In [1]:

```
import pandas as pd
import numpy as np
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 [3]:

```
df=pd.read_csv(r"C:\Users\Svijayalakshmi\Downloads\framingham.csv")
df
```

Out[3]:

	male	age	education	currentSmoker	cigsPerDay	BPMeds	prevalentStroke	prevalentHyp	diabetes
0	1	39	4.0	0	0.0	0.0	0	0	0
1	0	46	2.0	0	0.0	0.0	0	0	0
2	1	48	1.0	1	20.0	0.0	0	0	0
3	0	61	3.0	1	30.0	0.0	0	1	0
4	0	46	3.0	1	23.0	0.0	0	0	0
...	...	...	...	...	...	...	...	...	...
4233	1	50	1.0	1	1.0	0.0	0	1	0
4234	1	51	3.0	1	43.0	0.0	0	0	0
4235	0	48	2.0	1	20.0	NaN	0	0	0
4236	0	44	1.0	1	15.0	0.0	0	0	0
4237	0	52	2.0	0	0.0	0.0	0	0	0

4238 rows × 16 columns

In [4]:

```
df.head()
```

Out[4]:

	male	age	education	currentSmoker	cigsPerDay	BPMeds	prevalentStroke	prevalentHyp	diabetes	totC
0	1	39	4.0	0	0.0	0.0	0	0	0	1
1	0	46	2.0	0	0.0	0.0	0	0	0	2
2	1	48	1.0	1	20.0	0.0	0	0	0	2
3	0	61	3.0	1	30.0	0.0	0	1	0	2
4	0	46	3.0	1	23.0	0.0	0	0	0	2

In [5]:

```
df.tail()
```

Out[5]:

	male	age	education	currentSmoker	cigsPerDay	BPMeds	prevalentStroke	prevalentHyp	diabetes
4233	1	50	1.0	1	1.0	0.0	0	1	0
4234	1	51	3.0	1	43.0	0.0	0	0	0
4235	0	48	2.0	1	20.0	NaN	0	0	0
4236	0	44	1.0	1	15.0	0.0	0	0	0
4237	0	52	2.0	0	0.0	0.0	0	0	0

In [6]:

```
df.shape
```

Out[6]:

```
(4238, 16)
```

In [7]:

```
df.describe()
```

Out[7]:

	male	age	education	currentSmoker	cigsPerDay	BPMeds	prevalentStroke	pr
count	4238.000000	4238.000000	4133.000000	4238.000000	4209.000000	4185.000000	4238.000000	4
mean	0.429212	49.584946	1.978950	0.494101	9.003089	0.029630	0.005899	
std	0.495022	8.572160	1.019791	0.500024	11.920094	0.169584	0.076587	
min	0.000000	32.000000	1.000000	0.000000	0.000000	0.000000	0.000000	
25%	0.000000	42.000000	1.000000	0.000000	0.000000	0.000000	0.000000	
50%	0.000000	49.000000	2.000000	0.000000	0.000000	0.000000	0.000000	
75%	1.000000	56.000000	3.000000	1.000000	20.000000	0.000000	0.000000	
max	1.000000	70.000000	4.000000	1.000000	70.000000	1.000000	1.000000	

In [8]:

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 4238 entries, 0 to 4237
Data columns (total 16 columns):
 #   Column           Non-Null Count  Dtype  
---  ------           --------------  -----  
 0   male             4238 non-null   int64  
 1   age              4238 non-null   int64  
 2   education        4133 non-null   float64
 3   currentSmoker    4238 non-null   int64  
 4   cigsPerDay       4209 non-null   float64
 5   BPMeds           4185 non-null   float64
 6   prevalentStroke  4238 non-null   int64  
 7   prevalentHyp     4238 non-null   int64  
 8   diabetes         4238 non-null   int64  
 9   totChol          4188 non-null   float64
 10  sysBP            4238 non-null   float64
 11  diaBP            4238 non-null   float64
 12  BMI              4219 non-null   float64
 13  heartRate        4237 non-null   float64
 14  glucose          3850 non-null   float64
 15  TenYearCHD       4238 non-null   int64  
dtypes: float64(9), int64(7)
memory usage: 529.9 KB
```

