In [1]:
```
!wget https://raw.githubusercontent.com/keerthy456/Machine-Learning-Final-Project-Vakkalagadda-Keerthi/main/heart_disease.csv
```

```
--2022-04-28 00:04:25--  https://raw.githubusercontent.com/keerthy456/Machine-Learning-Final-Project-Vakkalagadda-Keerthi/main/heart_disease.csv
Resolving raw.githubusercontent.com (raw.githubusercontent.com)... 185.199.108.133, 185.199.109.133, 185.199.110.133, ...
Connecting to raw.githubusercontent.com (raw.githubusercontent.com)|185.199.108.133|:443... connected.
HTTP request sent, awaiting response... 200 OK
Length: 25189554 (24M) [text/plain]
Saving to: ‘heart_disease.csv.2’

heart_disease.csv.2 100%[===================>]  24.02M   133MB/s    in 0.2s

2022-04-28 00:04:26 (133 MB/s) - ‘heart_disease.csv.2’ saved [25189554/25189554]
```

In [ ]:
```
pip install dython
```

In [4]:
```
import numpy as np
import pandas as pd
from dython.nominal import associations
import os
%matplotlib inline
import matplotlib.pyplot as plt
import seaborn as sns
```

In [5]:
```
heart_df = pd.read_csv('heart_disease.csv')
```

In [6]:
```
heart_df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
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
```

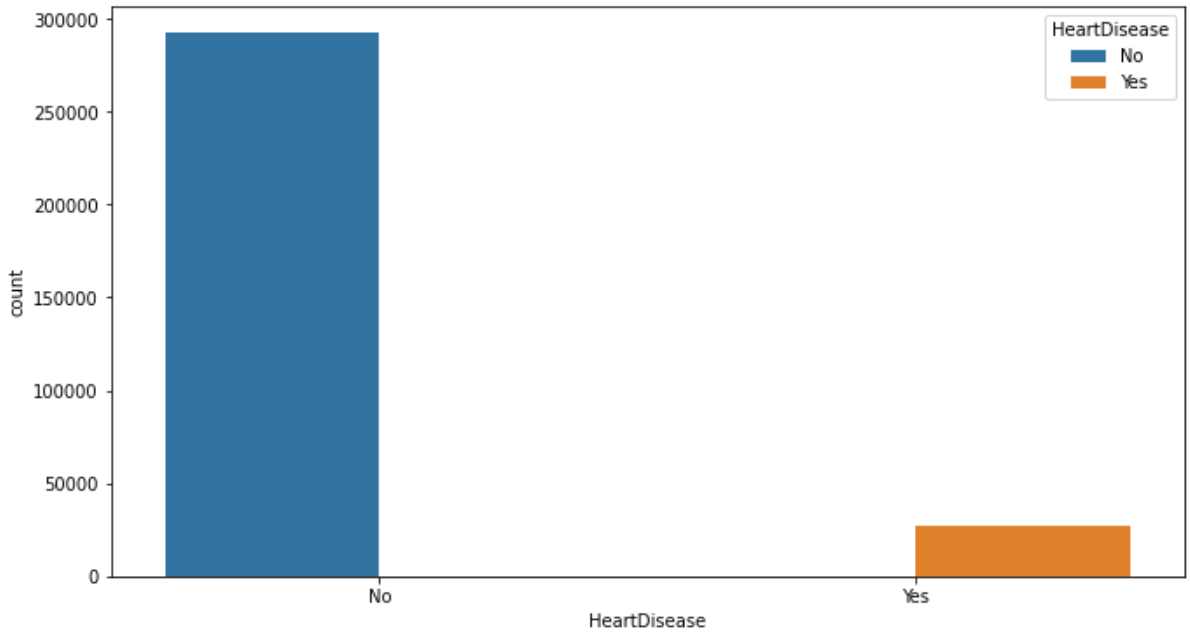
Target Variable Analysis

In [237]:
```
heart_df['HeartDisease'].value_counts()
```

Out[237]:
```
No     292422
Yes     27373
Name: HeartDisease, dtype: int64
```

In [236]:
```
plt.figure(figsize = (11,6))
sns.countplot(x = heart_df['HeartDisease'], hue = 'HeartDisease', data = heart_df)
```

Out[236]: <AxesSubplot:xlabel='HeartDisease', ylabel='count'>



In [ ]:
```

```

In [8]:
```
numeric_features = heart_df.select_dtypes(include=[np.number])
```

In [9]:
```
numerical_df = heart_df[['BMI', 'PhysicalHealth', 'MentalHealth', 'SleepTime', 'HeartDisease']]
```

In [10]:
```
categorical_features= [col for col in heart_df.columns if heart_df[col].dtypes == 'object']
```

In [11]:
```
for feature in categorical_features:
    print(feature, ":", heart_df[feature].unique())
    print()
```

```
HeartDisease : ['No' 'Yes']

Smoking : ['Yes' 'No']

AlcoholDrinking : ['No' 'Yes']

Stroke : ['No' 'Yes']

DiffWalking : ['No' 'Yes']

Sex : ['Female' 'Male']

AgeCategory : ['55-59' '80 or older' '65-69' '75-79' '40-44' '70-74' '60-64'
'50-54'
 '45-49' '18-24' '35-39' '30-34' '25-29']

Race : ['White' 'Black' 'Asian' 'American Indian/Alaskan Native' 'Other'
 'Hispanic']

Diabetic : ['Yes' 'No' 'No, borderline diabetes' 'Yes (during pregnancy)']

PhysicalActivity : ['Yes' 'No']

GenHealth : ['Very good' 'Fair' 'Good' 'Poor' 'Excellent']

Asthma : ['Yes' 'No']

KidneyDisease : ['No' 'Yes']

SkinCancer : ['Yes' 'No']
```

Zero Null values are detected in dataset

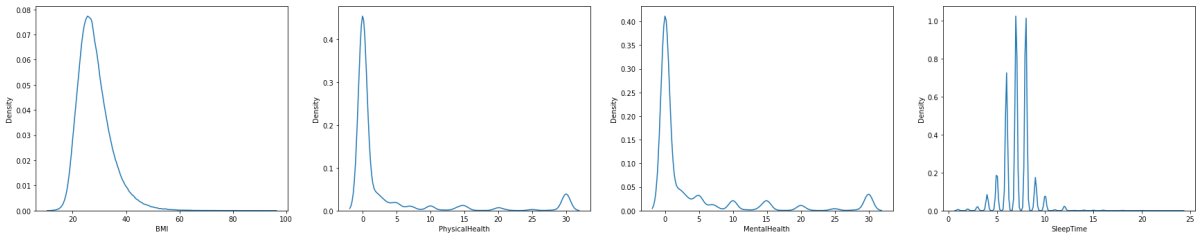
In [168]:
```
heart_df[heart_df.isnull().any(axis=1)]
```

Out[168]:

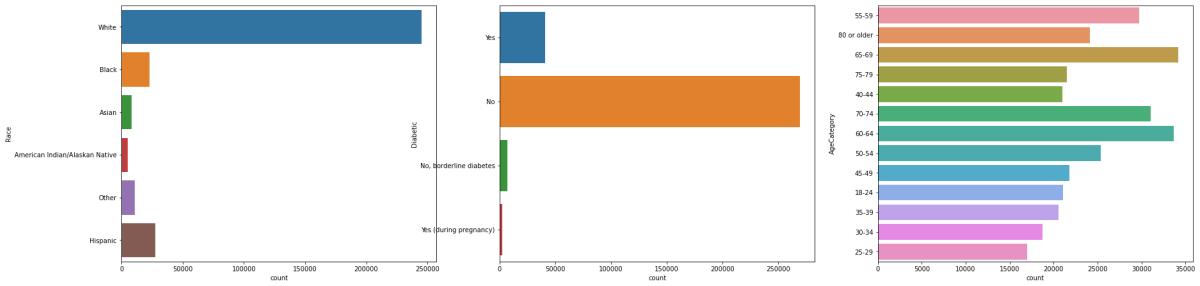
	HeartDisease	BMI	Smoking	AlcoholDrinking	Stroke	PhysicalHealth	MentalHealth	DiffWalking

