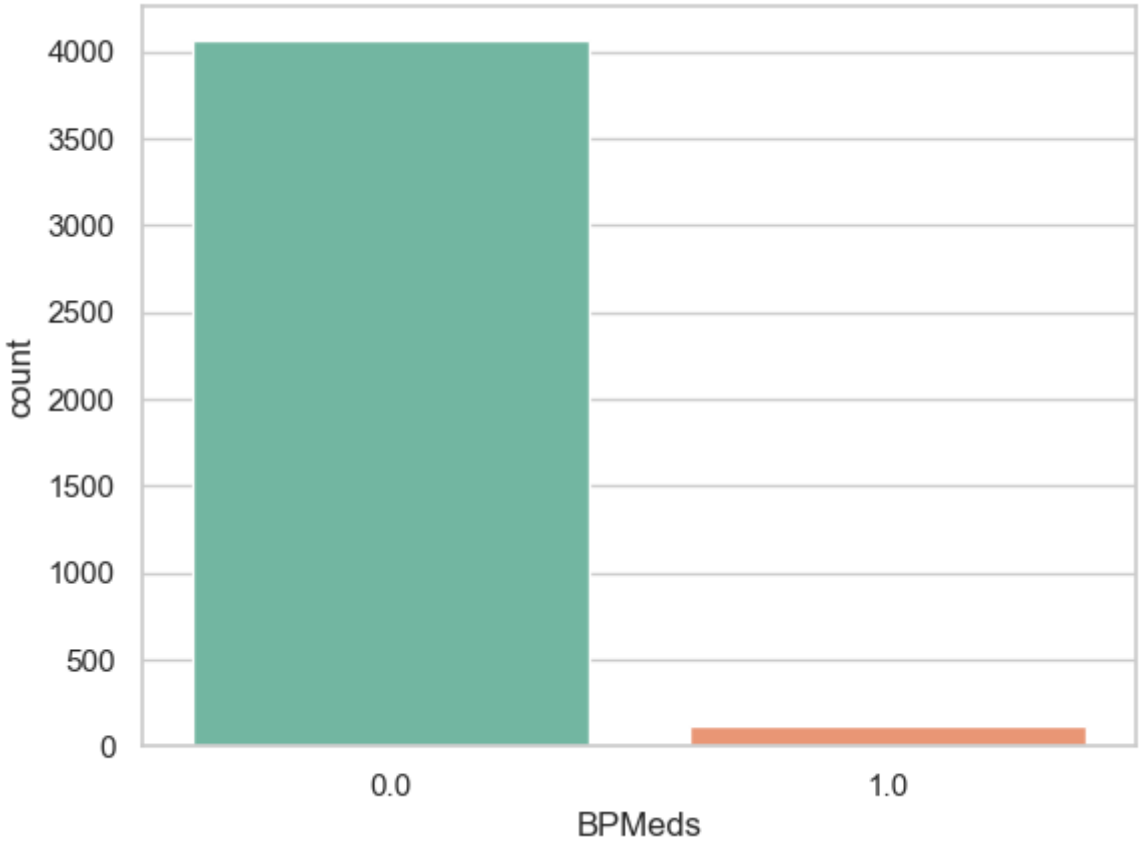
```
9.155261915998112
```

In [15]:
```
print((df['totChol'].isnull().sum()/df.shape[0]*100))
```

```
1.1798017932987257
```

In [16]:
```
data=df.copy
print(df['BPMeds'].value_counts())
sns.countplot(x='BPMeds',data=df,palette='Set2')
plt.show()
```

