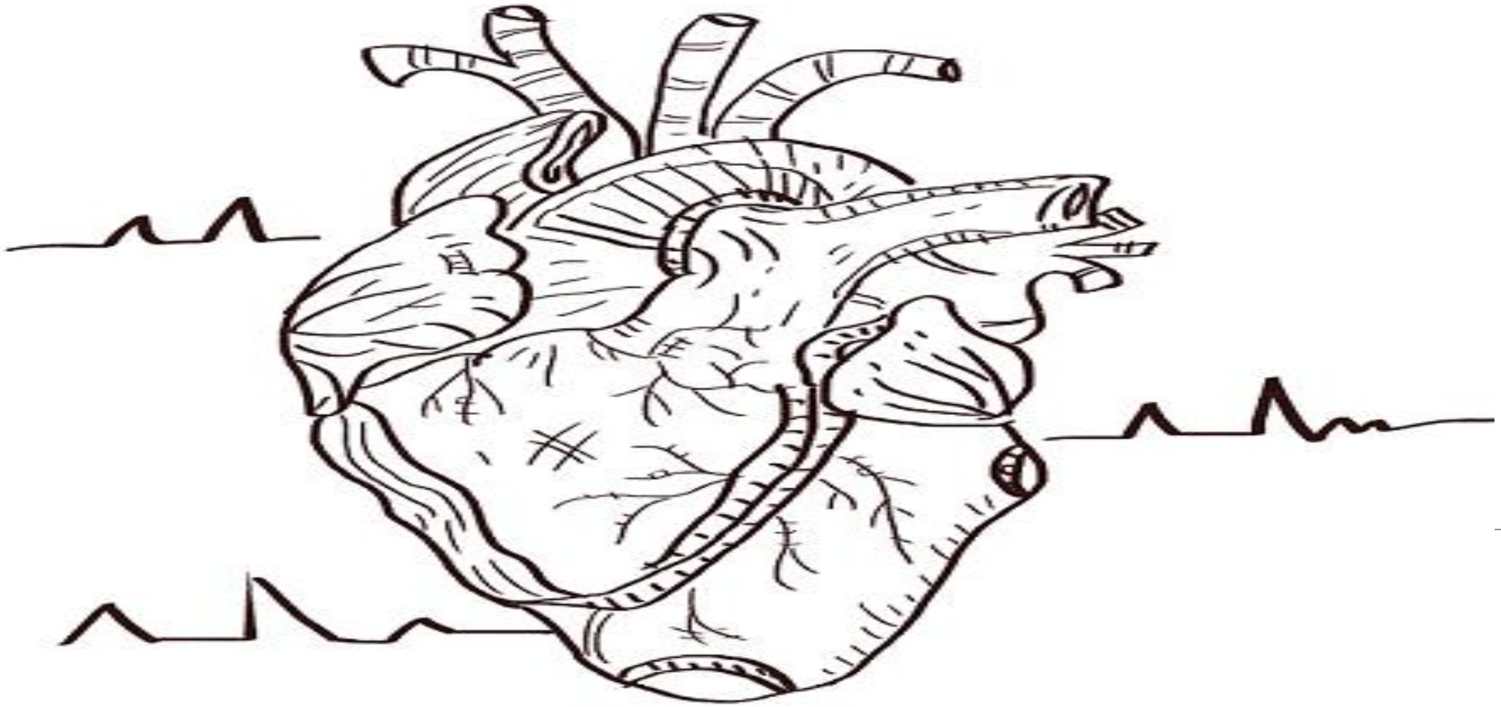
Heart Disease Prediction



Introduction

Several health conditions, your lifestyle, and your age and family history can increase your risk for heart disease. These are called risk factors. About half of all Americans (47%) have at least 1 of 3 key risk factors for heart disease: high blood pressure, high cholesterol, and smoking. Other key indicator like diabetic status, drinking too much alcohol. Detecting and preventing the factors that have the greatest impact on heart disease is very important in healthcare.

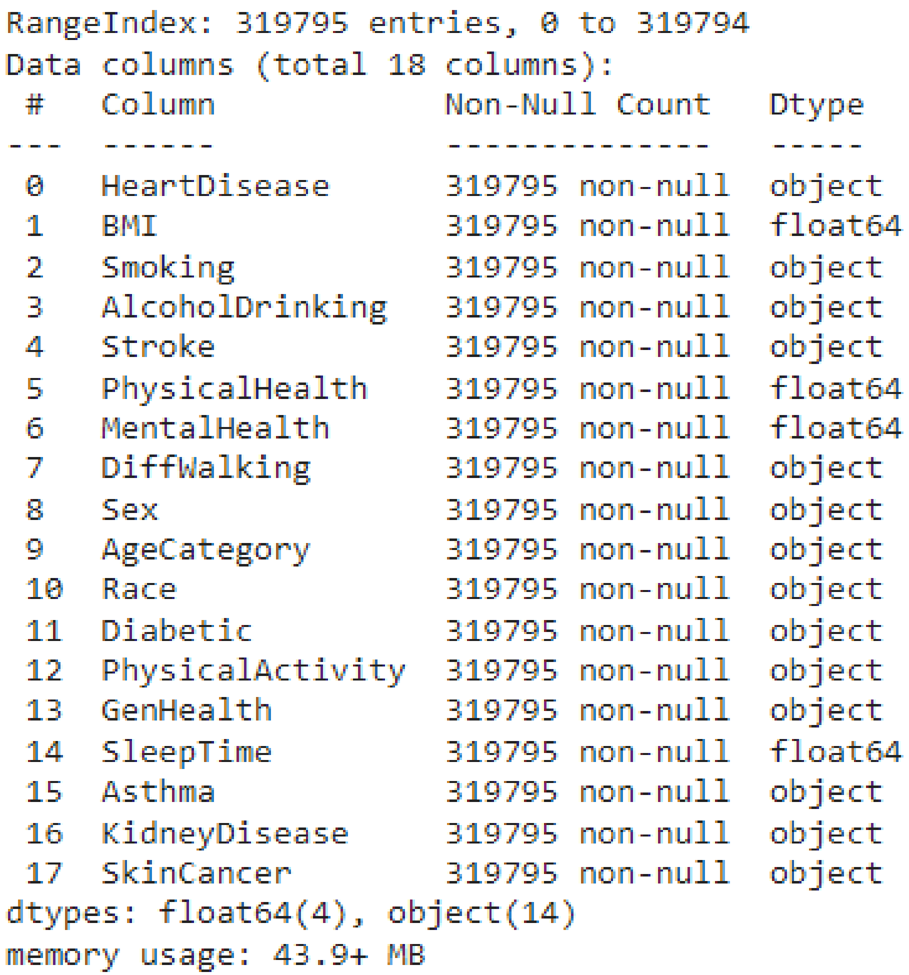
Dataset

The dataset come from the Centers for Disease control and prevention (CDC), which conducts annual telephone surveys to gather data on the health status of U.S. residents.
The dataset contains 18 variables and 319795 entries.