Distribution of Each Variable

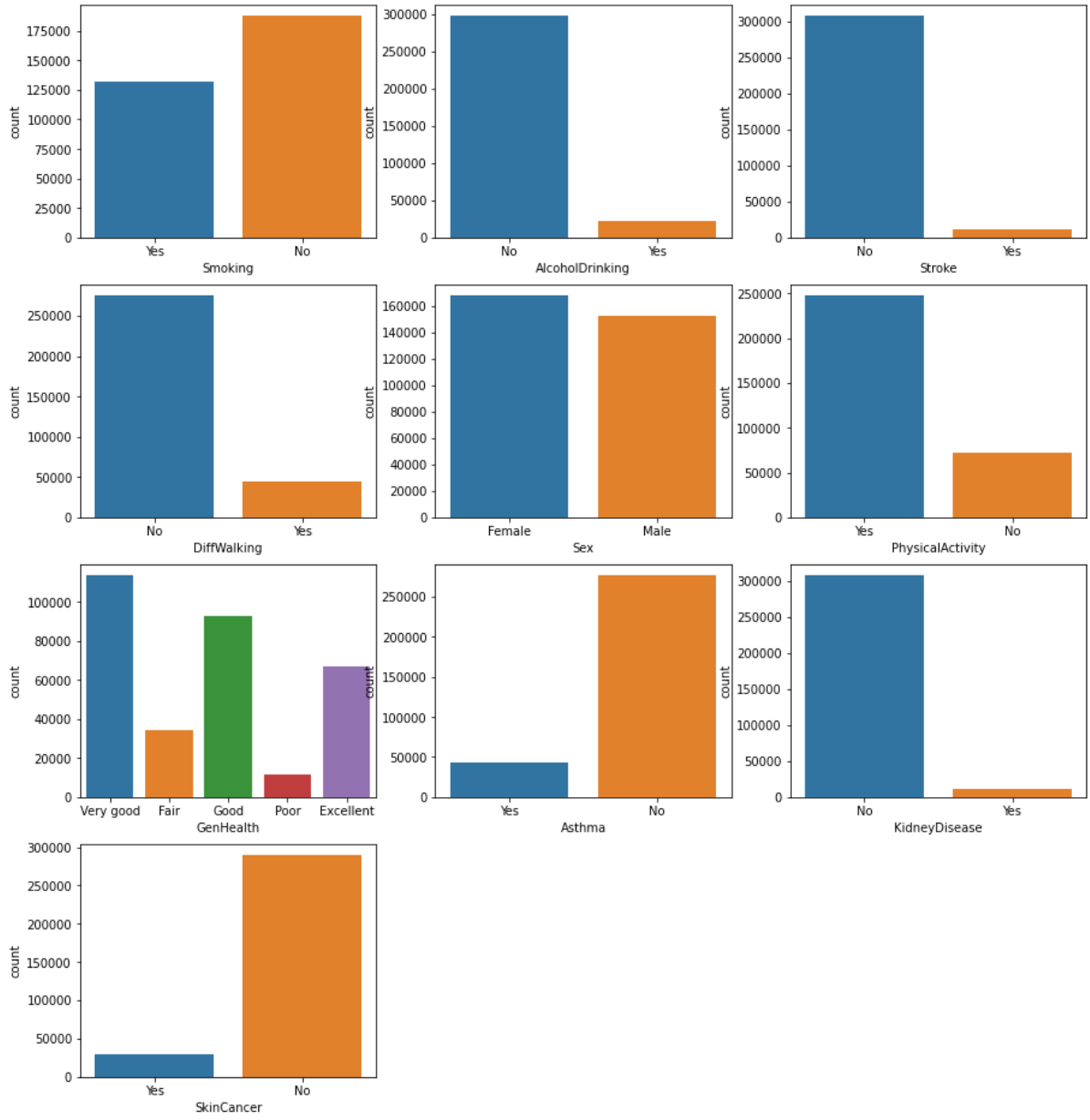
In [261]:
```
size = 1
plt.figure(figsize = (30,25))
for feature in numeric_features:
    plt.subplot(4,4,size)
    sns.kdeplot(x = feature , data = heart_df)
    size = size+1
```



In [258]:
```
size = 1
plt.figure(figsize = (30,25))
for feature in [ 'Race','Diabetic', 'AgeCategory']:
    plt.subplot(3,3,size)
    sns.countplot(y = feature, data = heart_df)
    size = size+1
```

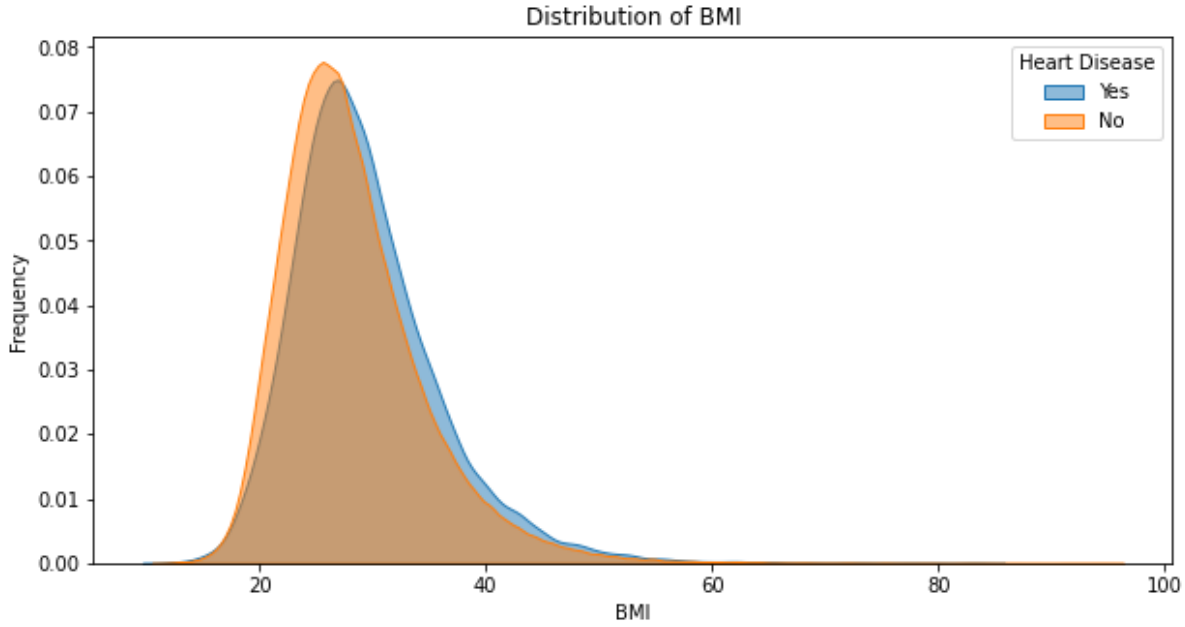


In [244]:
```
size = 1
plt.figure(figsize = (15,25))
for feature in categorical_features:
  if(not(feature in ['HeartDisease', 'Race','Diabetic', 'AgeCategory'])):
    plt.subplot(6,3,size)
    sns.countplot(x = feature , data = heart_df)
    size = size+1
```