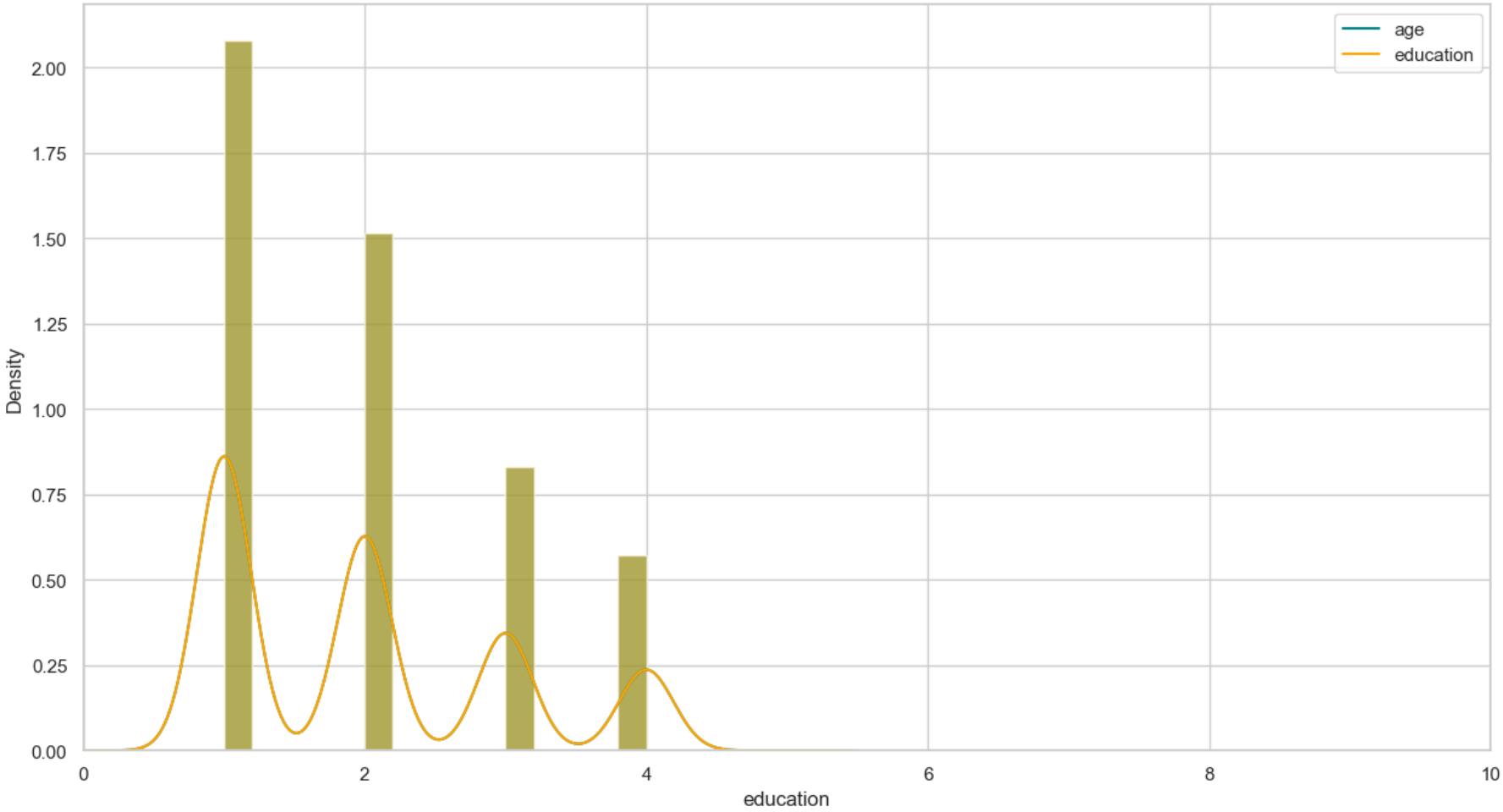
```
BPMeds
0.0    4061
1.0     124
Name: count, dtype: int64
```



In [17]:
```
print(df['heartRate'].value_counts().idxmax())
```

```
75.0
```

In [18]:
```
plt.figure(figsize=(15,8))
ax=df["education"].hist(bins=15,density=True,stacked=True,color='teal',alpha=0.6)
df["education"].plot(kind='density',color='teal')
ax=df["education"].hist(bins=15,density=True,stacked=True,color='orange',alpha=0.5)
df["education"].plot(kind='density',color='orange')
ax.legend(["age","education"])
ax.set(xlabel='education')
plt.xlim(-0,10)
plt.show()
```