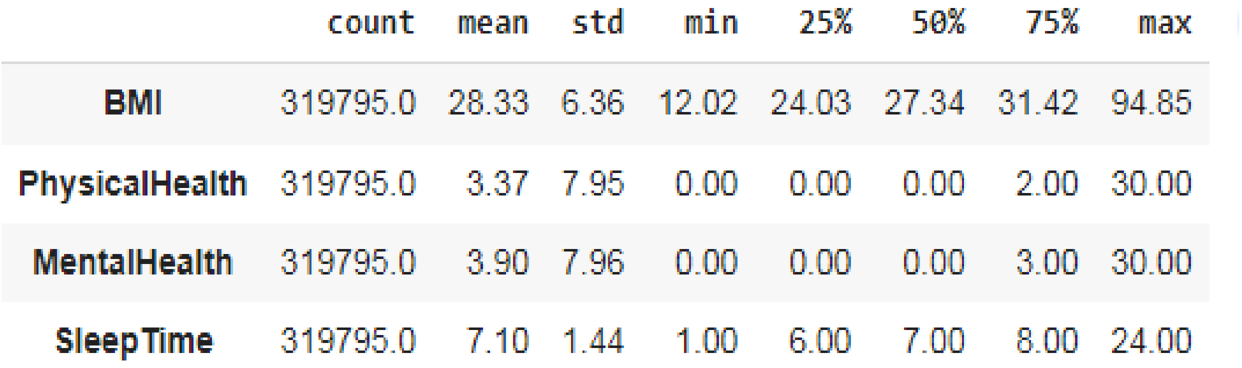
Explanation of the features of the dataset

1. HeartDisease (target) : take value yes or no.
2. Smoking : Have you smoked at least 100 cigarettes in your entire life? (The answer Yes or No).
3. AlcoholDrinking : Heavy drinkers (adult men having more than 14 drinks per week and adult women having more than 7 drinks per week
4. Stroke : (Ever told) (you had) a stroke?
5. PhysicalHealth : Now thinking about your physical health, which includes physical illness and injury, for how many days during the past 30 days was your physical health not good? (0-30 days).
6. MentalHealth : Thinking about your mental health, for how many days during the past 30 days was your mental health not good? (0-30 days).
7. DiffWalking : Do you have serious difficulty walking or climbing stairs?
8. Sex : Are you male or female?
9. AgeCategory: Fourteen-level age category.
10. Race : Imputed race/ethnicity value.
11. Diabetic : (Ever told) (you had) diabetes?
12. PhysicalActivity : Adults who reported doing physical activity or exercise during the past 30 days other than their regular job.
13. GenHealth : Would you say that in general your health is...
14. SleepTime : On average, how many hours of sleep do you get in a 24-hour period?
15. Asthma : (Ever told) (you had) asthma?
16. KidneyDisease : Not including kidney stones, bladder infection or incontinence, were you ever told you had kidney disease?
17. SkinCancer : (Ever told) (you had) skin cancer?

Information for data



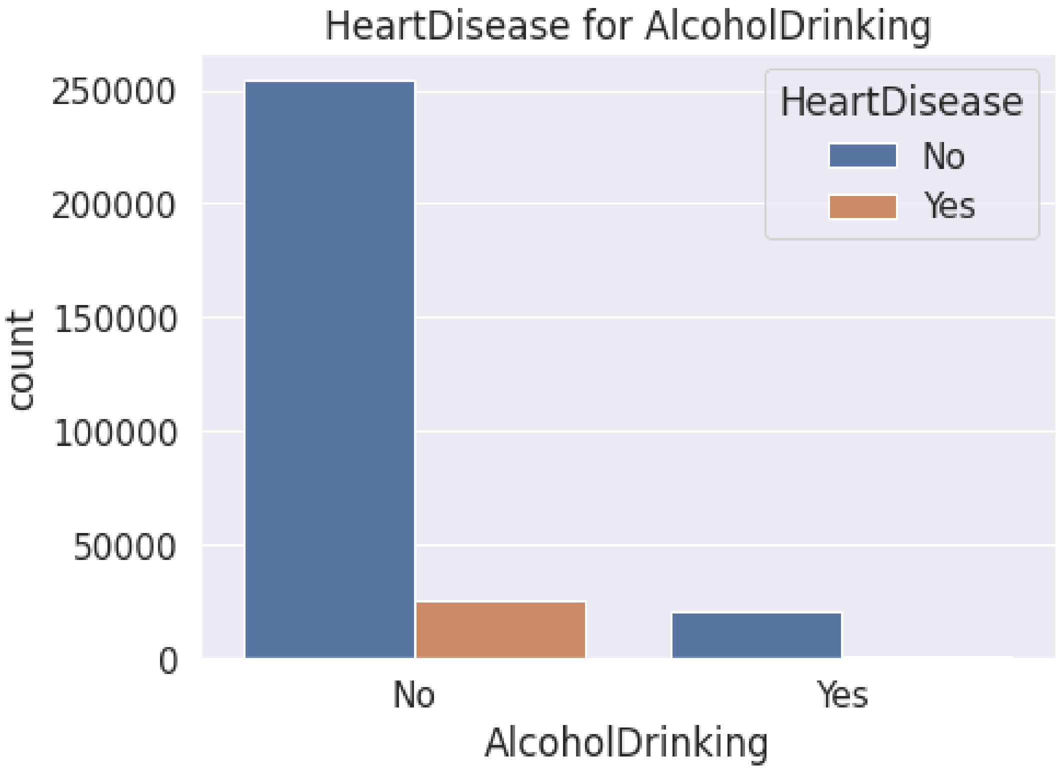
```
RangeIndex: 319795 entries, 0 to 319794
Data columns (total 18 columns):
 #   Column            Non-Null Count   Dtype  
---  ------            --------------   -----  
 0   HeartDisease      319795 non-null  object 
 1   BMI               319795 non-null  float64
 2   Smoking           319795 non-null  object 
 3   AlcoholDrinking   319795 non-null  object 
 4   Stroke            319795 non-null  object 
 5   PhysicalHealth    319795 non-null  float64
 6   MentalHealth      319795 non-null  float64
 7   DiffWalking       319795 non-null  object 
 8   Sex               319795 non-null  object 
 9   AgeCategory       319795 non-null  object 
 10  Race              319795 non-null  object 
 11  Diabetic          319795 non-null  object 
 12  PhysicalActivity  319795 non-null  object 
 13  GenHealth         319795 non-null  object 
 14  SleepTime         319795 non-null  float64
 15  Asthma            319795 non-null  object 
 16  KidneyDisease     319795 non-null  object 
 17  SkinCancer        319795 non-null  object 
dtypes: float64(4), object(14)
memory usage: 43.9+ MB
```



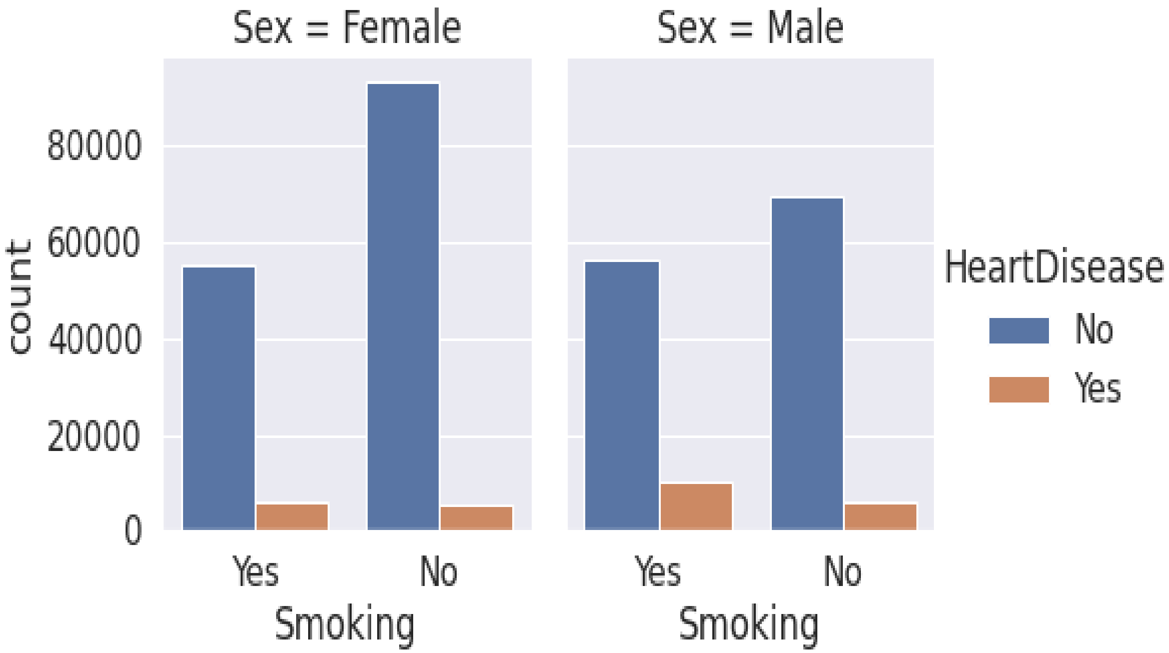
	count	mean	std	min	25%	50%	75%	max
BMI	319795.0	28.33	6.36	12.02	24.03	27.34	31.42	94.85
PhysicalHealth	319795.0	3.37	7.95	0.00	0.00	0.00	2.00	30.00
MentalHealth	319795.0	3.90	7.96	0.00	0.00	0.00	3.00	30.00
SleepTime	319795.0	7.10	1.44	1.00	6.00	7.00	8.00	24.00