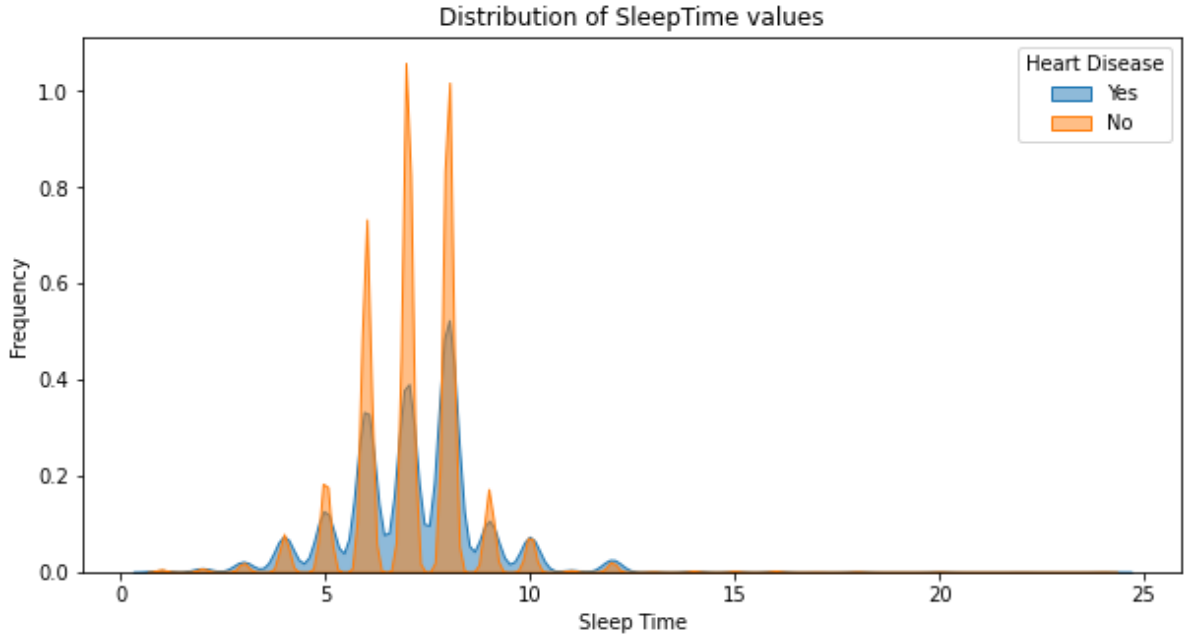


Distribution of Numerical Features based on Target Feature

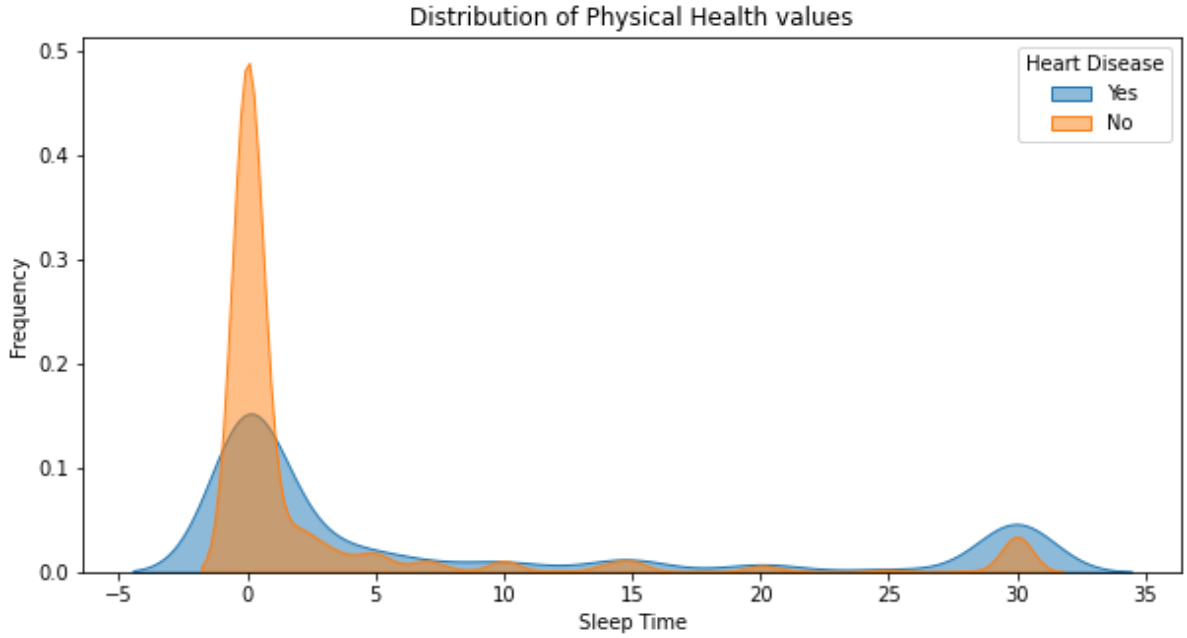
In [175]:
```
fig, axes = plt.subplots(figsize = (10,5))
sns.kdeplot(heart_df[heart_df["HeartDisease"]=='Yes']["BMI"], alpha=0.5,shade
= True, label="Yes", ax = axes)
sns.kdeplot(heart_df[heart_df["HeartDisease"]=='No']["BMI"], alpha=0.5,shade =
True, label="No", ax = axes)
plt.title('Distribution of BMI')
axes.set_xlabel("BMI")
axes.set_ylabel("Frequency")
axes.legend(title='Heart Disease')
plt.show()
```



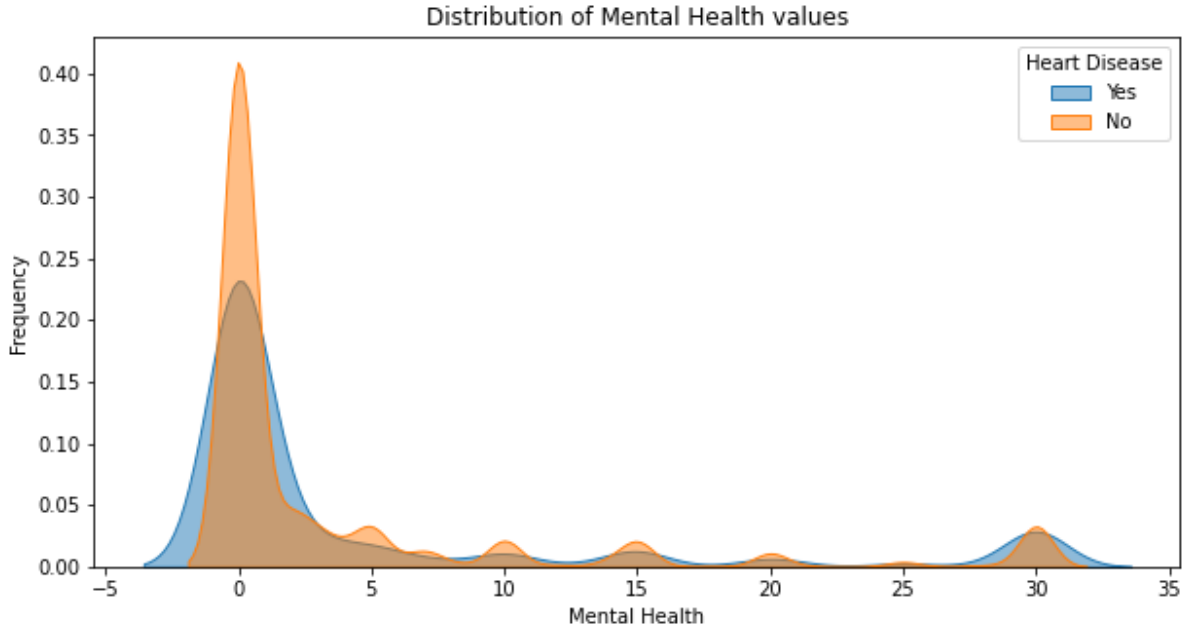
In [174]:
```
fig, axes = plt.subplots(figsize = (10,5))
sns.kdeplot(heart_df[heart_df["HeartDisease"]=='Yes']["SleepTime"], alpha=0.5,
shade = True, label="Yes", ax = axes)
sns.kdeplot(heart_df[heart_df["HeartDisease"]=='No']["SleepTime"], alpha=0.5,s
hade = True, label="No", ax = axes)
plt.title('Distribution of SleepTime values')
axes.set_xlabel("Sleep Time")
axes.set_ylabel("Frequency")
axes.legend(title='Heart Disease')
plt.show()
```



In [173]:
```
fig, axes = plt.subplots(figsize = (10,5))
sns.kdeplot(heart_df[heart_df["HeartDisease"]=='Yes']["PhysicalHealth"], alpha
=0.5,shade = True, label="Yes", ax = axes)
sns.kdeplot(heart_df[heart_df["HeartDisease"]=='No']["PhysicalHealth"], alpha=
0.5,shade = True, label="No", ax = axes)
plt.title('Distribution of Physical Health values')
axes.set_xlabel("Sleep Time")
axes.set_ylabel("Frequency")
axes.legend(title='Heart Disease')
plt.show()
```

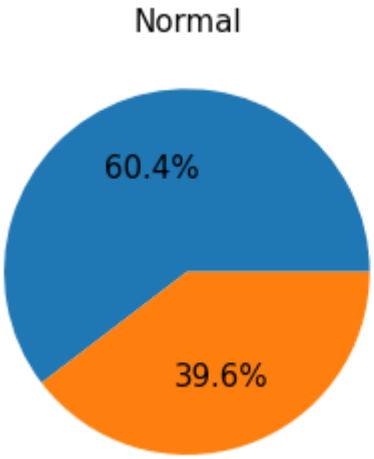


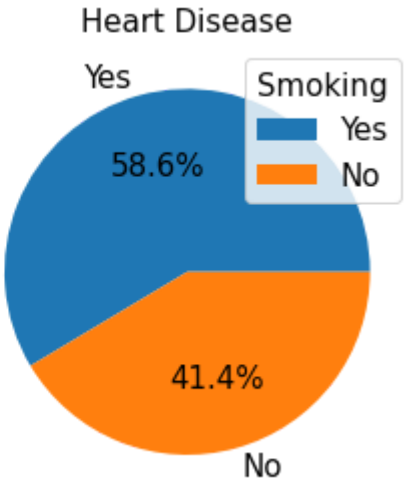
In [172]:
```
fig, axes = plt.subplots(figsize = (10,5))
sns.kdeplot(heart_df[heart_df["HeartDisease"]=='Yes']["MentalHealth"], alpha=
0.5,shade = True,  label="Yes", ax = axes)
sns.kdeplot(heart_df[heart_df["HeartDisease"]=='No']["MentalHealth"], alpha=0.
5,shade = True, label="No", ax = axes)
plt.title('Distribution of Mental Health values')
axes.set_xlabel("Mental Health")
axes.set_ylabel("Frequency")
axes.legend(title='Heart Disease')
plt.show()
```

