problem statement:

```
 To predict and analyze which age people are more effecting to heartdiseases an
d what are reasons to getting heartdisease
```

In [2]:

```
#import libraries
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='darkgrid',color_codes=True)
import warnings
warnings.simplefilter(action='ignore')
```

In [3]:

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

Out[3]:

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

4238 rows × 16 columns

In [4]:

```
df.head()
```

Out[4]:

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

In [5]:

```
df.shape
```

Out[5]:

```
(4238, 16)
```

In [6]:

```
df.describe()
```

Out[6]:

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

In [7]:

```
df.info
```

Out[7]:

```
<bound method DataFrame.info 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    BM
I  
0                   0             0         0    195.0  106.0   70.0  26.9
7  \
1                   0             0         0    250.0  121.0   81.0  28.7
3  
2                   0             0         0    245.0  127.5   80.0  25.3
4  
3                   0             1         0    225.0  150.0   95.0  28.5
8  
4                   0             0         0    285.0  130.0   84.0  23.1
0  
...               ...           ...       ...      ...    ...    ...    
...  
4233                0             1         0    313.0  179.0   92.0  25.9
7  
4234                0             0         0    207.0  126.5   80.0  19.7
1  
4235                0             0         0    248.0  131.0   72.0  22.0
0  
4236                0             0         0    210.0  126.5   87.0  19.1
6  
4237                0             0         0    269.0  133.5   83.0  21.4
7  

      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 [8]:

```
df.size
```

Out[8]:

```
67808
```

In [9]:

```
df.isna().any()
```

Out[9]:

```
male               False
age                False
education           True
currentSmoker      False
cigsPerDay          True
BPMeds              True
prevalentStroke    False
prevalentHyp       False
diabetes           False
totChol             True
sysBP              False
diaBP              False
BMI                 True
heartRate           True
glucose             True
TenYearCHD         False
dtype: bool
```

In [10]:

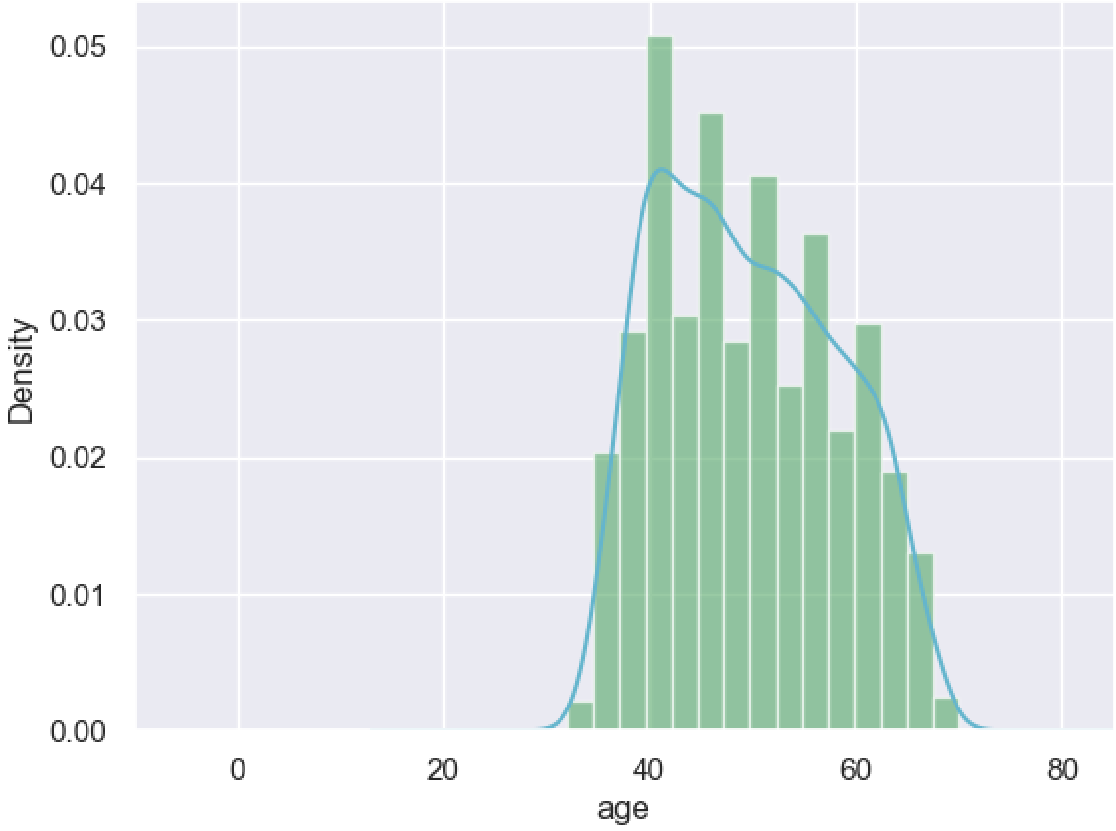
```
df.isnull().sum()
```

Out[10]:

```
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 [11]:

```
ax=df["age"].hist(bins=15,density=True,stacked=True,color='g',alpha=0.6)
df["age"].plot(kind='density',color='c')
ax.set(xlabel='age')
plt.xlim(-10,85)
plt.show()
```



In [12]:

```
from sklearn.model_selection import train_test_split

print(df['age'].mean(skipna=True))
print(df['age'].median(skipna=True))
```

```
49.58494572911751
49.0
```

In [13]:

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

```
9.155261915998112
```

In [14]:

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

```
2.4775837659273243
```

In [15]:

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

```
0.684285040113261
```

In [46]:

```
df
```

Out[46]:

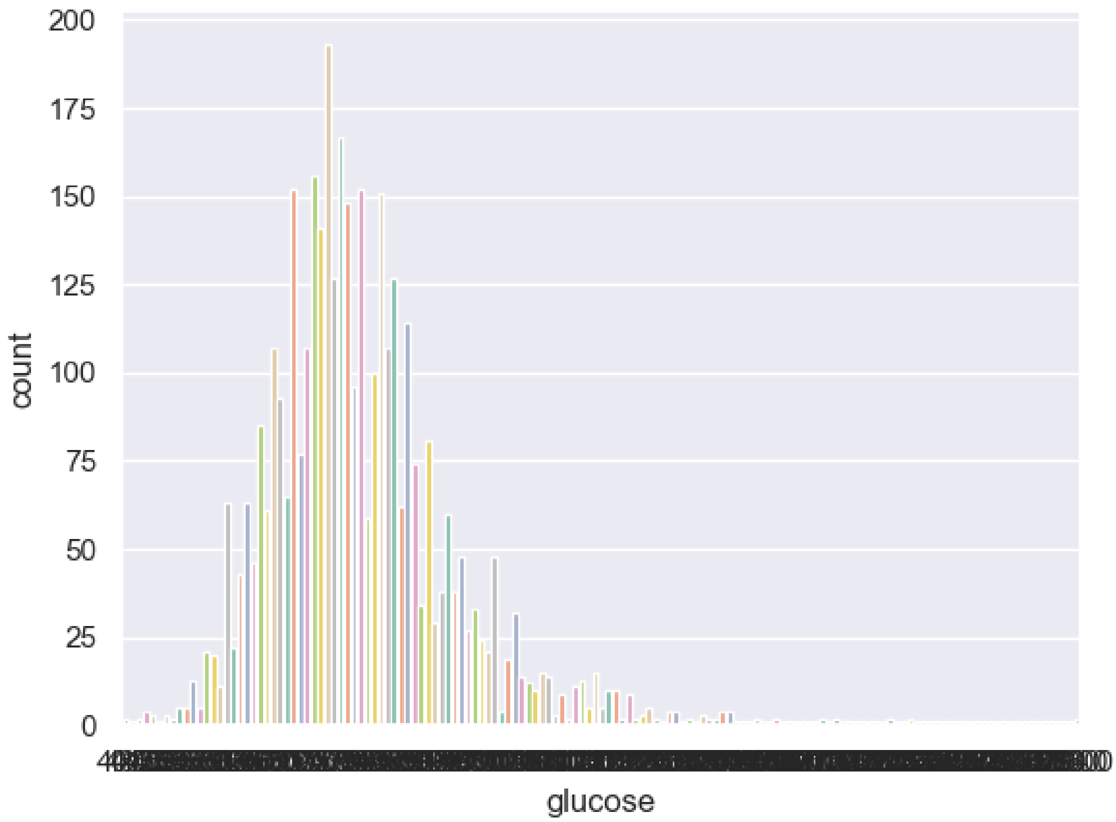
	male	age	education	currentSmoker	prevalentStroke	prevalentHyp	diabetes	sysBP
0	1	39	4.0	0	0	0	0	106.0
1	0	46	2.0	0	0	0	0	121.0
2	1	48	1.0	1	0	0	0	127.5
3	0	61	3.0	1	0	1	0	150.0
4	0	46	3.0	1	0	0	0	130.0
...	...	...	...	...	...	...	...	...
4233	1	50	1.0	1	0	1	0	179.0
4234	1	51	3.0	1	0	0	0	126.5
4235	0	48	2.0	1	0	0	0	131.0
4236	0	44	1.0	1	0	0	0	126.5
4237	0	52	2.0	0	0	0	0	133.5