Analysis Questions

1-The count of heart disease for AlcoholDrinking people?

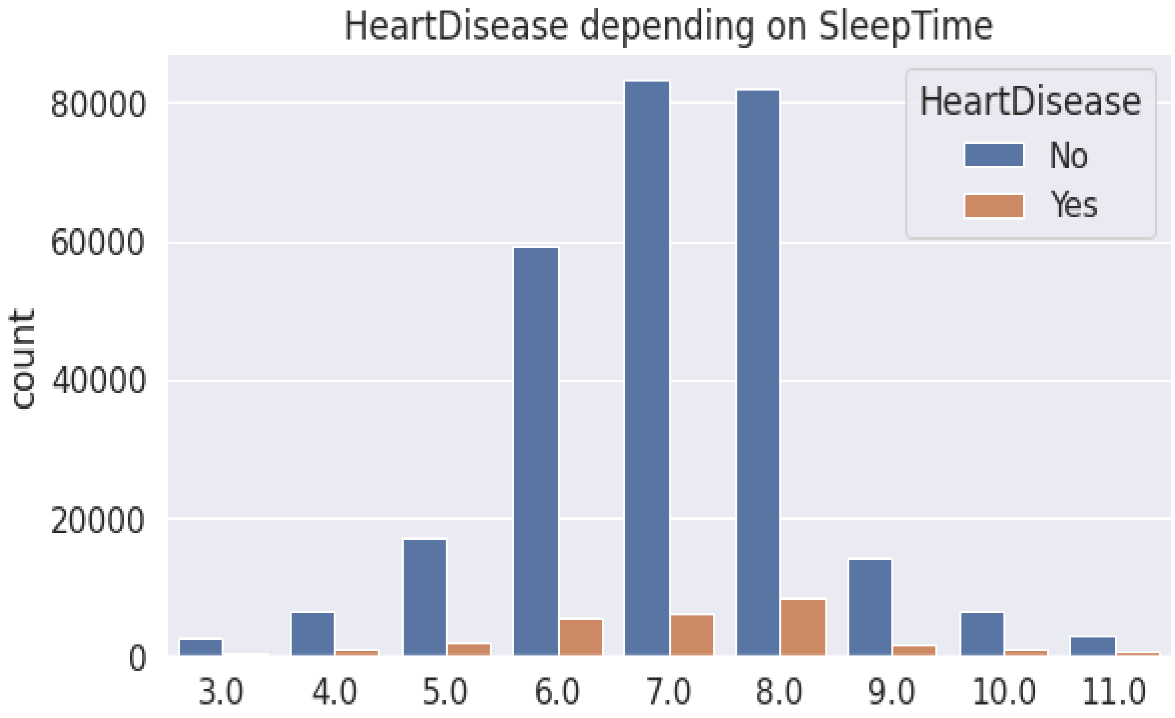


2-The count of heart disease for Smoking people for each sex?

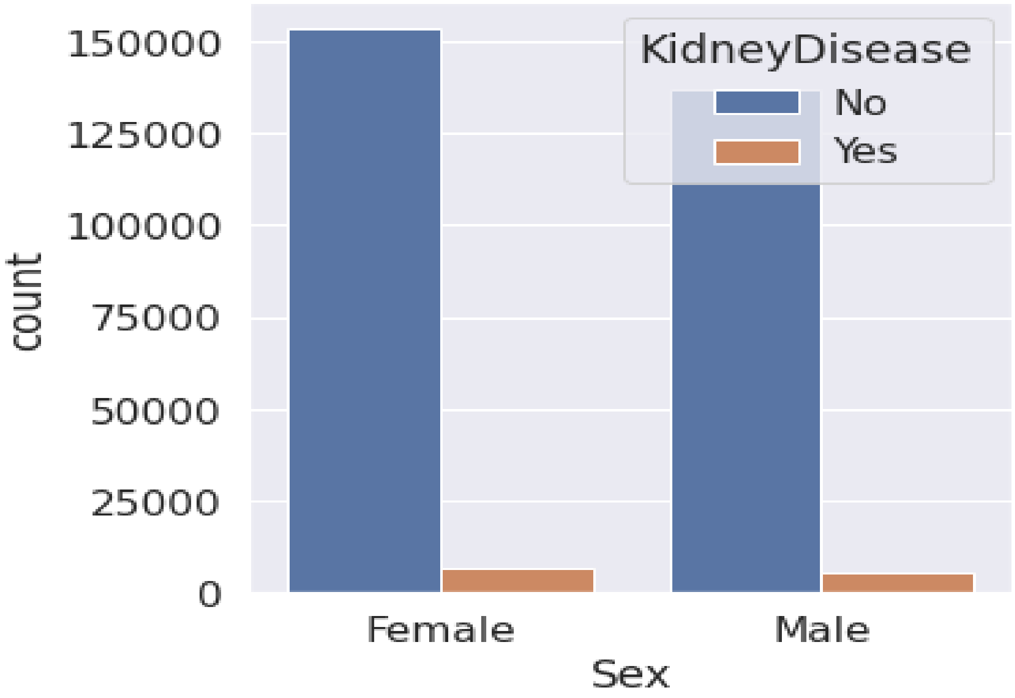


Analysis Questions

3-The count of heart disease depending on SleepTime?