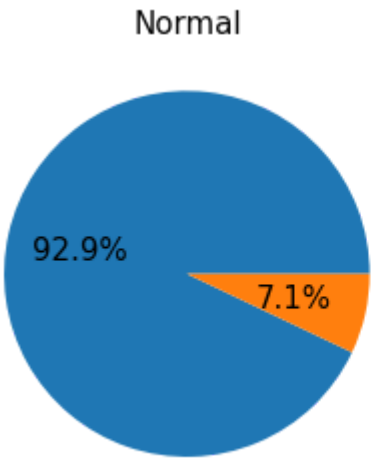


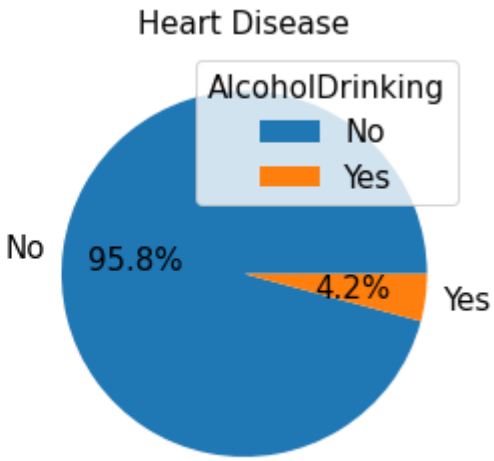
Distribution of Categorical Features based on Target Feature

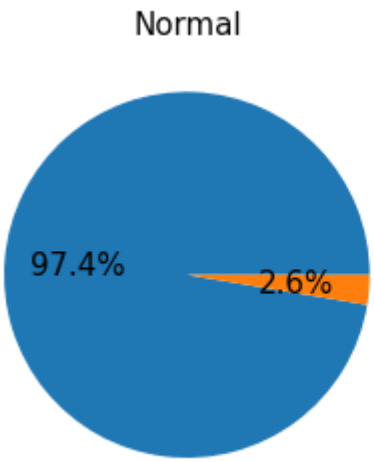
In [43]:
```
for feature in categorical_features:
  if (not(feature in ['Race', 'AgeCategory', 'Diabetic', 'HeartDisease'])):
    fig,axes = plt.subplots(1,2,figsize=(9,8))
    labels = heart_df[feature].unique()
    textprops = {"fontsize":15}

    axes[0].pie(heart_df[heart_df.HeartDisease=="No"][feature].value_counts(),
autopct='%1.1f%%',textprops =textprops)
    axes[0].set_title('Normal',fontsize=15)
    axes[1].pie(heart_df[heart_df.HeartDisease=="Yes"][feature].value_counts()
, labels = labels, autopct='%1.1f%%',textprops =textprops)
    axes[1].set_title('Heart Disease',fontsize=15)

    plt.legend(title = feature, fontsize=15, title_fontsize=15)
    plt.show()
```