In [19]:
```
df['Disease']=np.where((df["prevalentHyp"]+df["prevalentStroke"])>0,0,1)
df.drop('prevalentHyp',axis=1,inplace=True)
df.drop('prevalentStroke',axis=1,inplace=True)
```

In [20]:
```
training=pd.get_dummies(df,columns=["currentSmoker","totChol","sysBP"])
training.drop("TenYearCHD",axis=1,inplace=True)
training.drop("male",axis=1,inplace=True)
training.drop("diaBP",axis=1,inplace=True)

final_train=training
final_train.head()
```

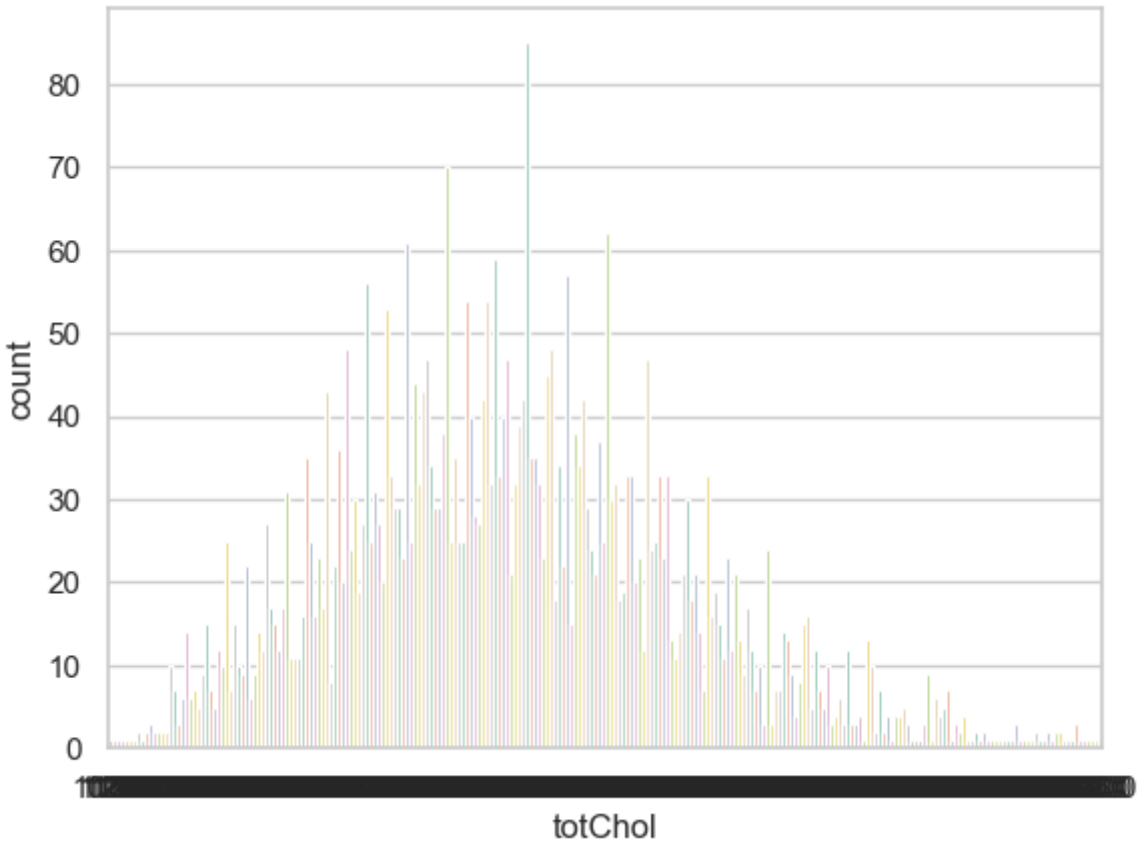
Out[20]:

	age	education	cigsPerDay	BPMeds	diabetes	BMI	heartRate	glucose	Disease	currentSmoker_0	...	sysBP_215.0	sysBP_217.0	sysBP_2
0	39	4.0	0.0	0.0	0	26.97	80.0	77.0	1	True	...	False	False	F
1	46	2.0	0.0	0.0	0	28.73	95.0	76.0	1	True	...	False	False	F
2	48	1.0	20.0	0.0	0	25.34	75.0	70.0	1	False	...	False	False	F
3	61	3.0	30.0	0.0	0	28.58	65.0	103.0	0	False	...	False	False	F
4	46	3.0	23.0	0.0	0	23.10	85.0	85.0	1	False	...	False	False	F