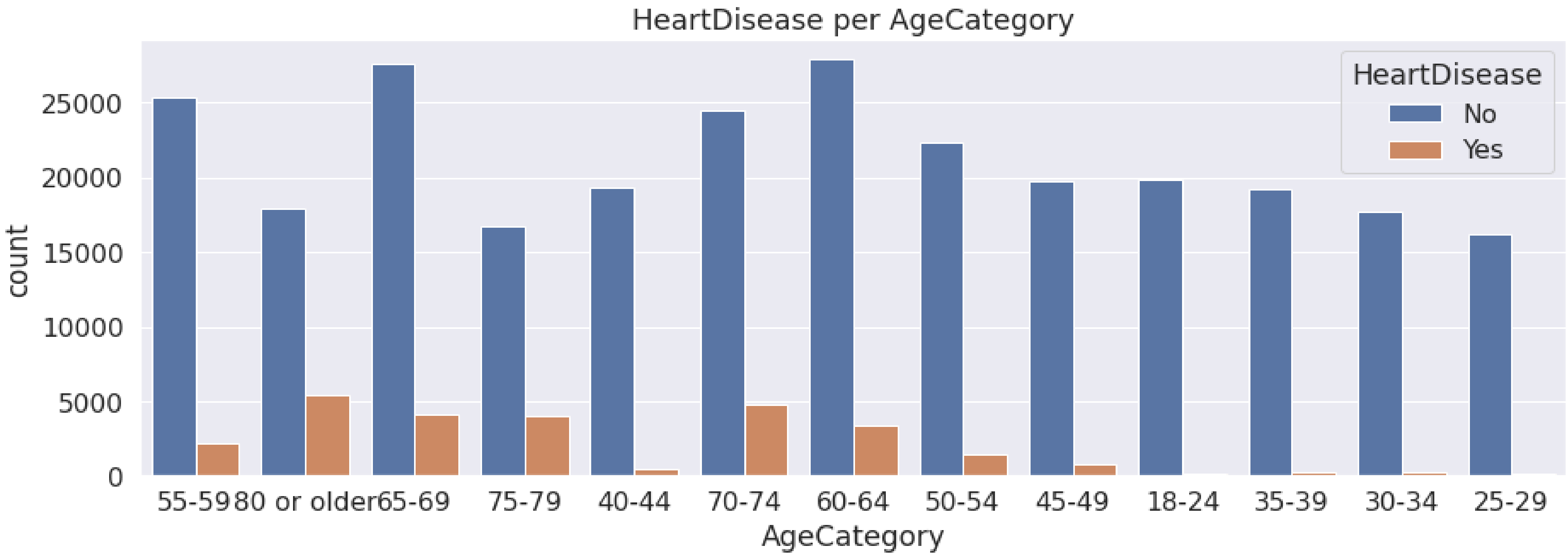


4-The count of heartdisease for KidneyDisease for each sex?

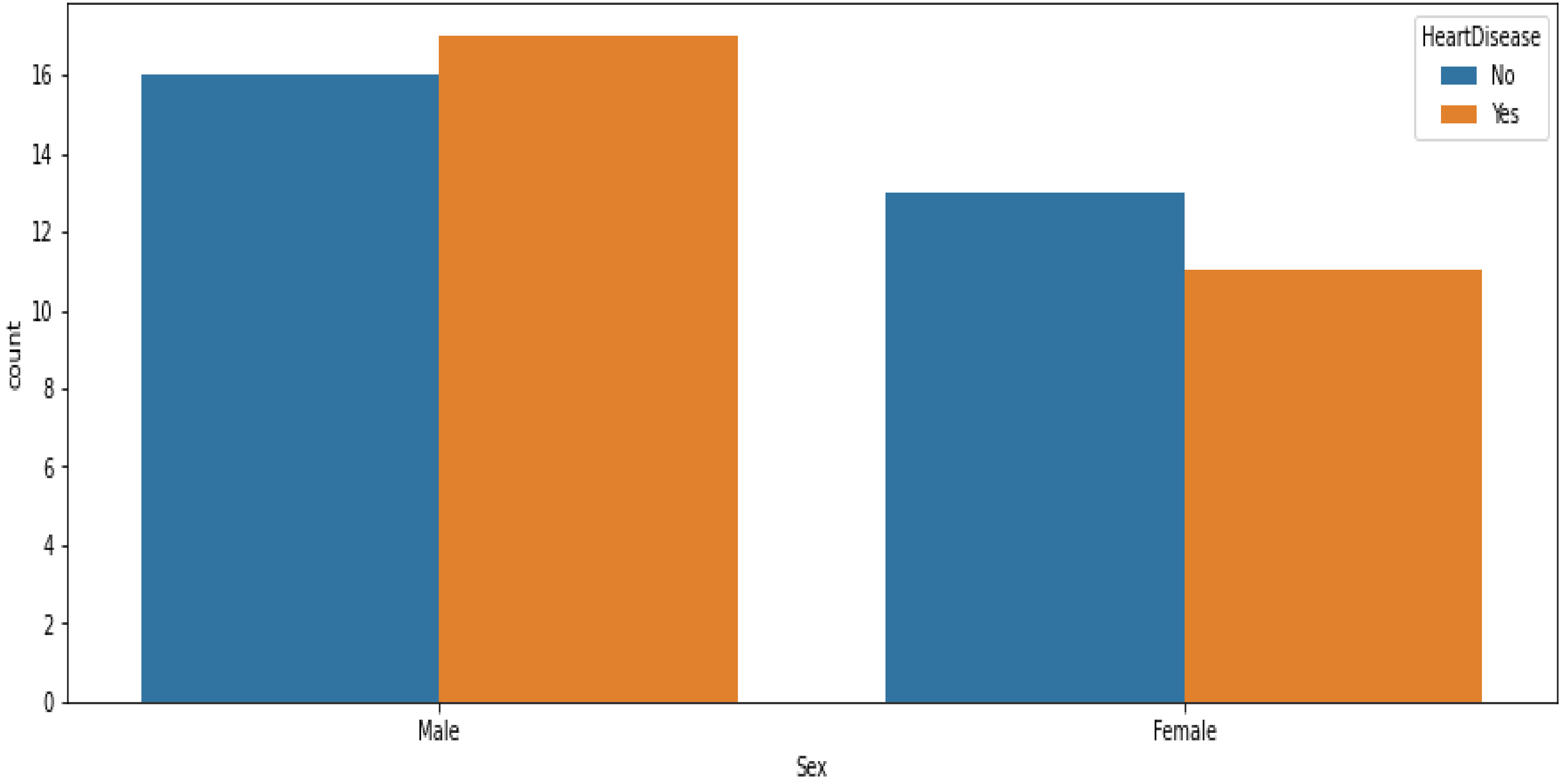


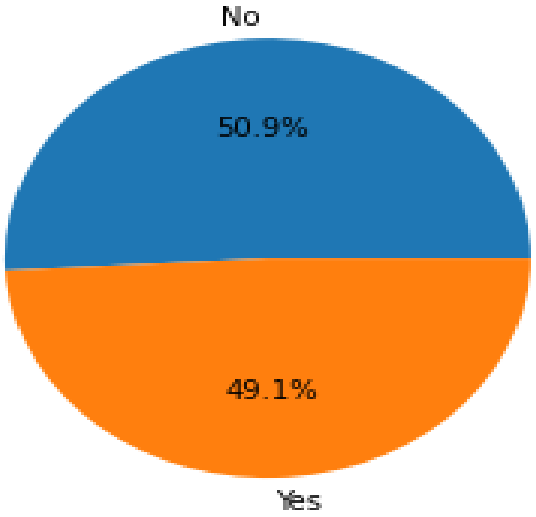
Analysis Questions

5-The count of heart disease for each AgeCategory?