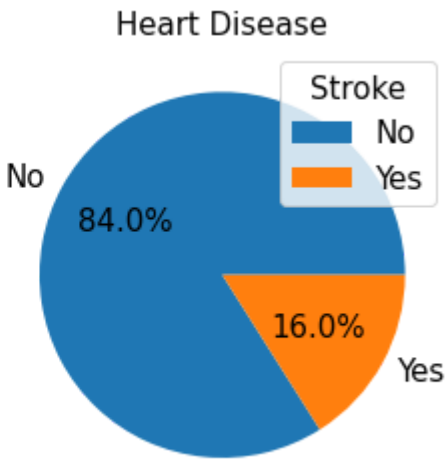

Normal
60.4%
39.6%

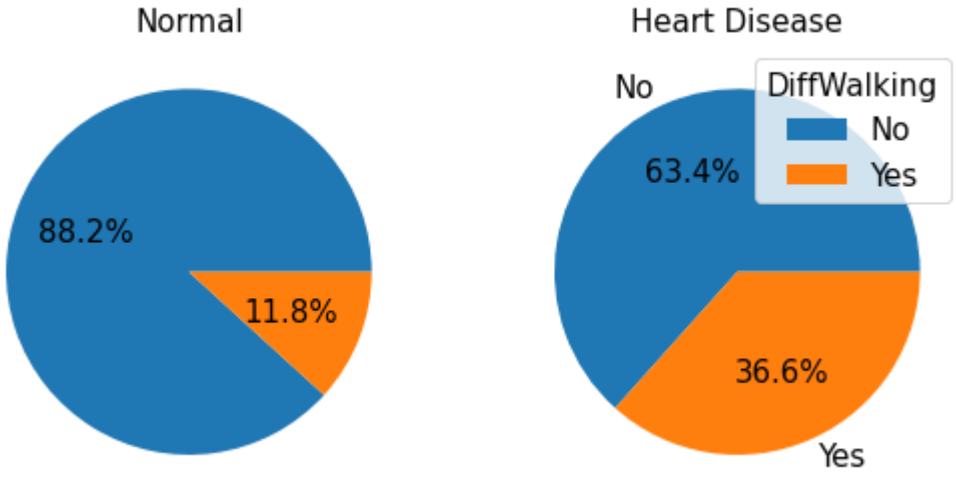

Heart Disease
Yes
Smoking
Yes
No
58.6%
41.4%
No


Normal
92.9%
7.1%

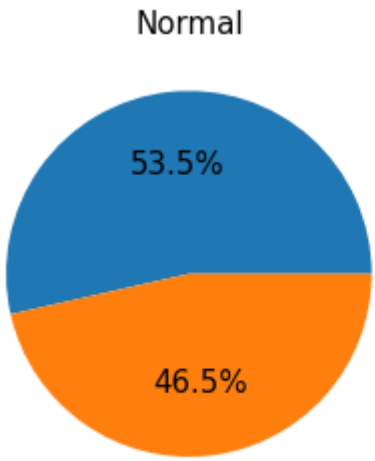

Heart Disease
AlcoholDrinking
No
Yes
No
95.8%
4.2%
Yes

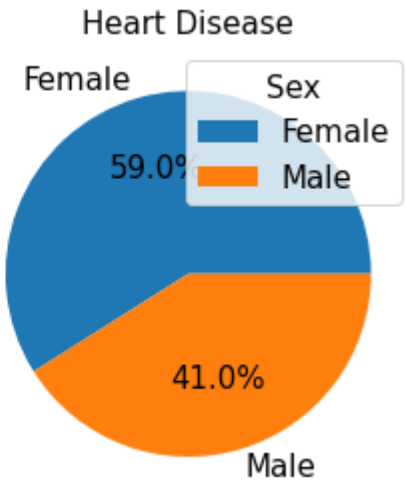

Normal
97.4%
2.6%

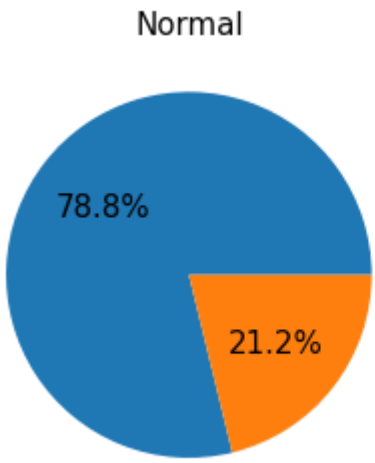

Heart Disease
Stroke
No
Yes
No
84.0%
16.0%
Yes

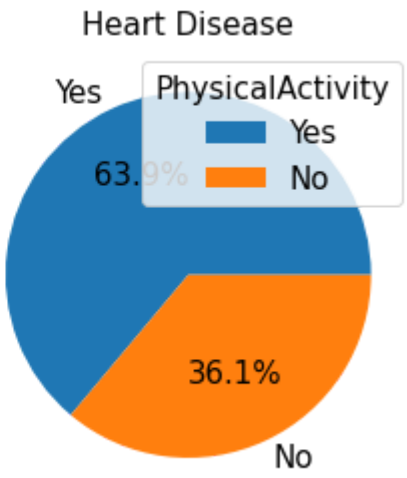

Normal
88.2%
11.8%

Heart Disease
No
63.4%
36.6%
Yes
DiffWalking
No
Yes

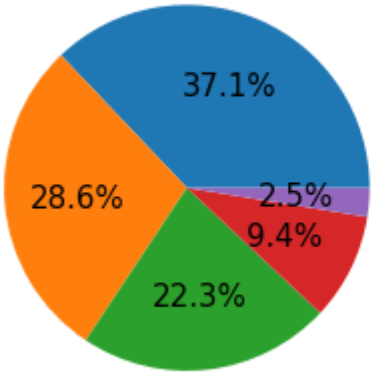

Normal
53.5%
46.5%


Heart Disease
Female
59.0%
41.0%
Male
Sex
Female
Male

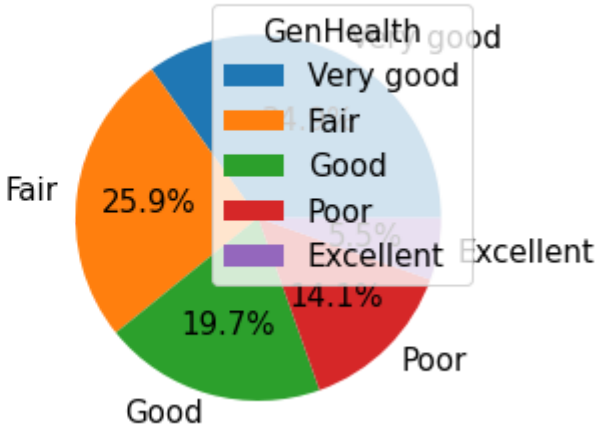

Normal
78.8%
21.2%


Heart Disease
Yes
36.1%
No
PhysicalActivity
Yes
No

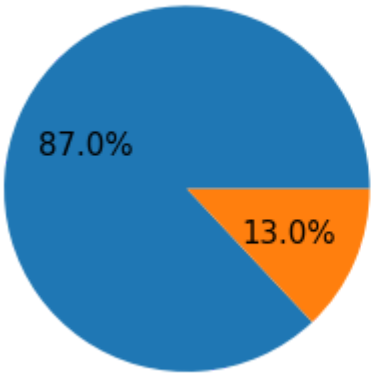
Normal



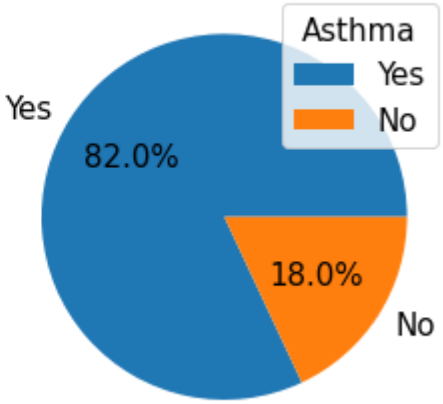
Heart Disease



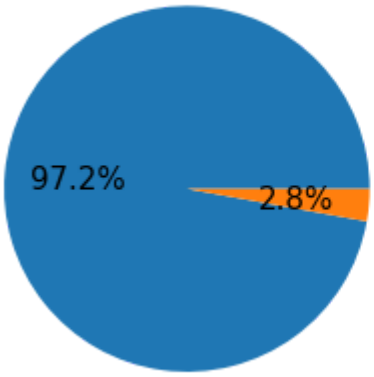
Normal



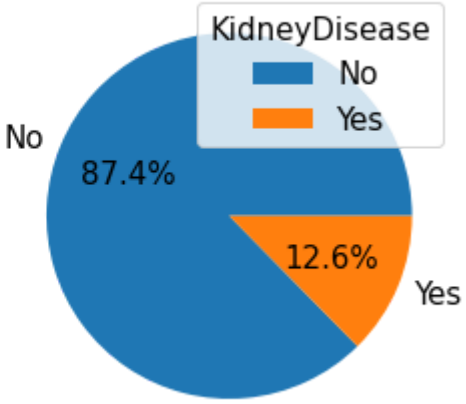
Heart Disease

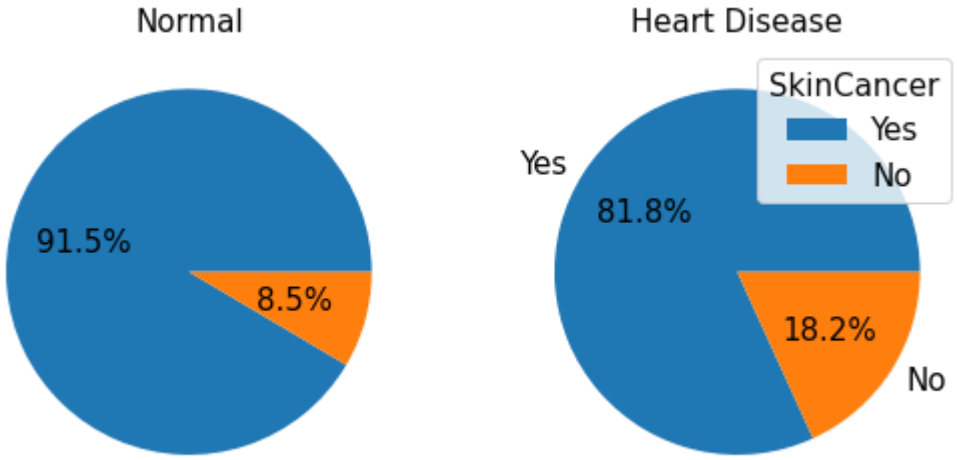


Normal

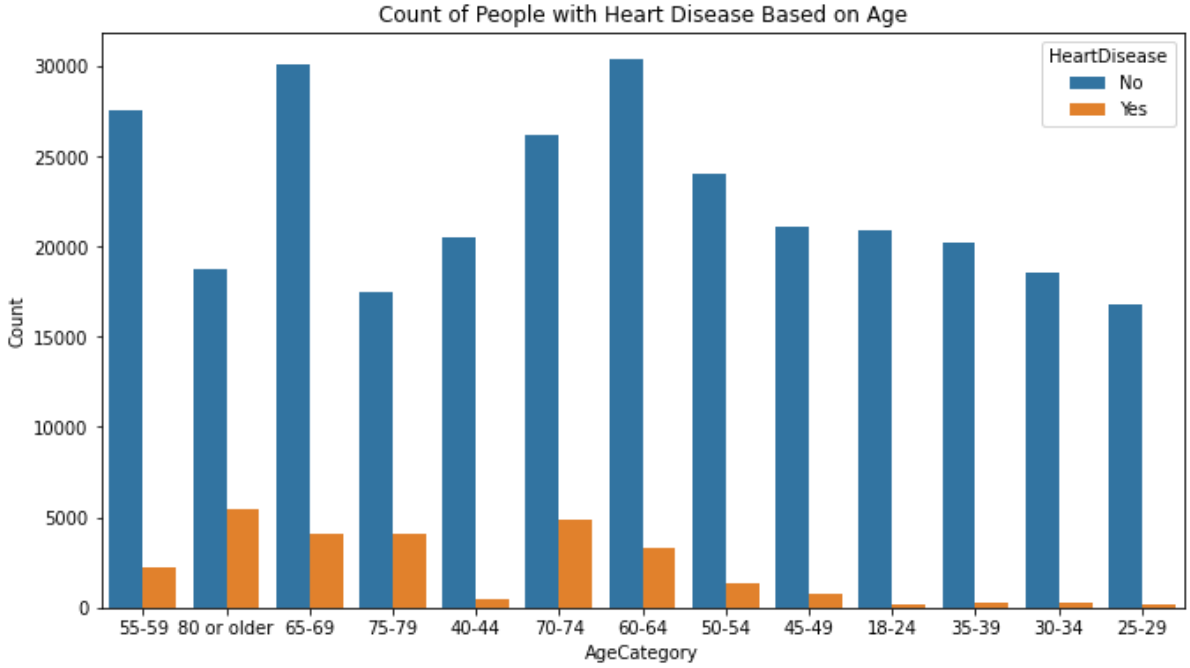


Heart Disease

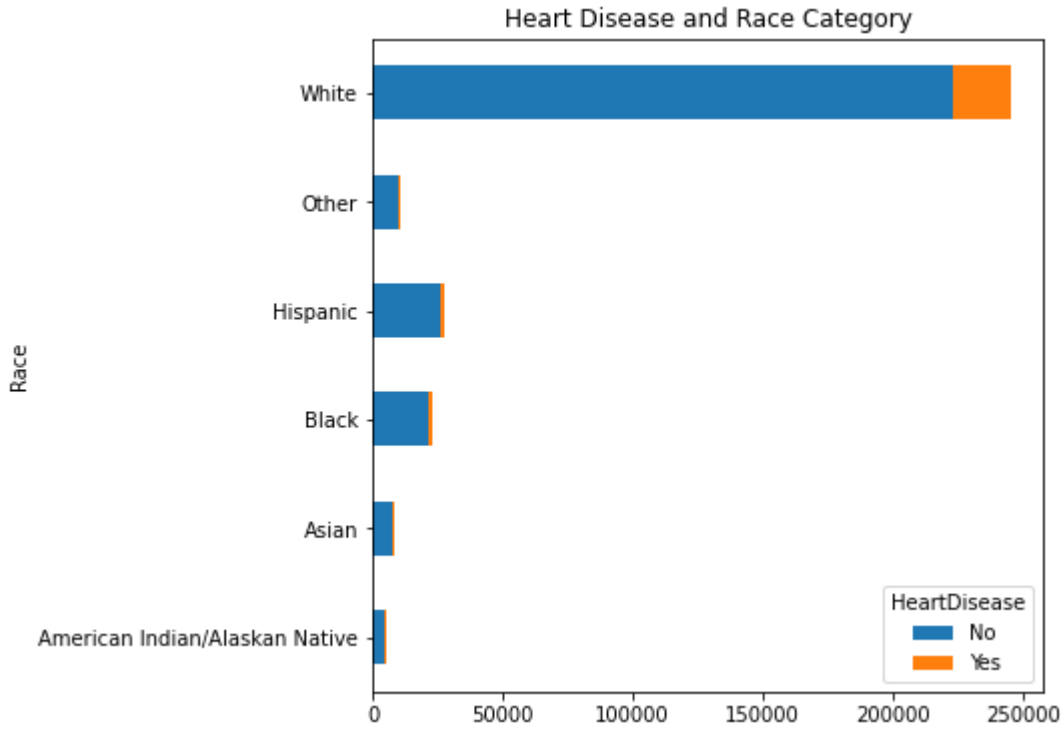




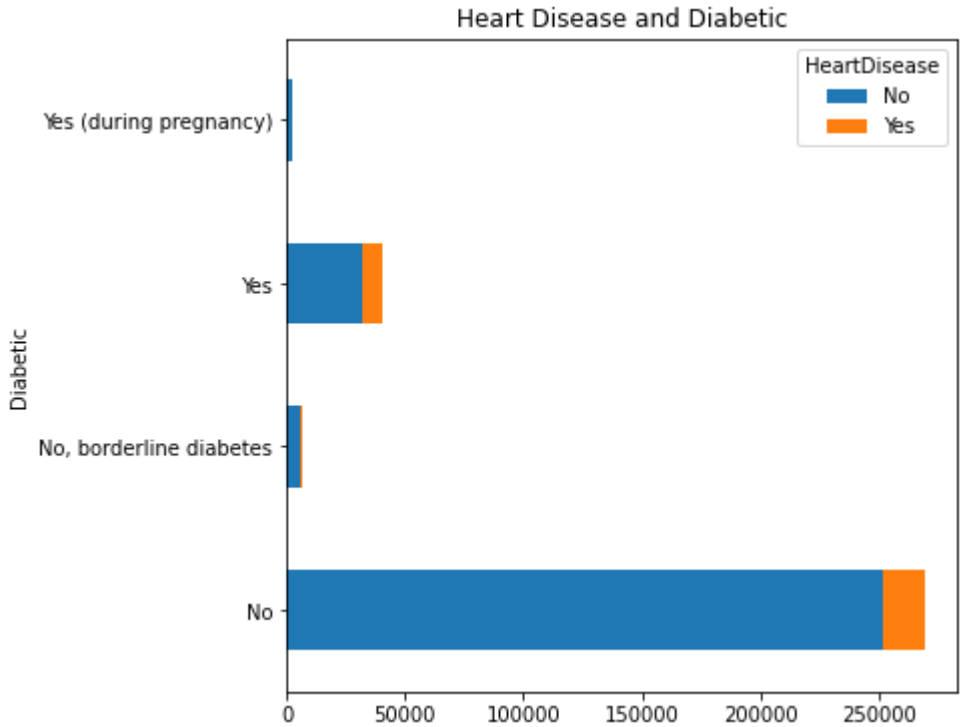
In [176]:
```
plt.figure(figsize = (11,6))
sns.countplot(x = heart_df['AgeCategory'], hue = 'HeartDisease', data = heart_