In [9]:

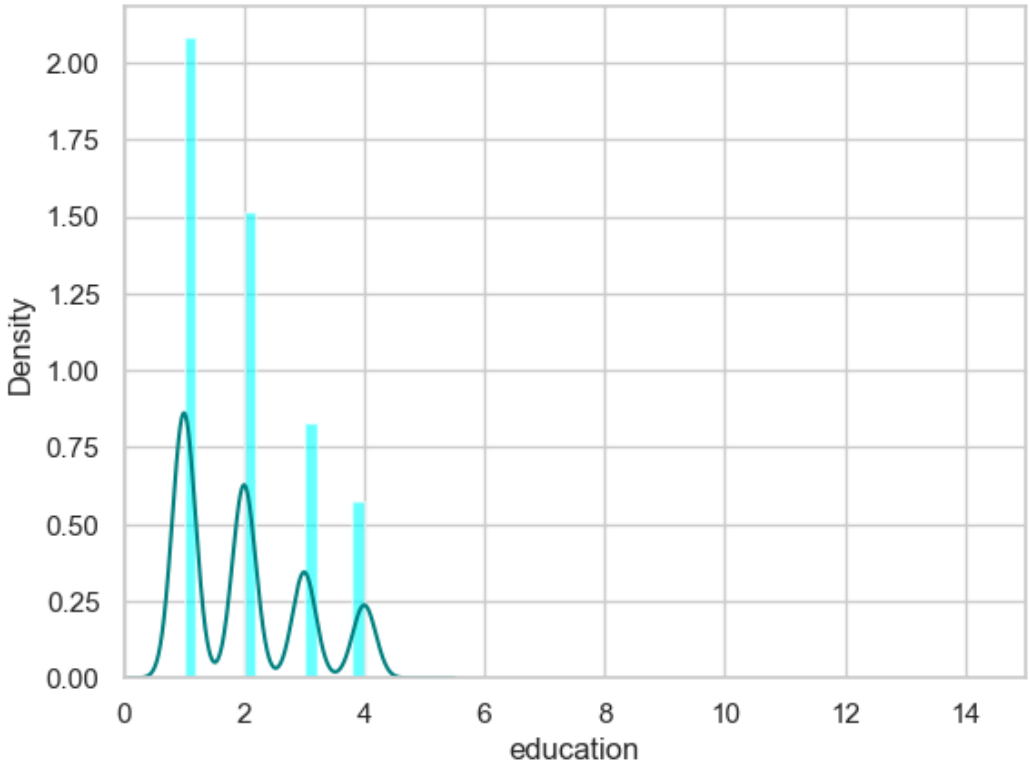
```
df.isnull().sum()
```

Out[9]:

```
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 [10]:

```
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 [11]:

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

```
1.9789499153157513
2.0
```

In [12]:

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

```
9.155261915998112
```

In [13]:

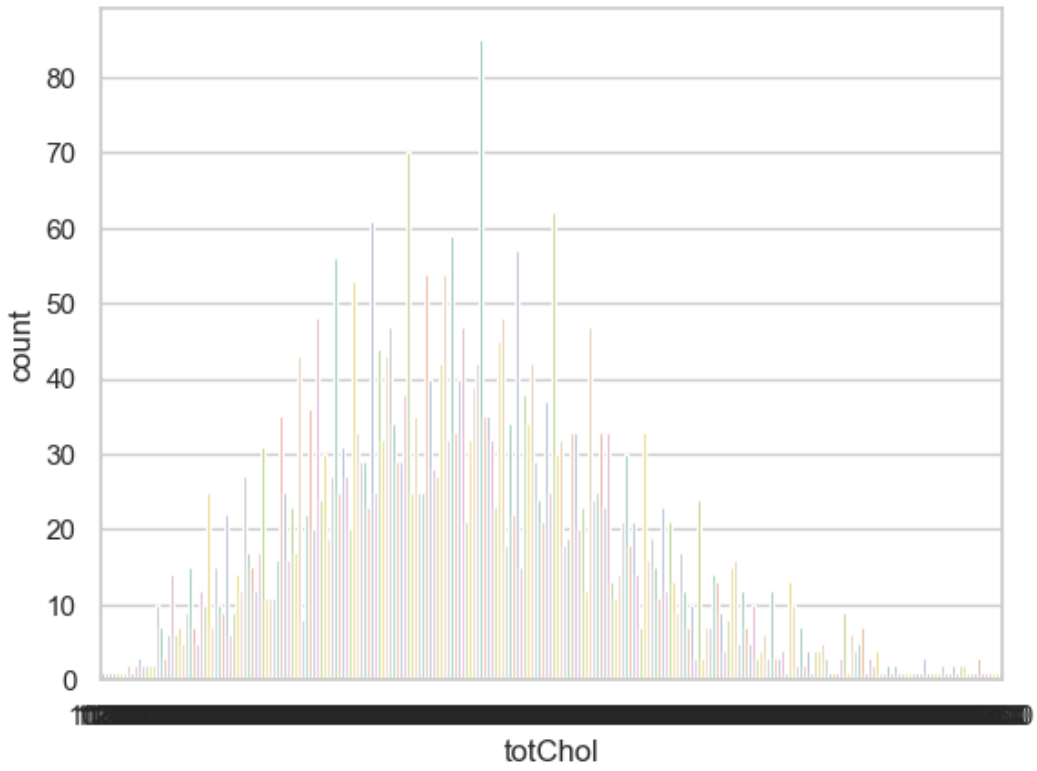
```
print(df['totChol'].isnull().sum()/df.shape[0]*100)
```

```
1.1798017932987257
```

In [14]:

```
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
```



In [15]:

```
print(df['totChol'].value_counts().idxmax())
```

```
240.0
```

In [16]:

```
data=df.copy()
data["education"].fillna(df["education"].median(skipna=True),inplace=True)
data["totChol"].fillna(df["totChol"].value_counts().idxmax(),inplace=True)
data.drop('glucose',axis=1,inplace=True)
```

In [17]:

```
data.isnull().sum()
```

Out[17]:

```
male               0
age                0
education          0
currentSmoker      0
cigsPerDay        29
BPMeds            53
prevalentStroke    0
prevalentHyp       0
diabetes           0
totChol            0
sysBP              0
diaBP              0
BMI               19
heartRate          1
TenYearCHD         0
dtype: int64
```

In [18]:

```
pd.set_option('display.max_rows',4238)
pd.set_option('display.max_columns',16)
```

In [19]:

```
pd.set_option('display.width',50)
```

In [20]:

```
print('This DataFrame has %d Rows and %d Columns'%(df.shape))
```

```
This DataFrame has 4238 Rows and 16 Columns
```

In [21]:

```
features_matrix=df.iloc[:,0:15]
```

In [22]:

```
target_vector=df.iloc[:,-2]
```

In [23]:

```
print('The Features Matrix Has %d Rows And %d Column(s)'%(features_matrix.shape))
```

```
The Features Matrix Has 4238 Rows And 15 Column(s)
```

In [24]:

```
print('The Target Matrix Has %d Rows And %d Column(s)'%(np.array(target_vector).reshape(-1,1).shape))
```

```
The Target Matrix Has 4238 Rows And 1 Column(s)
```

In [25]:

```
df['education'].mean()
```

Out[25]:

```
1.9789499153157513
```

In [26]:

```
df['cigsPerDay'].mean()
```

Out[26]:

```
9.003088619624615
```

In [27]:

```
df['heartRate'].median()
```

Out[27]:

```
75.0
```

In [28]:

```
df['BPMeds'].mean()
```

Out[28]:

```
0.02962962962962963
```

In [29]:

```
df["glucose"].fillna(df["glucose"].median(skipna=True),inplace=True)
df
```