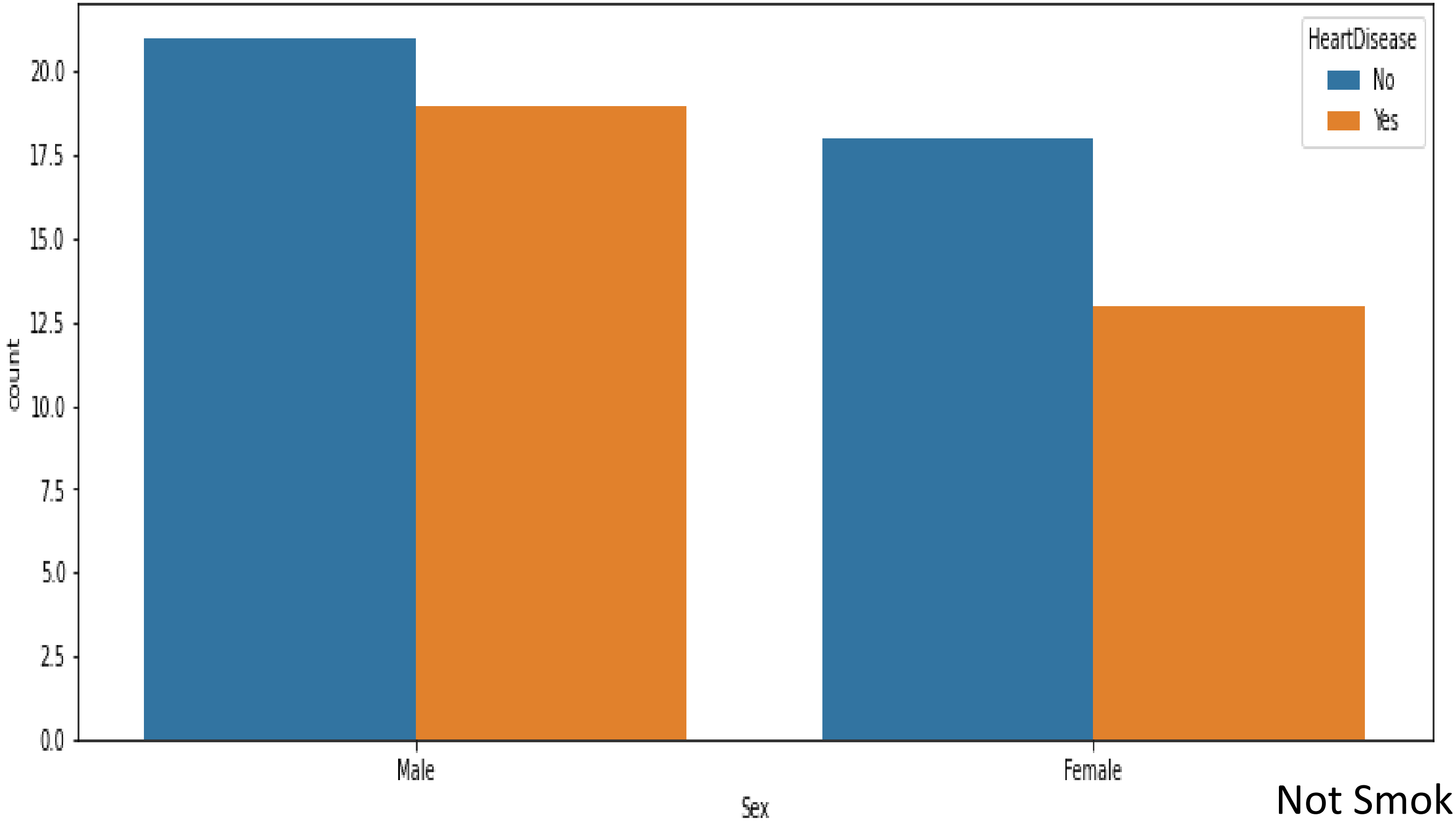
6-The count of heartdisease of smoking people that have age older than 80 and not walking and genhealth poor and have kidney or stroke disease for each sex?

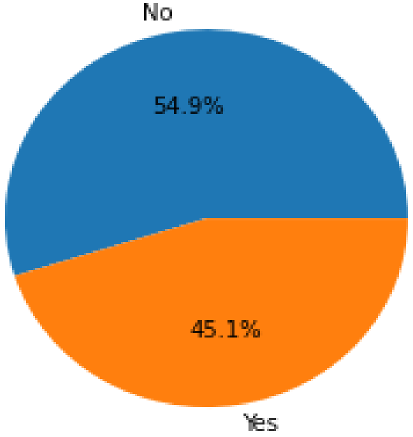




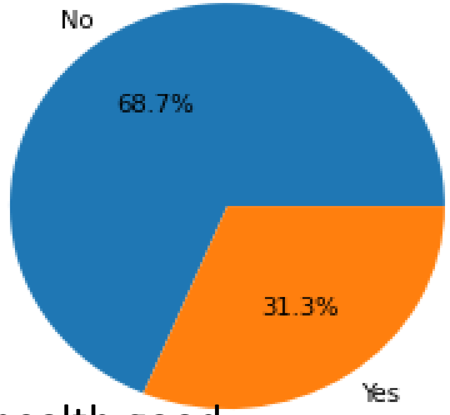
Percentage for all sex

7-The count of heart disease of smoking people that have age older than 80 and not walking and genhealth poor and have kidney or censer disease for each sex?



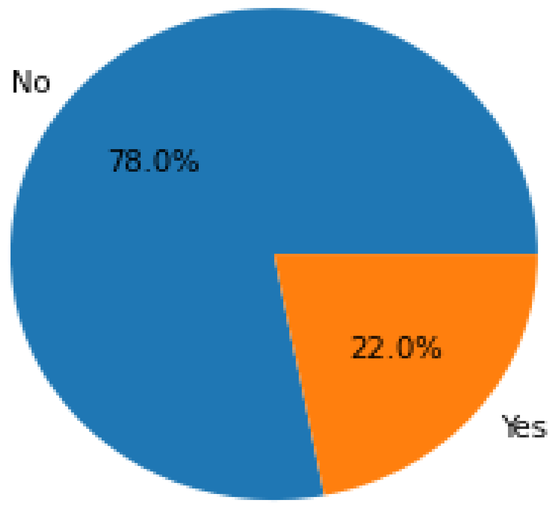


Smoking &general health poor

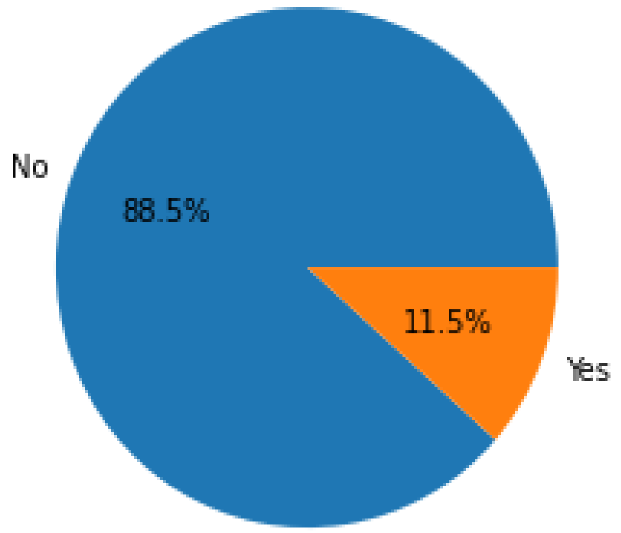


Not Smoking &general health good

8-The count of heart disease that have Diabetic?



9-The count of heart disease that have Asthma?