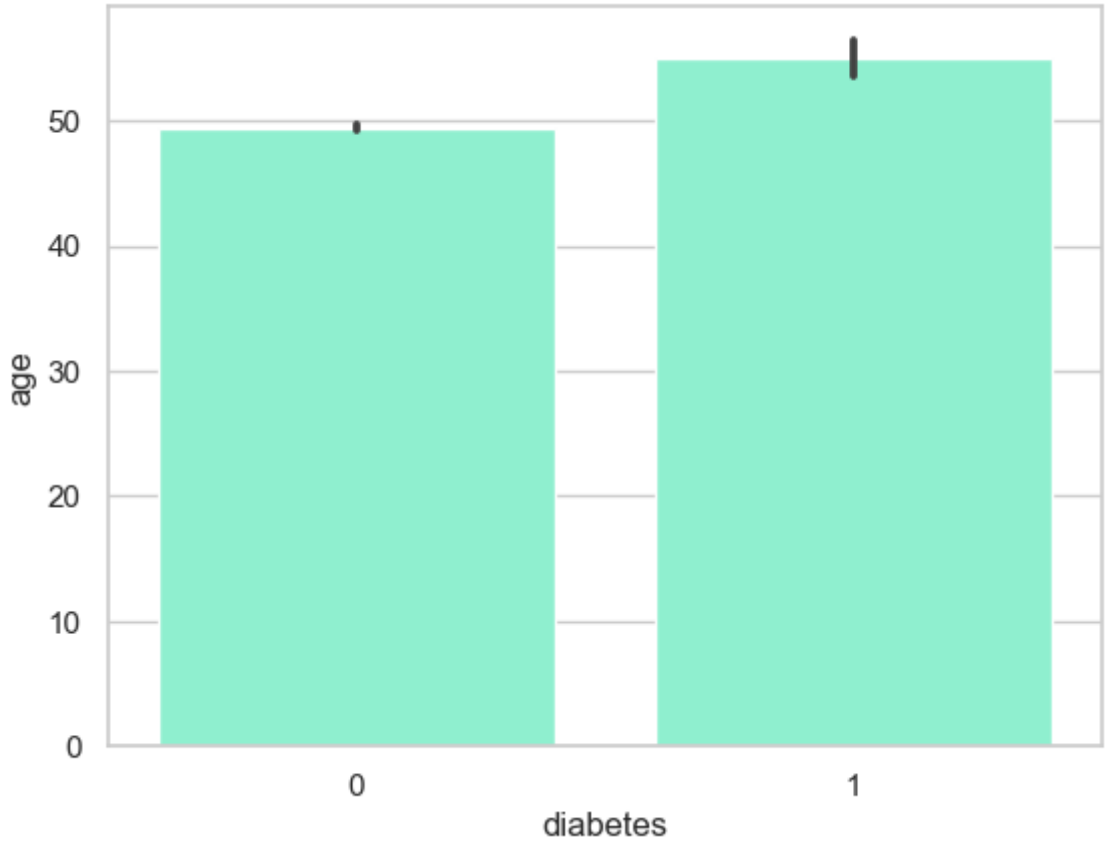
5 rows × 493 columns

In [21]:
```
print(df['totChol'].value_counts())
sns.countplot(x= 'totChol',data=df,palette='Set2')
plt.show()
```