53	0	62	1.0	0	0.0	0.0	0	0	0	240.0	145
54	0	39	2.0	1	20.0	0.0	0	0	0	209.0	115
55	0	46	1.0	1	10.0	0.0	0	0	0	250.0	116
56	0	54	1.0	1	9.0	0.0	0	0	1	266.0	114
57	1	49	1.0	1	2.0	0.0	0	1	0	255.0	143
58	1	44	2.0	0	0.0	0.0	0	0	0	185.0	115
59	0	40	4.0	1	20.0	0.0	0	0	0	205.0	158
60	1	56	4.0	1	20.0	0.0	0	0	0	270.0	121
61	0	67	1.0	0	0.0	0.0	0	1	0	254.0	157
62	1	53	1.0	1	20.0	0.0	0	0	0	220.0	123
63	0	57	1.0	1	3.0	0.0	0	0	0	235.0	126
64	1	57	1.0	0	0.0	0.0	0	0	0	220.0	136

In [31]:

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

Out[31]:

```
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             0
TenYearCHD          0
dtype: int64
```

In [32]:

```
df['education'].fillna(df['education'].median(skipna=True),inplace=True)
```

In [33]:

```
df['totChol'].fillna(df['totChol'].median(skipna=True),inplace=True)
```

In [34]:

```
df['BMI'].fillna(df['BMI'].median(skipna=True),inplace=True)
```

In [35]:

```
df['heartRate'].fillna(df['heartRate'].median(skipna=True),inplace=True)
```

In [36]:

```
df['BPMeds'].fillna(df['BPMeds'].median(skipna=True),inplace=True)
```

In [37]:

```
df['cigsPerDay'].fillna(df['cigsPerDay'].median(skipna=True),inplace=True)
```

In [38]:

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

Out[38]:

```
male               0
age                0
education          0
currentSmoker      0
cigsPerDay         0
BPMeds             0
prevalentStroke    0
prevalentHyp       0
diabetes           0
totChol            0
sysBP              0
diaBP              0
BMI                0
heartRate          0
glucose            0
TenYearCHD         0
dtype: int64
```

In [39]:

```
df.drop('glucose',axis=1,inplace=True)
```

In [40]:

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

Out[40]:

```
male               0
age                0
education          0
currentSmoker      0
cigsPerDay         0
BPMeds             0
prevalentStroke    0
prevalentHyp       0
diabetes           0
totChol            0
sysBP              0
diaBP              0
BMI                0
heartRate          0
TenYearCHD         0
dtype: int64
```

In [41]:

```
print(df["cigsPerDay"].mean(skipna=True))
print(df["cigsPerDay"].median(skipna=True))
```

```
8.941481831052384
0.0
```

In [42]:

```
print((df['BPMeds'].isnull().sum()/df.shape[0]*100))
```

```
0.0
```

In [43]:

```
print((df['BMI'].isnull().sum()/df.shape[0]*100))
```

```
0.0
```

In [44]:

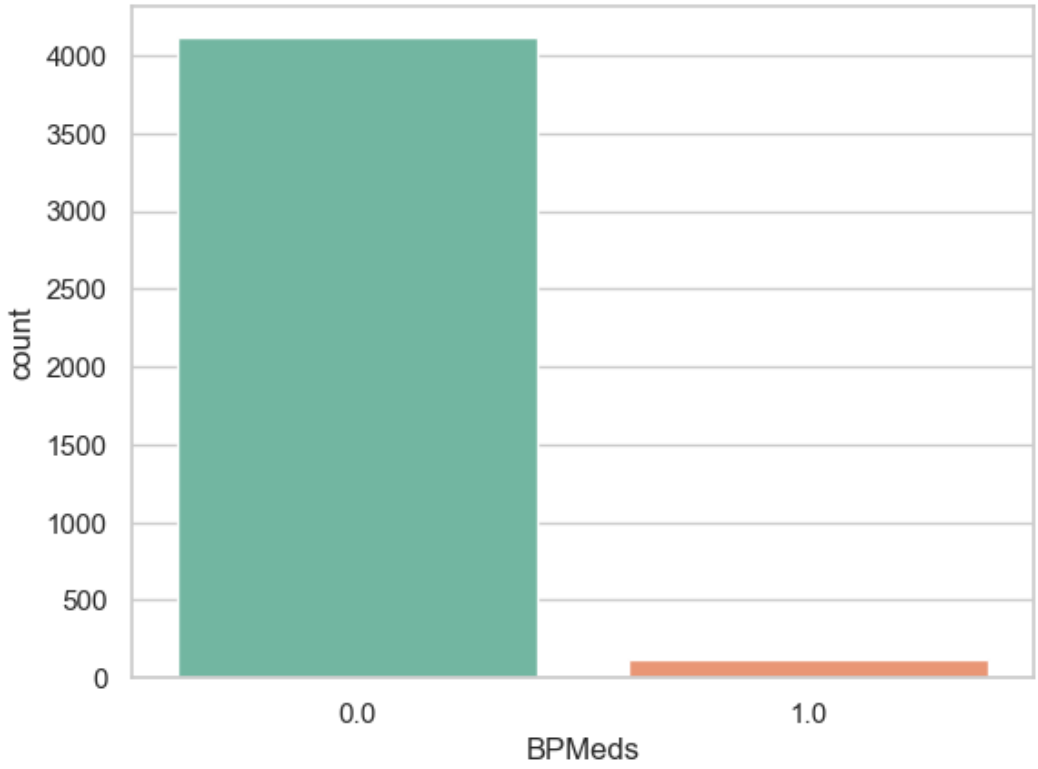
```
print((df['heartRate'].isnull().sum()/df.shape[0]*100))
```

```
0.0
```

In [45]:

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

```
BPMeds
0.0    4114
1.0     124
Name: count, dtype: int64
```



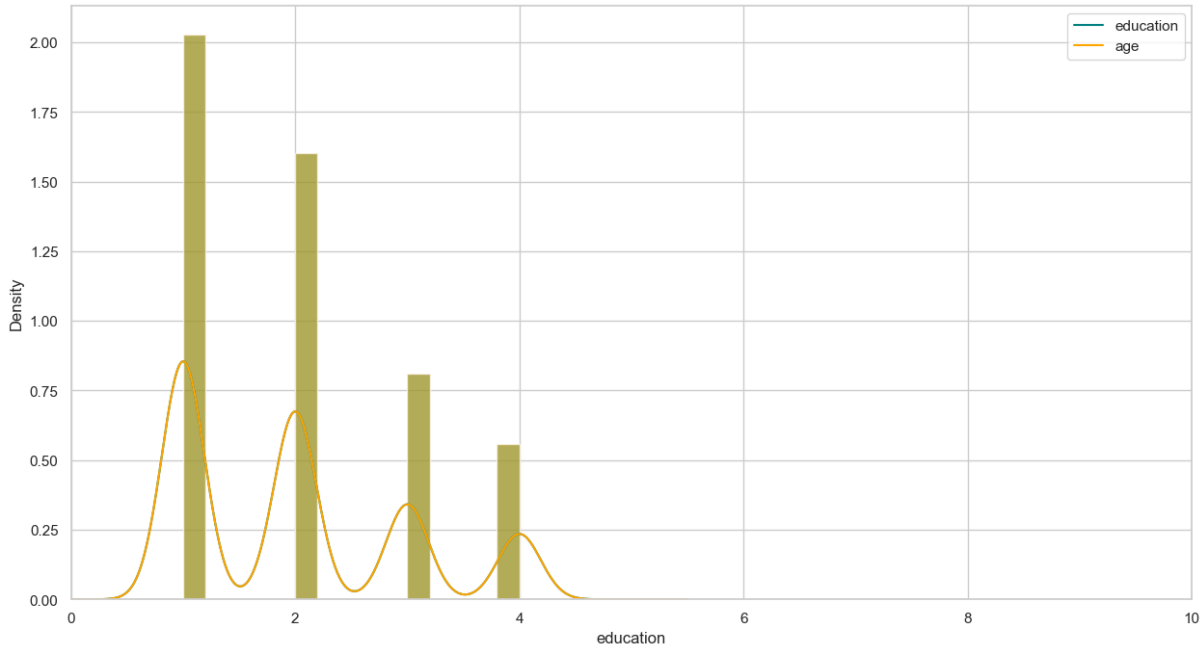
In [46]:

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

```
75.0
```

In [47]:

```
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=data["education"].hist(bins=15,density=True,stacked=True,color='orange',alpha=0.5)
data["education"].plot(kind='density',color='orange')
ax.legend(["education","age"])
ax.set(xlabel='education')
plt.xlim(-0,10)
plt.show()
```



In [48]:

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

In [49]:

```
training=pd.get_dummies(data,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()
```

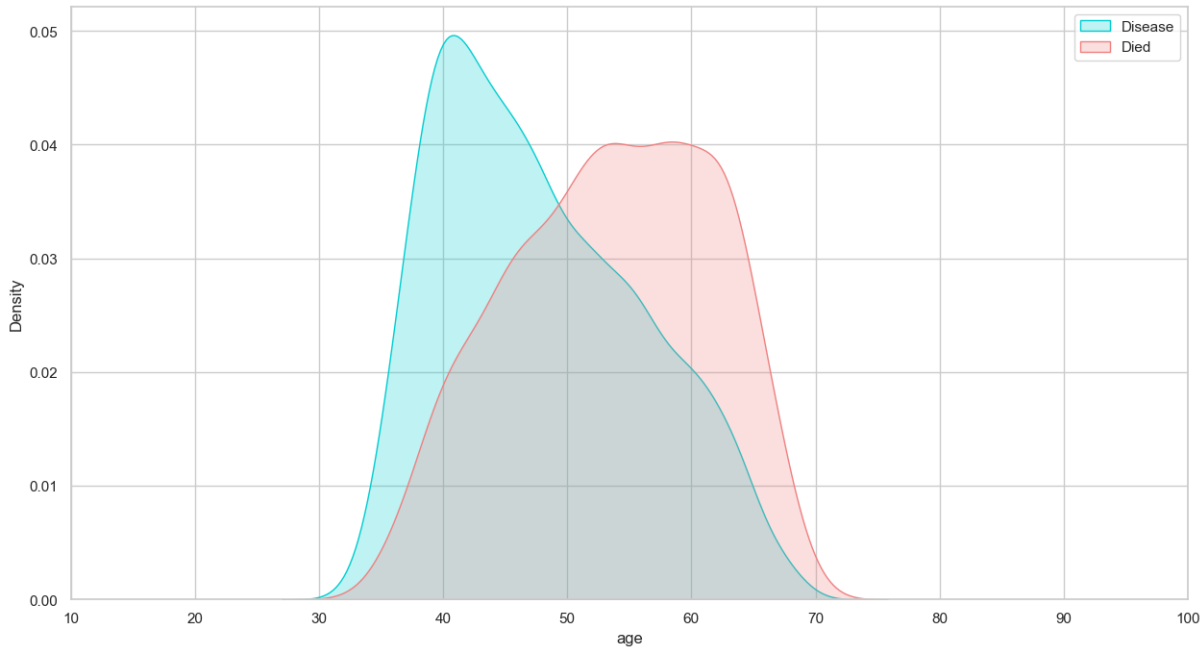
Out[49]:

	age	education	cigsPerDay	BPMeds	diabetes	BMI	heartRate	Disease	...	sysBP_220.0	sysBP_230.0
0	39	4.0	0.0	0.0	0	26.97	80.0	1	...	False	False
1	46	2.0	0.0	0.0	0	28.73	95.0	1	...	False	False
2	48	1.0	20.0	0.0	0	25.34	75.0	1	...	False	False
3	61	3.0	30.0	0.0	0	28.58	65.0	0	...	False	False
4	46	3.0	23.0	0.0	0	23.10	85.0	1	...	False	False