4238 rows × 11 columns

In [47]:

```
print('Boarded passengers grouped by part of embarkation(C=Cherbourg,Q=Queenstown,S=Sout
print(df['glucose'].value_counts())
sns.countplot(x='glucose',data=df,palette= 'Set2')
plt.show()
```

```
Boarded passengers grouped by part of embarkation(C=Cherbourg,Q=Queenstow
n,S=Southampton:)
glucose
75.0     193
77.0     167
73.0     156
80.0     152
70.0     152
        ...
386.0      1
155.0      1
147.0      1
205.0      1
260.0      1
Name: count, Length: 143, dtype: int64
```



In [48]:

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

```
40
```

In [49]:

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

In [50]:

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

Out[50]:

```
male               0
age                0
currentSmoker      0
prevalentStroke    0
prevalentHyp       0
diabetes           0
sysBP              0
diaBP              0
glucose            0
TenYearCHD         0
dtype: int64
```

In [52]:

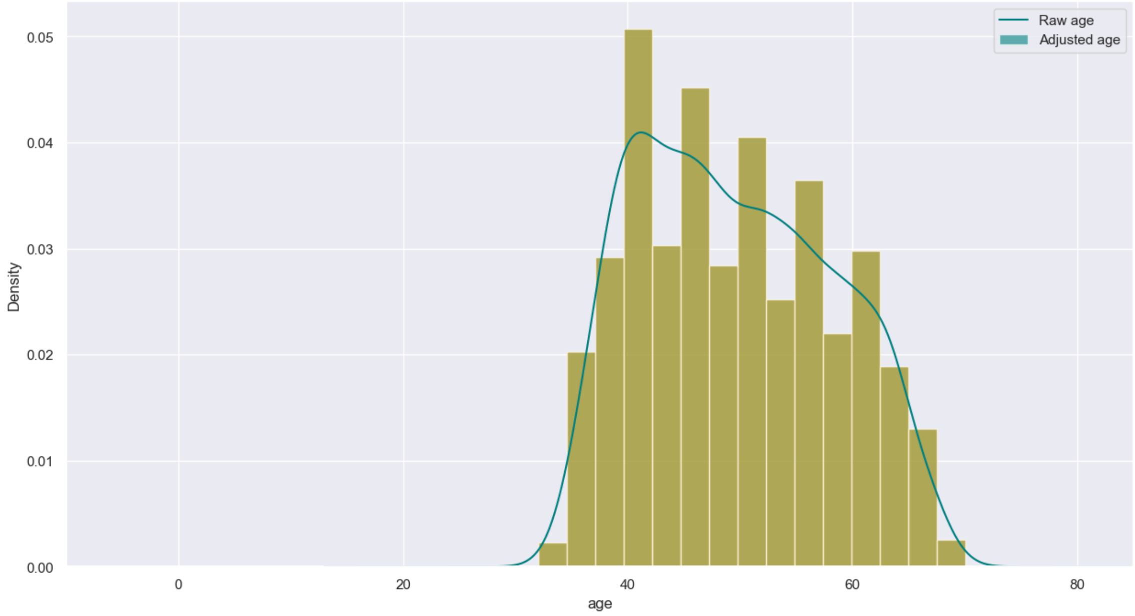
```
train_data.head()
```

Out[52]:

	male	age	currentSmoker	prevalentStroke	prevalentHyp	diabetes	sysBP	diaBP	glucos
0	1	39	0	0	0	0	106.0	70.0	77
1	0	46	0	0	0	0	121.0	81.0	76
2	1	48	1	0	0	0	127.5	80.0	70
3	0	61	1	0	1	0	150.0	95.0	103
4	0	46	1	0	0	0	130.0	84.0	85

In [55]:

```
plt.figure(figsize=(15,8))
ax=df['age'].hist(bins=15,density=-True,stacked=True,color='teal',alpha=0.6)
df['age'].plot(kind='density',color='teal')
ax=train_data["age"].hist(bins=15,density=True,stacked=True,color='orange',alpha=0.5)
ax.legend(['Raw age','Adjusted age'])
ax.set(xlabel='age')
plt.xlim(-10,85)
plt.show()
```



In [56]:

```
train_data['TravalAlone']=np.where((train_data["diaBP"]+train_data["sysBP"])>0,0,1)
train_data.drop('diaBP',axis=1,inplace=True)
train_data.drop('sysBP',axis=1,inplace=True)
```