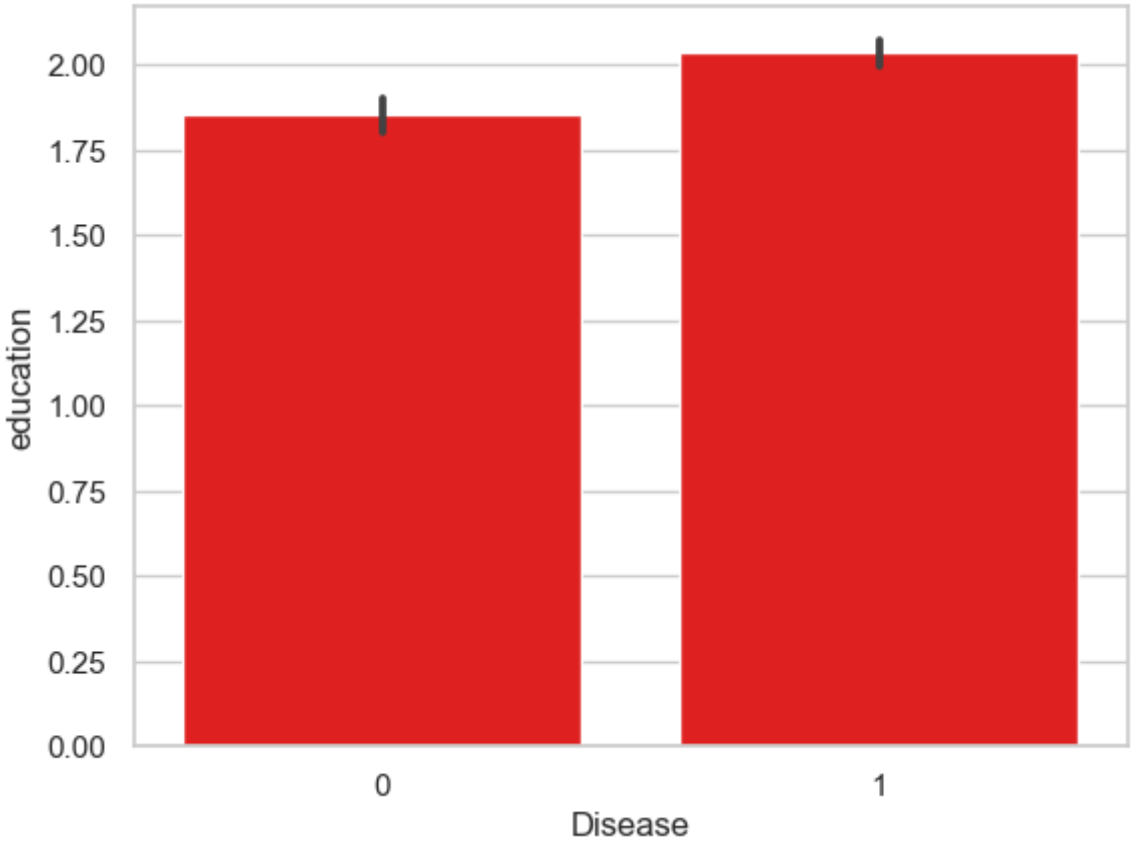
```
totChol
240.0    85
220.0    70
260.0    62
210.0    61
232.0    59
         ..
392.0     1
405.0     1
359.0     1
398.0     1
119.0     1
Name: count, Length: 248, dtype: int64
```