5 rows × 492 columns

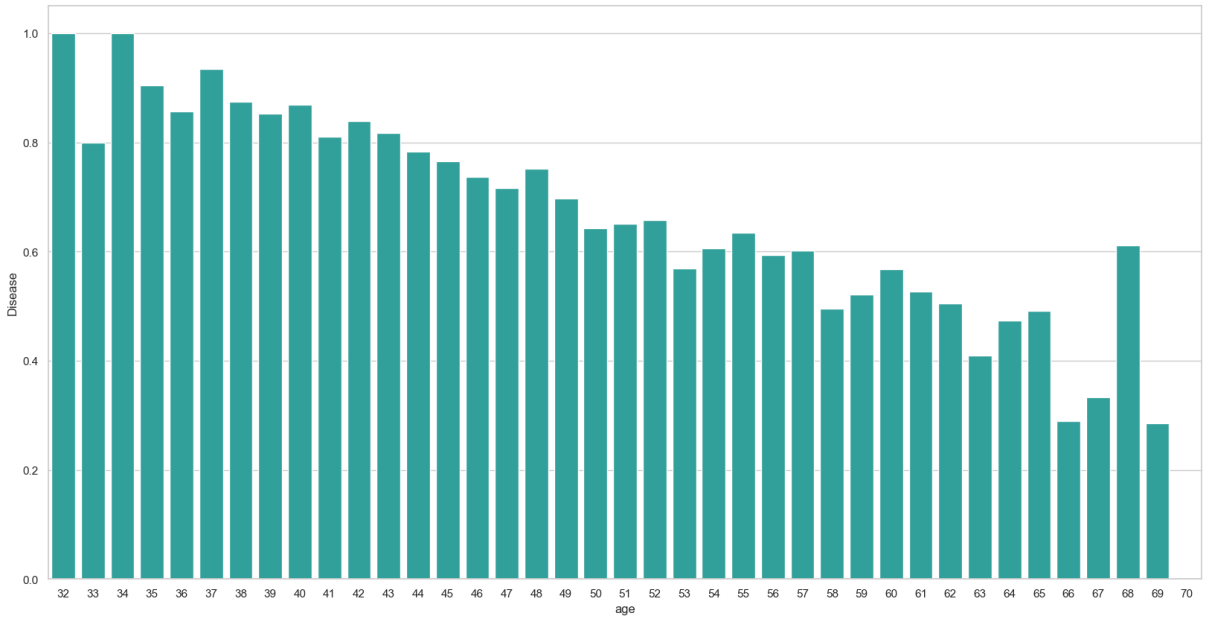
In [50]:

```
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 [51]:

```
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()
```



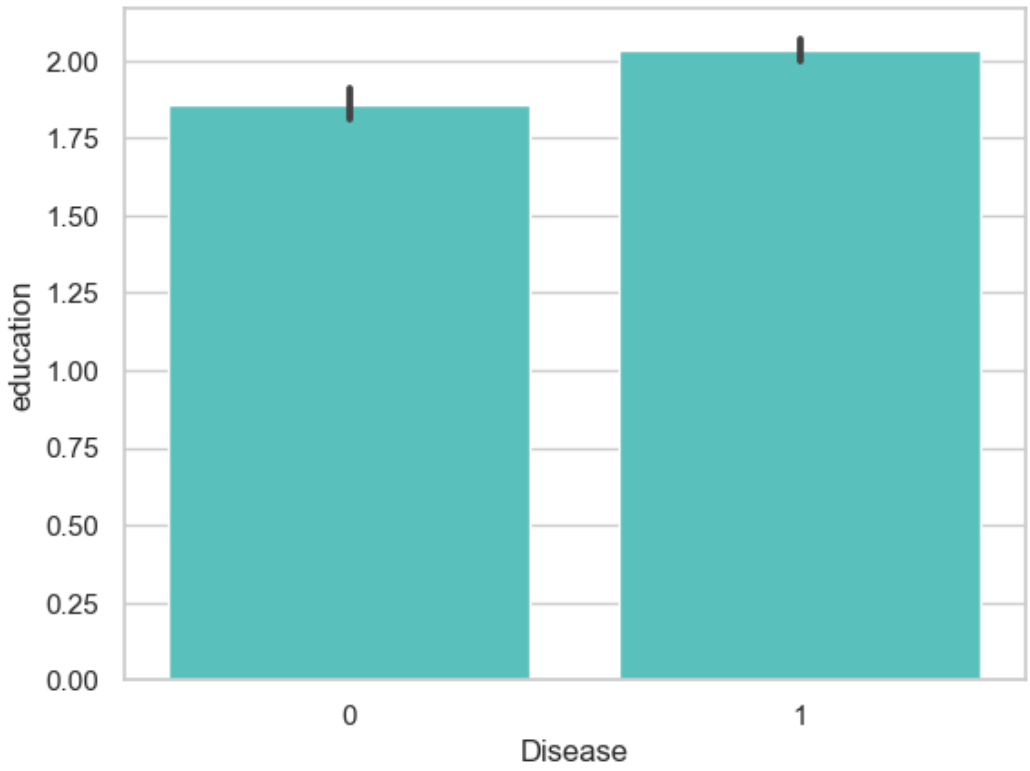
In [52]:

```
final_train['IsMinor']=np.where(final_train['age']<=16,1,0)
print(final_train['IsMinor'])
```

```
12      0
13      0
14      0
15      0
16      0
17      0
18      0
19      0
20      0
21      0
22      0
23      0
24      0
25      0
26      0
27      0
28      0
29      0
30      0
31      0
```

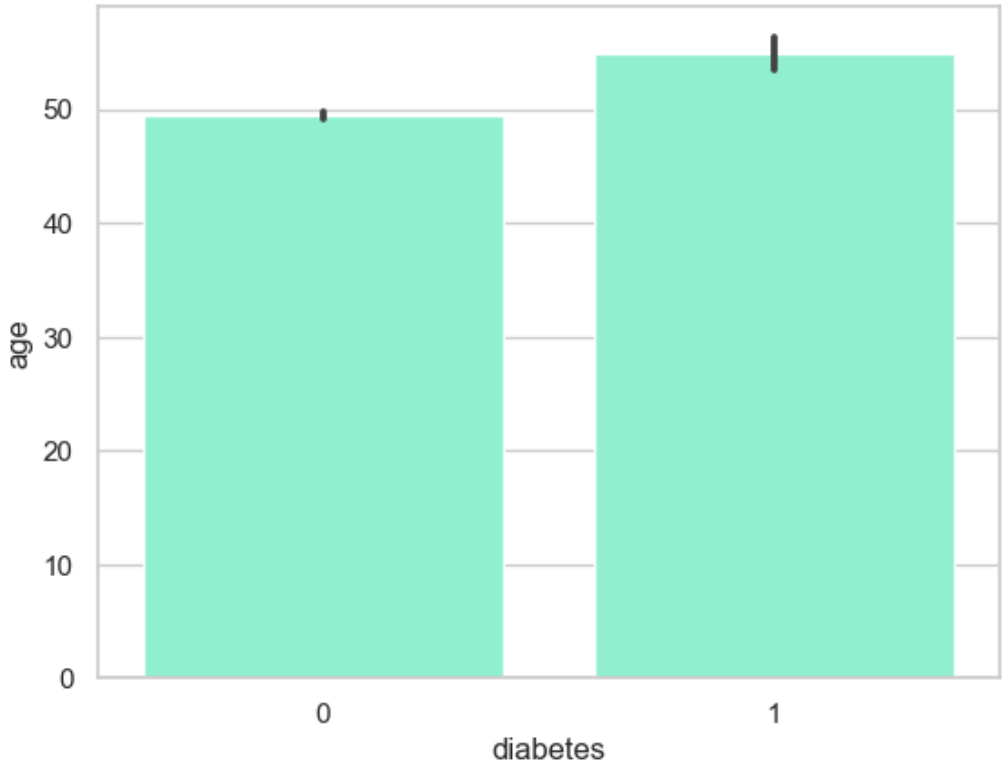
In [53]:

```
sns.barplot(x='Disease',y='education',data=final_train,color="mediumturquoise")
plt.show()
```



In [54]:

```
import seaborn as sns
import matplotlib.pyplot as plt
sns.barplot(x='diabetes',y='age',data=df,color="aquamarine")
plt.show()
```



In [ ]:

```

```