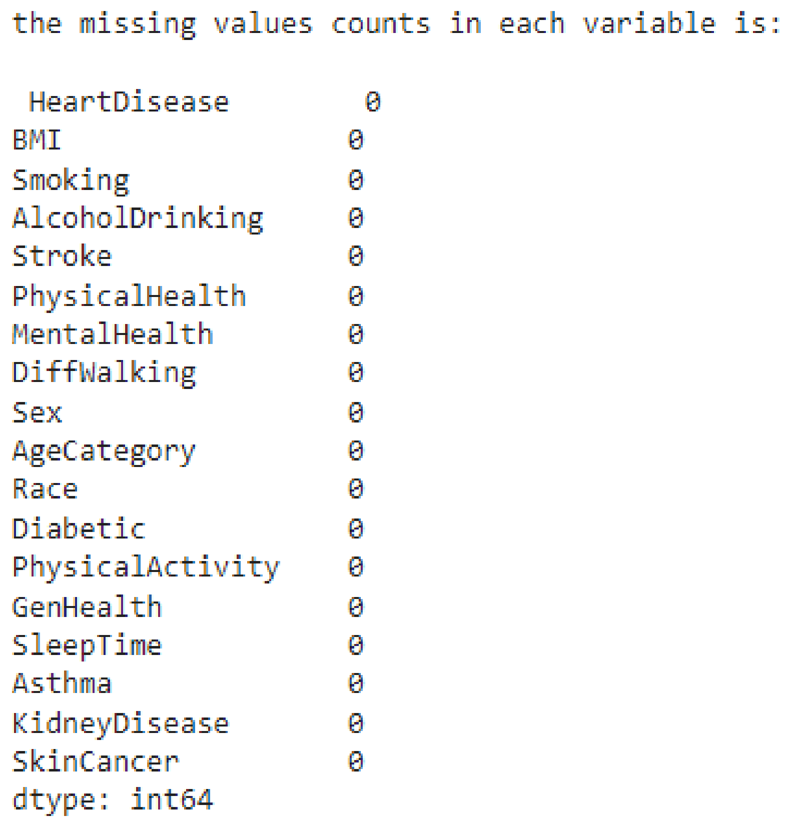


Decision

Throug the analysis of data, we found that the smoking people of age older than 80 and have diseses like kidney , canser and stoke and the general health is poor and they do not make any physical health or walking are the most people that have heart disese. so we make a Awareness for those people and another that not have any diseses to take off smoking and make physical activity and walking to preserve of a good health .

Processing

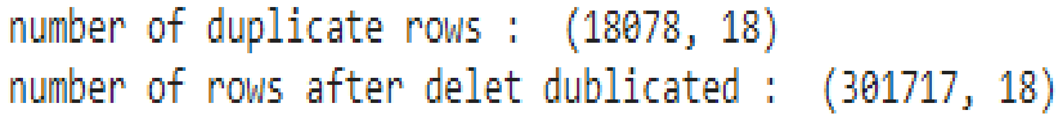
1-Missing values and Duplicated rows



```
the missing values counts in each variable is:

 HeartDisease        0
BMI                  0
Smoking              0
AlcoholDrinking      0
Stroke               0
PhysicalHealth       0
MentalHealth         0
DiffWalking          0
Sex                  0
AgeCategory          0
Race                 0
Diabetic             0
PhysicalActivity     0
GenHealth            0
SleepTime            0
Asthma               0
KidneyDisease        0
SkinCancer           0
dtype: int64
```

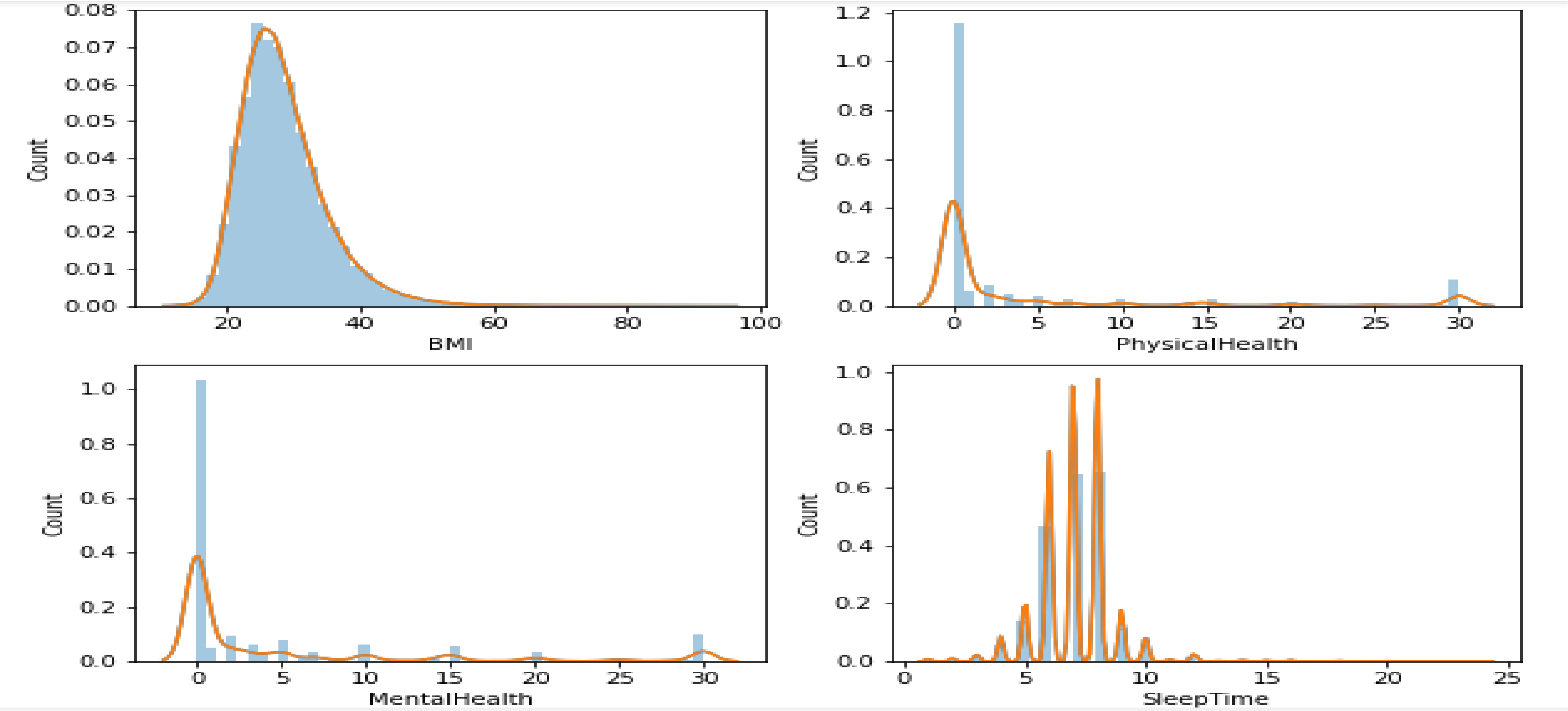
Duplicated rows



```
number of duplicate rows :  (18078, 18)
number of rows after delet dublicated :  (301717, 18)
```