df)
plt.title("Count of People with Heart Disease Based on Age")
plt.ylabel('Count')
plt.show()
```



In [177]:
```
age_h=pd.DataFrame(pd.crosstab(heart_df["Race"],heart_df["HeartDisease"])).res
et_index()
ax=age_h.plot(x="Race",kind='barh', stacked=True, title='Heart Disease and Rac
e Category',figsize=(6,6))
```



In [178]:
```
age_h=pd.DataFrame(pd.crosstab(heart_df["Diabetic"],heart_df["HeartDisease"]))
.reset_index()
ax=age_h.plot(x="Diabetic",kind='barh', stacked=True, title='Heart Disease and
Diabetic',figsize=(6,6))
```



Aggregate Relationship

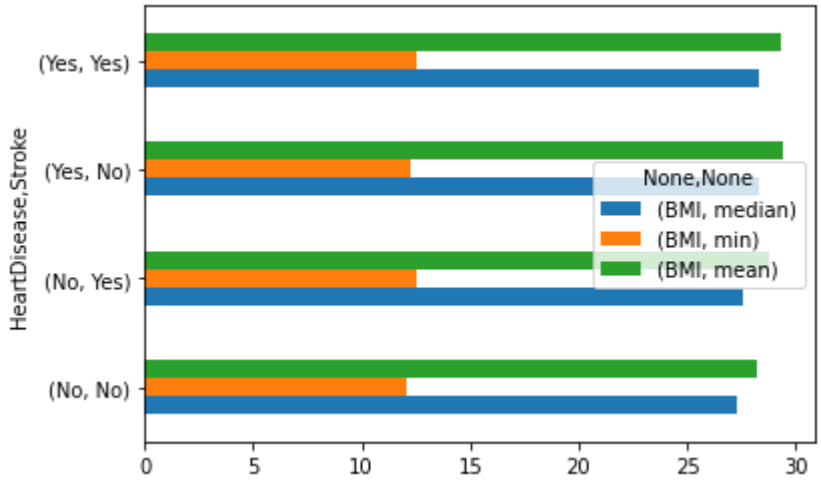
In [164]:
```
r = heart_df.groupby(['HeartDisease','Stroke'])[['BMI']].aggregate(['median',
'min','mean'])
r
```

Out[164]:

		BMI		
		median	min	mean
HeartDisease	Stroke			
No	No	27.25	12.02	28.210930
	Yes	27.60	12.53	28.733646
Yes	No	28.34	12.21	29.410951
	Yes	28.34	12.48	29.352581

In [165]:
```
r.plot(kind='barh')
```

Out[165]: <AxesSubplot:ylabel='HeartDisease,Stroke'>



from the above plot people with BMI value Higher than '28' has high probablity of getting a heart disease and stroke.

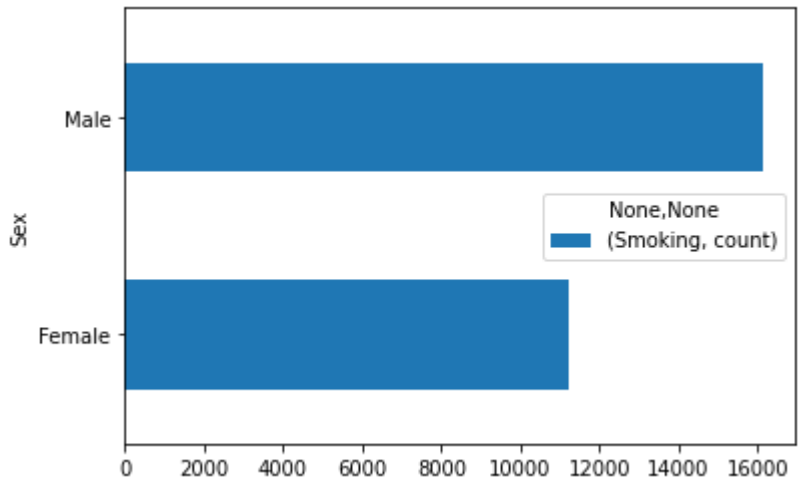
In [219]:
```
r1 = heart_df[heart_df["HeartDisease"]=='Yes'].groupby(['Sex'])[['Smoking']].aggregate(['count'])
r1
```

Out[219]:

Sex	Smoking count
Female	11234
Male	16139

In [221]:
```
r1.plot(kind='barh')
```

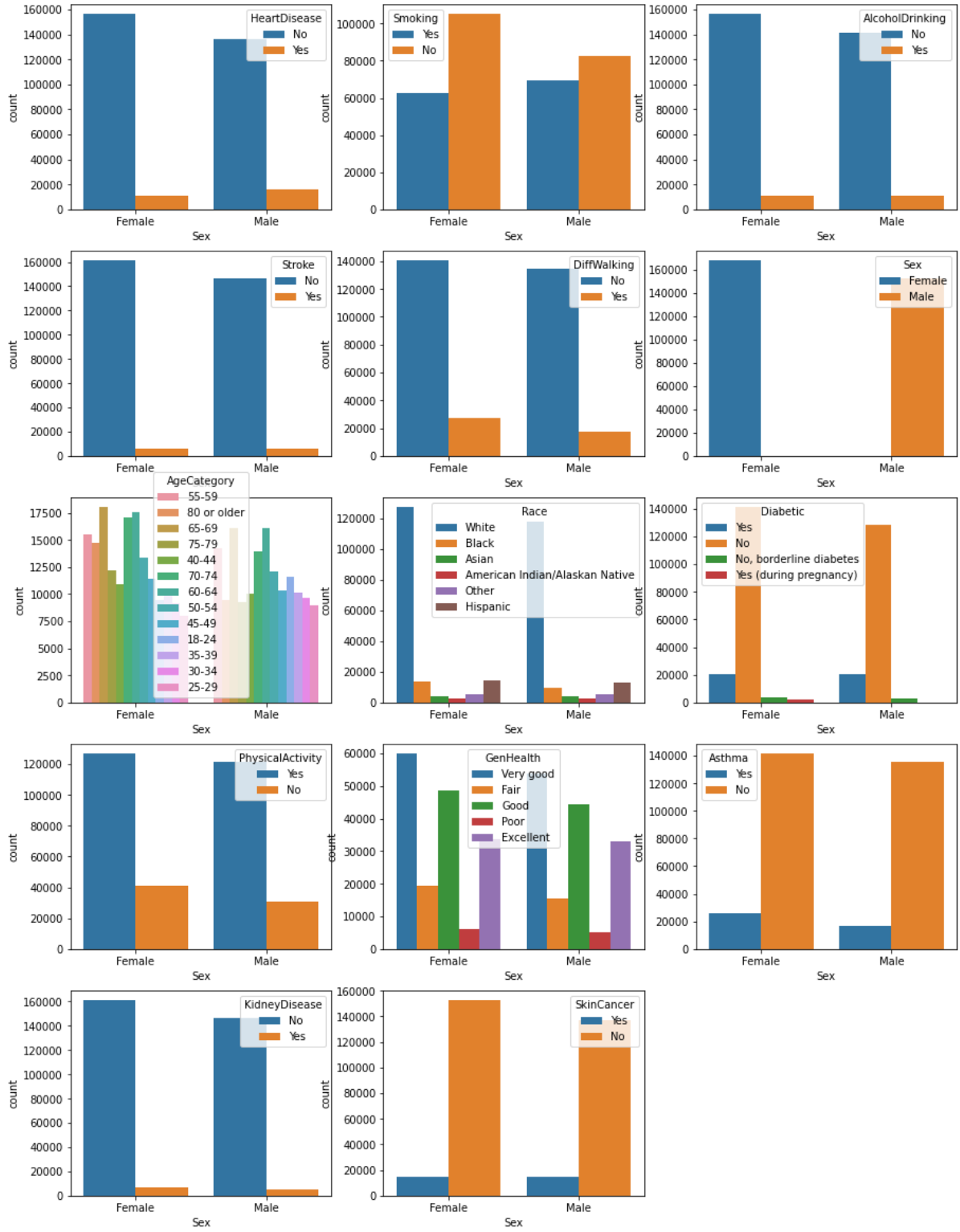
Out[221]: <AxesSubplot:ylabel='Sex'>



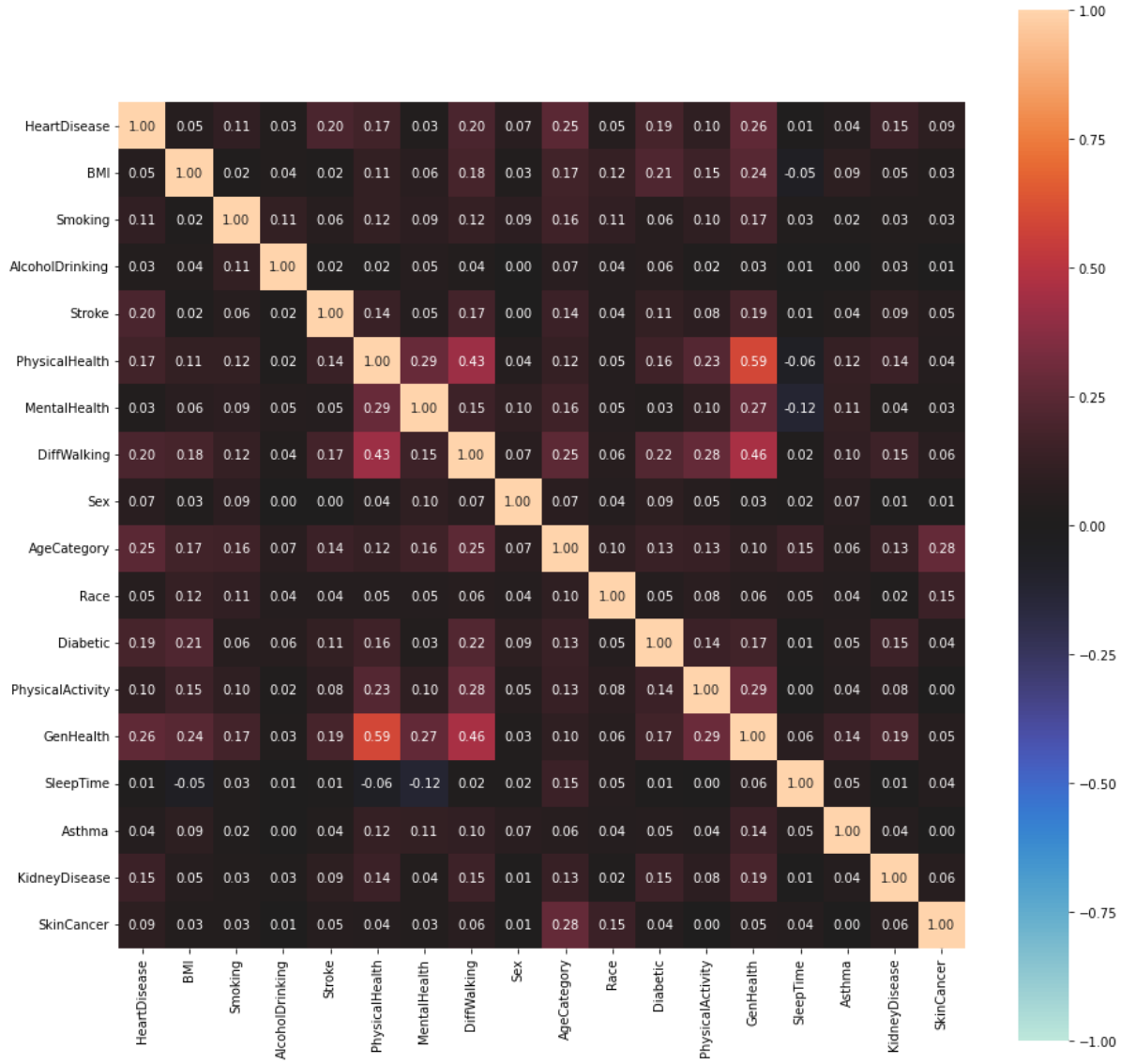
From the results male adults who have heart disease smoke more than female peers.

Analyzing the Distribution of Categorical variables depending on gender

In [21]:
```
size = 1
plt.figure(figsize = (15,25))
for feature in categorical_features:
    plt.subplot(6,3,size)
    sns.countplot(x = 'Sex',hue = heart_df[feature] , data = heart_df)
    size = size+1
```