80
70
60
50
40
30
20
10
0
count
totChol

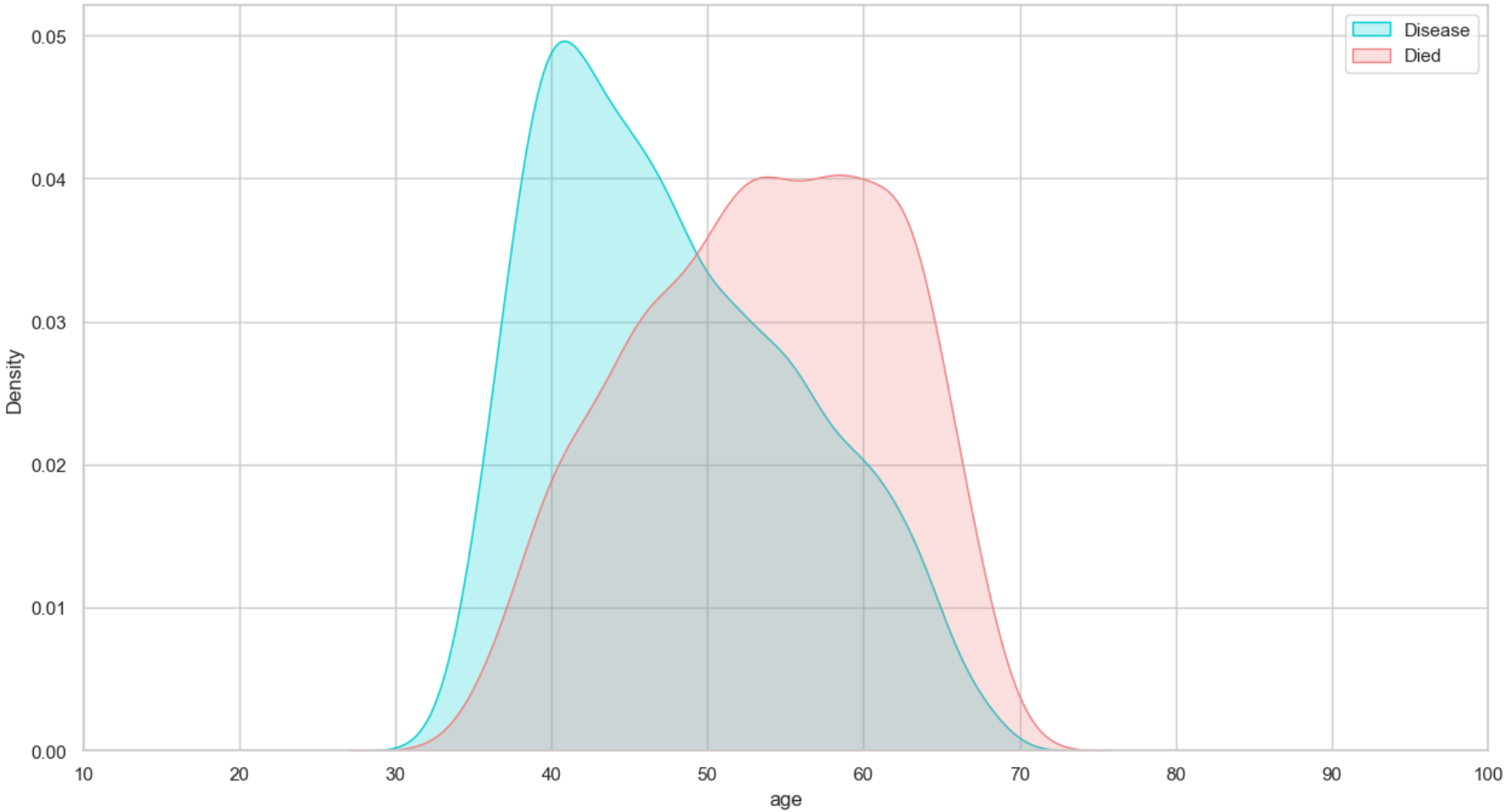
In [22]:
```
sns.barplot(x='diabetes',y='age',data=df,color="aquamarine")
plt.show()
```