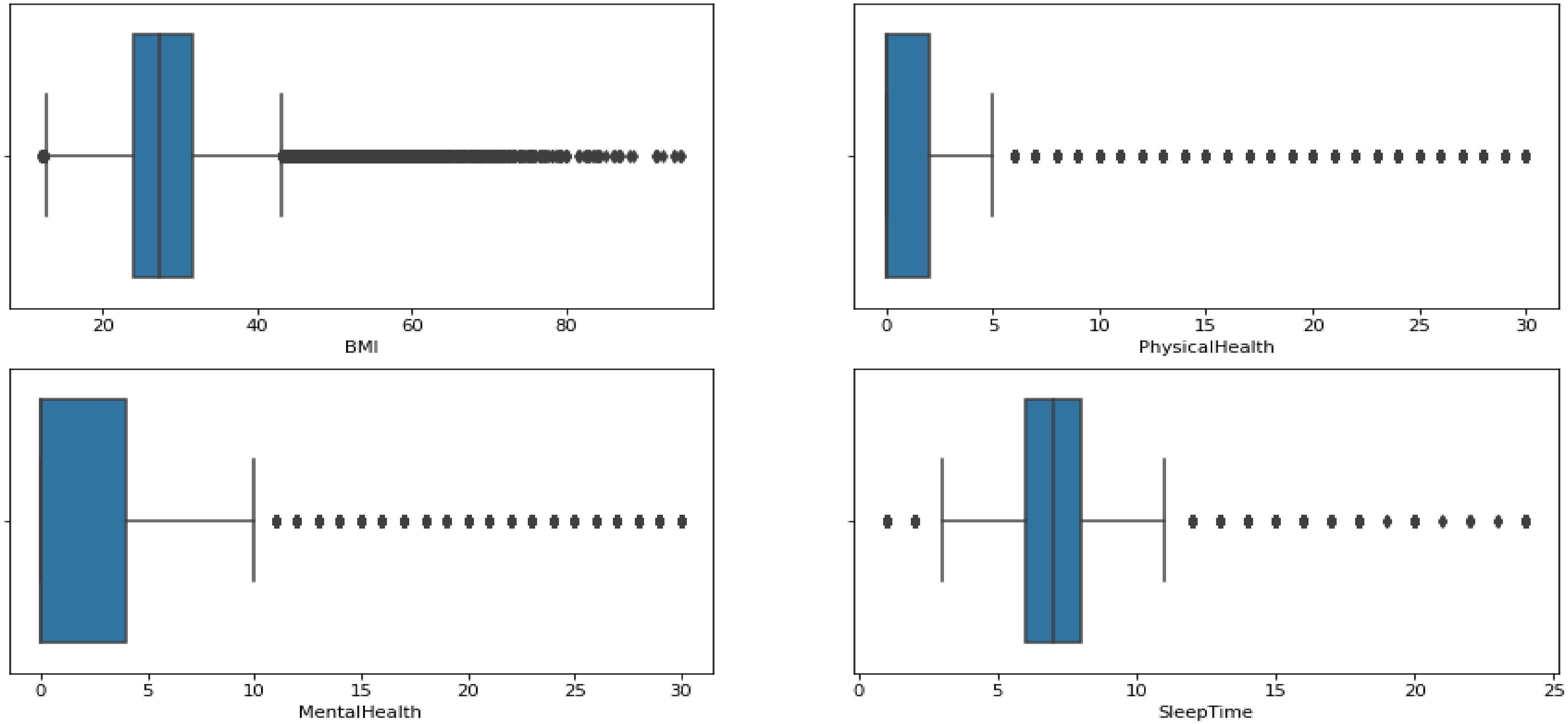
2-Numerical data distribution

Num_col=[BMI, PhysicalHealth, MentalHealth, SleepTime]



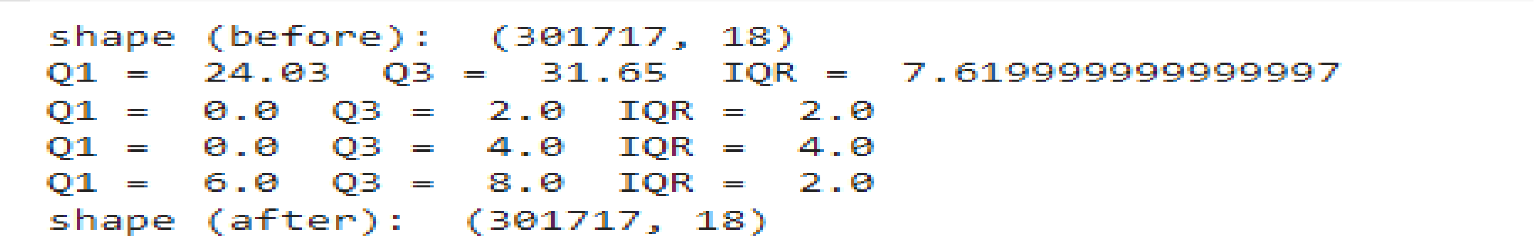
3-Handling outlier

```
the number of outliers: 79186
```

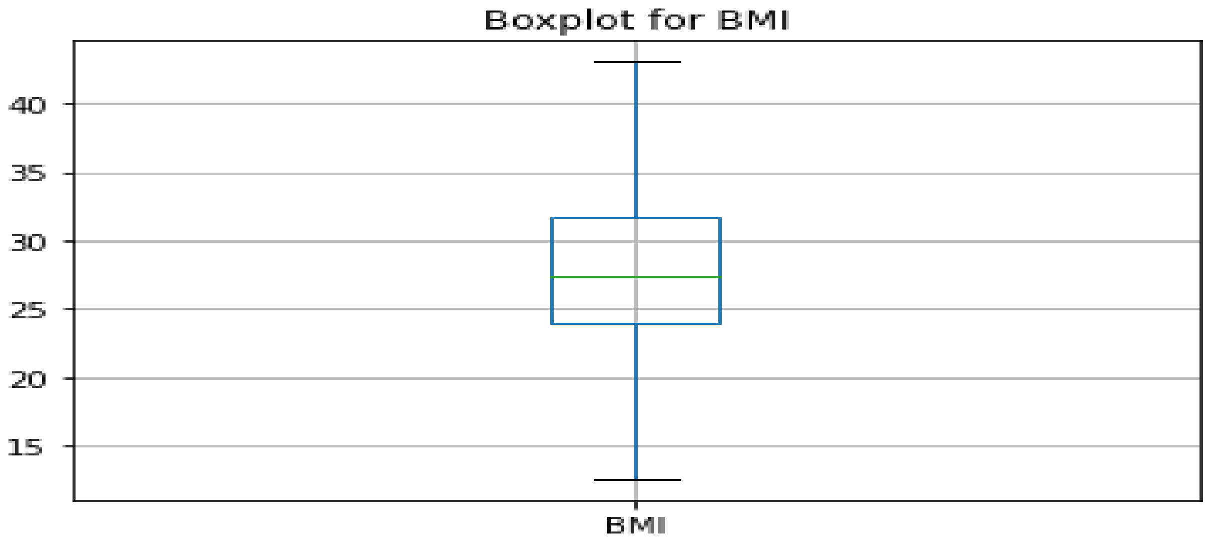


3-Handling outlier

Num_col=[BMI, PhysicalHealth, MentalHealth, SleepTime]



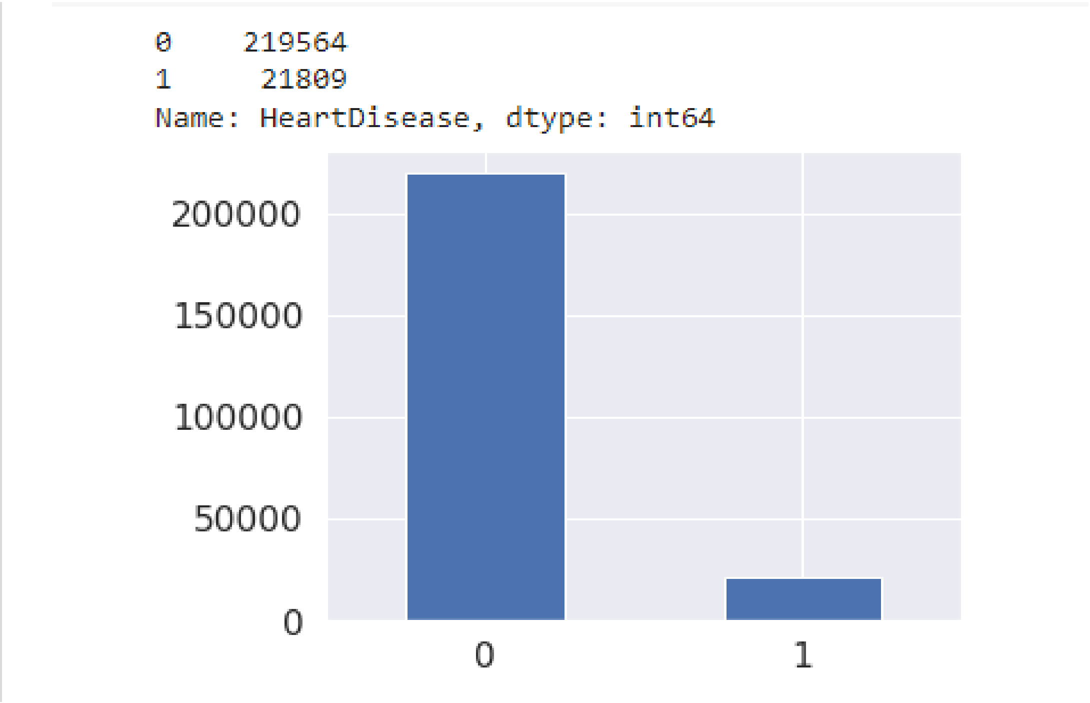
```
shape (before):  (301717, 18)
Q1 =  24.03  Q3 =  31.65  IQR =  7.619999999999997
Q1 =  0.0  Q3 =  2.0  IQR =  2.0
Q1 =  0.0  Q3 =  4.0  IQR =  4.0
Q1 =  6.0  Q3 =  8.0  IQR =  2.0
shape (after):  (301717, 18)
```