In [225]:
```
associations(heart_df, figsize=(15,15))
```


HeartDisease 1.00 0.05 0.11 0.03 0.20 0.17 0.03 0.20 0.07 0.25 0.05 0.19 0.10 0.26 0.01 0.04 0.15 0.09
BMI 0.05 1.00 0.02 0.04 0.02 0.11 0.06 0.18 0.03 0.17 0.12 0.21 0.15 0.24 -0.05 0.09 0.05 0.03
Smoking 0.11 0.02 1.00 0.11 0.06 0.12 0.09 0.12 0.09 0.16 0.11 0.06 0.10 0.17 0.03 0.02 0.03 0.03
AlcoholDrinking 0.03 0.04 0.11 1.00 0.02 0.02 0.05 0.04 0.00 0.07 0.04 0.06 0.02 0.03 0.01 0.00 0.03 0.01
Stroke 0.20 0.02 0.06 0.02 1.00 0.14 0.05 0.17 0.00 0.14 0.04 0.11 0.08 0.19 0.01 0.04 0.09 0.05
PhysicalHealth 0.17 0.11 0.12 0.02 0.14 1.00 0.29 0.43 0.04 0.12 0.05 0.16 0.23 0.59 -0.06 0.12 0.14 0.04
MentalHealth 0.03 0.06 0.09 0.05 0.05 0.29 1.00 0.15 0.10 0.16 0.05 0.03 0.10 0.27 -0.12 0.11 0.04 0.03
DiffWalking 0.20 0.18 0.12 0.04 0.17 0.43 0.15 1.00 0.07 0.25 0.06 0.22 0.28 0.46 0.02 0.10 0.15 0.06
Sex 0.07 0.03 0.09 0.00 0.00 0.04 0.10 0.07 1.00 0.07 0.04 0.09 0.05 0.03 0.02 0.07 0.01 0.01
AgeCategory 0.25 0.17 0.16 0.07 0.14 0.12 0.16 0.25 0.07 1.00 0.10 0.13 0.13 0.10 0.15 0.06 0.13 0.28
Race 0.05 0.12 0.11 0.04 0.04 0.05 0.05 0.06 0.04 0.10 1.00 0.05 0.08 0.06 0.05 0.04 0.02 0.15
Diabetic 0.19 0.21 0.06 0.06 0.11 0.16 0.03 0.22 0.09 0.13 0.05 1.00 0.14 0.17 0.01 0.05 0.15 0.04
PhysicalActivity 0.10 0.15 0.10 0.02 0.08 0.23 0.10 0.28 0.05 0.13 0.08 0.14 1.00 0.29 0.00 0.04 0.08 0.00
GenHealth 0.26 0.24 0.17 0.03 0.19 0.59 0.27 0.46 0.03 0.10 0.06 0.17 0.29 1.00 0.06 0.14 0.19 0.05
SleepTime 0.01 -0.05 0.03 0.01 0.01 -0.06 -0.12 0.02 0.02 0.15 0.05 0.01 0.00 0.06 1.00 0.05 0.01 0.04
Asthma 0.04 0.09 0.02 0.00 0.04 0.12 0.11 0.10 0.07 0.06 0.04 0.05 0.04 0.14 0.05 1.00 0.04 0.00
KidneyDisease 0.15 0.05 0.03 0.03 0.09 0.14 0.04 0.15 0.01 0.13 0.02 0.15 0.08 0.19 0.01 0.04 1.00 0.06
SkinCancer 0.09 0.03 0.03 0.01 0.05 0.04 0.03 0.06 0.01 0.28 0.15 0.04 0.00 0.05 0.04 0.00 0.06 1.00
HeartDisease
BMI
Smoking
AlcoholDrinking
Stroke
PhysicalHealth
MentalHealth
DiffWalking
Sex
AgeCategory
Race
Diabetic
PhysicalActivity
GenHealth
SleepTime
Asthma
KidneyDisease
SkinCancer
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00

Out[225]: {'ax': <AxesSubplot:>,
'corr':

	HeartDisease	BMI	Smoking	AlcoholDrinking	Stroke
HeartDisease	1.000000	0.051803	0.107738	0.032009	0.196798
BMI	0.051803	1.000000	0.023118	0.038816	0.019733
Smoking	0.107738	0.023118	1.000000	0.111741	0.061184
AlcoholDrinking	0.032009	0.038816	0.111741	1.000000	0.019746
Stroke	0.196798	0.019733	0.061184	0.019746	1.000000
PhysicalHealth	0.170721	0.109788	0.115352	0.017254	0.137014
MentalHealth	0.028591	0.064131	0.085157	0.051282	0.046467
DiffWalking	0.201234	0.181678	0.120052	0.035265	0.174111
Sex	0.070007	0.026940	0.085028	0.003796	0.002515
AgeCategory	0.245588	0.170749	0.164059	0.069702	0.143594
Race	0.051230	0.118244	0.108177	0.040267	0.040646
Diabetic	0.185101	0.212988	0.059369	0.059576	0.109737
PhysicalActivity	0.100001	0.150616	0.097150	0.017382	0.079416
GenHealth	0.259519	0.237771	0.174799	0.030491	0.185981
SleepTime	0.008327	-0.051822	0.030336	0.005065	0.011900
Asthma	0.041390	0.092345	0.024074	0.001282	0.038802
KidneyDisease	0.145157	0.050768	0.034858	0.028192	0.091106
SkinCancer	0.093281	0.033644	0.033920	0.005399	0.048055

	PhysicalHealth	MentalHealth	DiffWalking	Sex	\
HeartDisease	0.170721	0.028591	0.201234	0.070007	
BMI	0.109788	0.064131	0.181678	0.026940	
Smoking	0.115352	0.085157	0.120052	0.085028	
AlcoholDrinking	0.017254	0.051282	0.035265	0.003796	
Stroke	0.137014	0.046467	0.174111	0.002515	
PhysicalHealth	1.000000	0.287987	0.428373	0.040904	
MentalHealth	0.287987	1.000000	0.152235	0.100058	
DiffWalking	0.428373	0.152235	1.000000	0.068828	
Sex	0.040904	0.100058	0.068828	1.000000	
AgeCategory	0.118165	0.156797	0.251497	0.074206	
Race	0.046159	0.046583	0.061017	0.040316	
Diabetic	0.161181	0.034621	0.221032	0.087368	
PhysicalActivity	0.232283	0.095808	0.278508	0.048207	
GenHealth	0.588780	0.266917	0.457933	0.030617	
SleepTime	-0.061387	-0.119717	0.022216	0.015704	
Asthma	0.117907	0.114008	0.103194	0.069159	

```
 KidneyDisease           0.142197      0.037281     0.153030  0.008893
 SkinCancer              0.041700      0.033412     0.064801  0.013306

                   AgeCategory      Race  Diabetic  PhysicalActivity  \
 HeartDisease         0.245588  0.051230  0.185101          0.100001
 BMI                  0.170749  0.118244  0.212988          0.150616
 Smoking              0.164059  0.108177  0.059369          0.097150
 AlcoholDrinking      0.069702  0.040267  0.059576          0.017382
 Stroke               0.143594  0.040646  0.109737          0.079416
 PhysicalHealth       0.118165  0.046159  0.161181          0.232283
 MentalHealth         0.156797  0.046583  0.034621          0.095808
 DiffWalking          0.251497  0.061017  0.221032          0.278508
 Sex                  0.074206  0.040316  0.087368          0.048207
 AgeCategory          1.000000  0.097284  0.134229          0.130170
 Race                 0.097284  1.000000  0.045334          0.078090
 Diabetic             0.134229  0.045334  1.000000          0.142829
 PhysicalActivity     0.130170  0.078090  0.142829          1.000000
 GenHealth            0.096581  0.059982  0.167802          0.293548
 SleepTime            0.147311  0.047946  0.014871          0.003849
 Asthma               0.060409  0.042394  0.050767          0.041477
 KidneyDisease        0.127348  0.023219  0.154846          0.081788
 SkinCancer           0.280929  0.146445  0.039383          0.000000

                   GenHealth  SleepTime    Asthma  KidneyDisease  SkinCancer
 HeartDisease       0.259519   0.008327  0.041390       0.145157    0.093281
 BMI                0.237771  -0.051822  0.092345       0.050768    0.033644
 Smoking            0.174799   0.030336  0.024074       0.034858    0.033920
 AlcoholDrinking    0.030491   0.005065  0.001282       0.028192    0.005399
 Stroke             0.185981   0.011900  0.038802       0.091106    0.048055
 PhysicalHealth     0.588780  -0.061387  0.117907       0.142197    0.041700
 MentalHealth       0.266917  -0.119717  0.114008       0.037281    0.033412
 DiffWalking        0.457933   0.022216  0.103194       0.153030    0.064801
 Sex                0.030617   0.015704  0.069159       0.008893    0.013306
 AgeCategory        0.096581   0.147311  0.060409       0.127348    0.280929
 Race               0.059982   0.047946  0.042394       0.023219    0.146445
 Diabetic           0.167802   0.014871  0.050767       0.154846    0.039383
 PhysicalActivity   0.293548   0.003849  0.041477       0.081788    0.000000
 GenHealth          1.000000   0.064153  0.141151       0.192892    0.053217
 SleepTime          0.064153   1.000000  0.048245       0.006238    0.041266
 Asthma             0.141151   0.048245  1.000000       0.039643    0.000000
 KidneyDisease      0.192892   0.006238  0.039643       1.000000    0.061762
 SkinCancer         0.053217   0.041266  0.000000       0.061762    1.000000
}
```

Summary

My conclusions after performing basic data analysis on dataset for predicting responsible key-features for having "Heart Disease" are:

The adults whose age is greater than or euqal to 80 have higher chances of getting a heart disease. In overall Dataset, most people who are diagnosed with heart disease are smokers and in that, percentage of male adults is high. White and Black people seem to have higher chance of getting heart disease. I did not see any relationship between heart disease and people who are Heavy drinkers/ asthma patients. Diabetic adults seem to have more chances of getting heart disease. However, the dataset is highly unbalanced and because of this some conculsions/plots needed to be further investiged and I'm hoping that by applying sampling techinques on dataset we can achieve noticable relationships between some features.