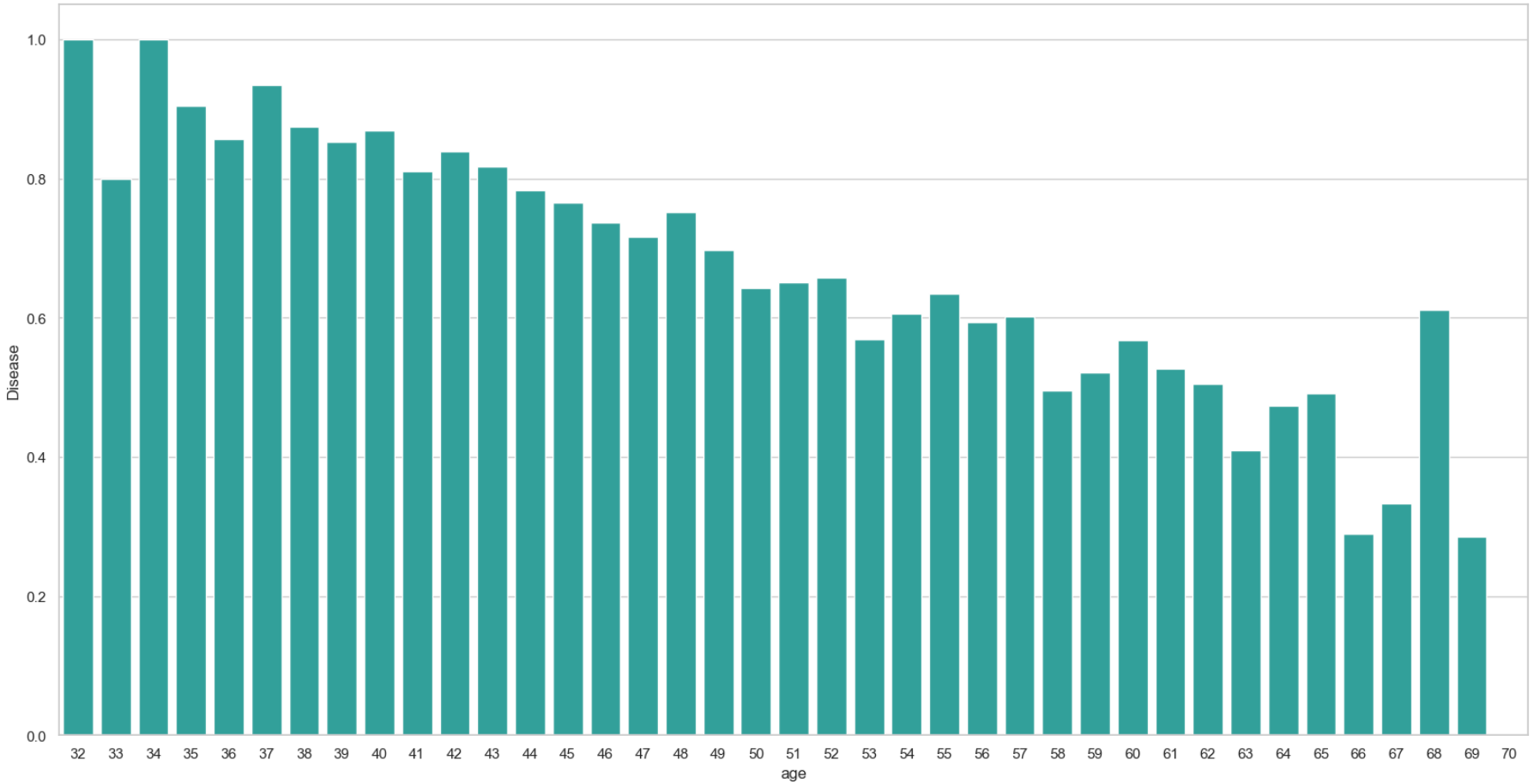
In [23]:
```
sns.barplot(x='Disease',y='education',data=df,color="red")
plt.show()
```



In [24]:
```
plt.figure(figsize=(15,8))
ax = sns.kdeplot(final_train["age"][final_train.Disease == 1],color="darkturquoise",shade=True)
sns.kdeplot(final_train["age"][final_train.Disease == 0],color="lightcoral",shade=True)
plt.legend(['Disease','Died'])
ax.set(xlabel='age')
plt.xlim(10,100)
plt.show()
```



In [25]:
```
plt.figure(figsize=(20,10))
avg_survival_byage=final_train[["age","Disease"]].groupby(['age'],as_index=False).mean()
g=sns.barplot(x='age',y='Disease',data=avg_survival_byage,color="LightSeaGreen")
plt.show()
```



In [ ]:
```

```