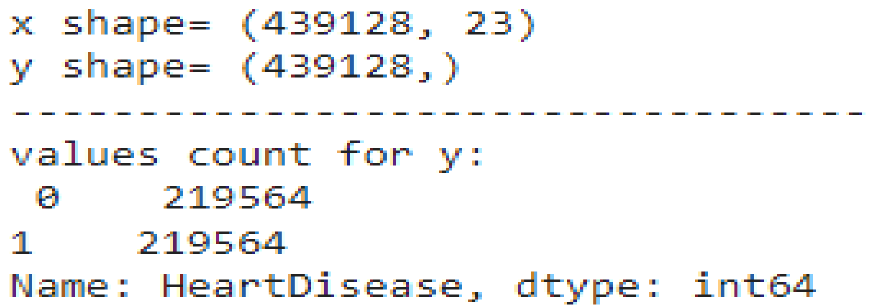
4-Handling categorical

```
df['GenHealth'].map({'Excellent':5,'Very good':4,'Good':3,'Fair':2,'Poor':1})
df['HeartDisease'].map({'Yes':1,'No':0})
df['AgeCategory']=df['AgeCategory'].map({'18-24':18,'25-29':25,'30-34':30,'35-
39':35,'40-44':40,'45-49':45,'50-54':50,'55-59':55,'60-64':60,'65-69':65,'70-
74':70,'75-79':75,'80 or older':80})
df = pd.get_dummies(data=df,drop_first=True)
```

Split data and handle imbalanced



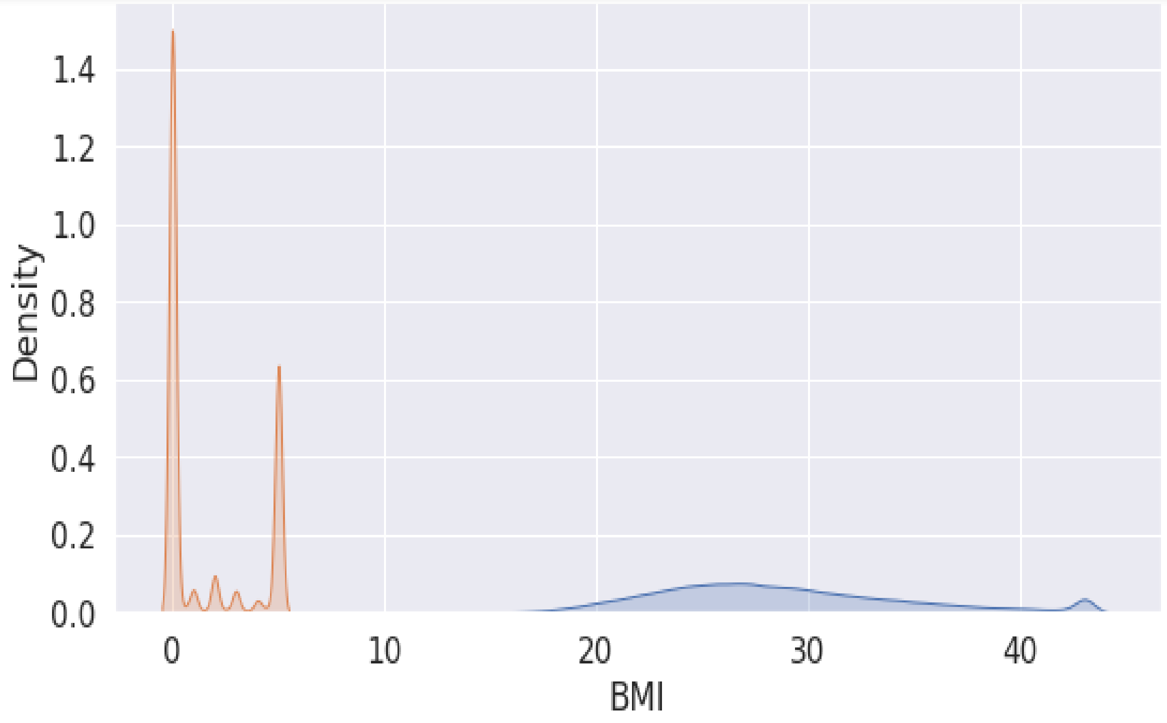
```
0    219564
1     21809
Name: HeartDisease, dtype: int64
```



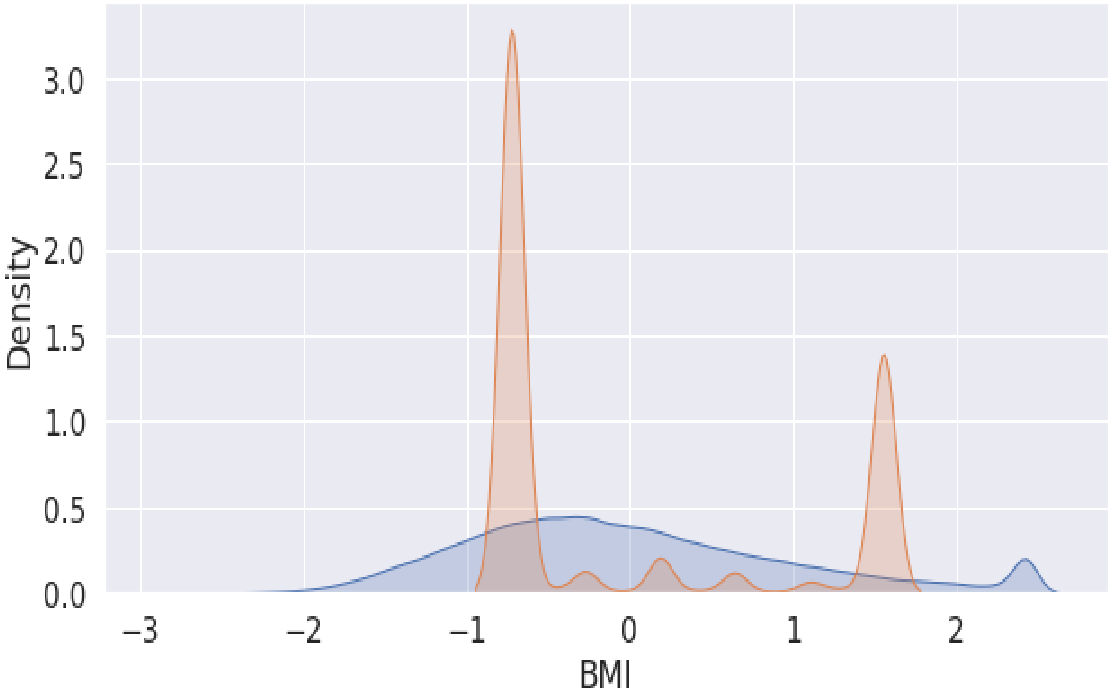
```
x shape= (439128, 23)
y shape= (439128,)
-----------------------------------
values count for y:
 0    219564
1    219564
Name: HeartDisease, dtype: int64
```

Feature scaling using standerdscaler

PhysicalHealth



before



after