heartdataanalysis

February 8, 2024

1 DATA CLEANING

[3]:
```
#Package import
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import plotly.express as px
import seaborn as sns
import plotly.express as px
```

[4]:
```
#data import
heart_data = pd.read_csv(r"C:\New folder\OneDrive - Conestoga␣
 ↪College\Desktop\heart data.csv")
#heart_data = pd.read_csv(r"heart data .csv")
print(heart_data)
```

```
       HeartDisease    BMI Smoking AlcoholDrinking Stroke  PhysicalHealth  \
0                No  16.60     Yes              No     No               3
1                No  20.34      No              No    Yes               0
2                No  26.58     Yes              No     No              20
3                No  24.21      No              No     No               0
4                No  23.71      No              No     No              28
…               …    …      …               …    …              …
319790          Yes  27.41     Yes              No     No               7
319791           No  29.84     Yes              No     No               0
319792           No  24.24      No              No     No               0
319793           No  32.81      No              No     No               0
319794           No  46.56      No              No     No               0

        MentalHealth DiffWalking     Sex  AgeCategory      Race Diabetic  \
0                 30          No  Female        55-59     White      Yes
1                  0          No  Female  80 or older     White       No
2                 30          No    Male        65-69     White      Yes
3                  0          No  Female        75-79     White       No
4                  0         Yes  Female        40-44     White       No
…               …          …      …          …       …      …
319790             0         Yes    Male        60-64  Hispanic      Yes
319791             0          No    Male        35-39  Hispanic       No
```

```
319792             0          No  Female        45-49  Hispanic       No
319793             0          No  Female        25-29  Hispanic       No
319794             0          No  Female  80 or older  Hispanic       No

       PhysicalActivity  GenHealth  SleepTime Asthma KidneyDisease SkinCancer
0                   Yes  Very good          5    Yes            No        Yes
1                   Yes  Very good          7     No            No         No
2                   Yes       Fair          8    Yes            No         No
3                    No       Good          6     No            No        Yes
4                   Yes  Very good          8     No            No         No
…                   …          …        …     …             …         …
319790               No       Fair          6    Yes            No         No
319791              Yes  Very good          5    Yes            No         No
319792              Yes       Good          6     No            No         No
319793               No       Good         12     No            No         No
319794              Yes       Good          8     No            No         No

[319795 rows x 18 columns]
```

2 DATA EXPLORATION

[6]:
```
heart_data.shape
```

[6]:
```
(319795, 18)
```

[7]:
```
#data description
heart_data.describe()
```

[7]:
```
                BMI  PhysicalHealth   MentalHealth      SleepTime
count  319795.000000    319795.00000  319795.000000  319795.000000
mean       28.325399         3.37171       3.898366       7.097075
std         6.356100         7.95085       7.955235       1.436007
min        12.020000         0.00000       0.000000       1.000000
25%        24.030000         0.00000       0.000000       6.000000
50%        27.340000         0.00000       0.000000       7.000000
75%        31.420000         2.00000       3.000000       8.000000
max        94.850000        30.00000      30.000000      24.000000
```

[8]:
```
#column name space removal
heart_data.columns = heart_data.columns.str.strip()
```

[9]:
```
#missing values check
heart_data.isna().sum()
```

[9]:
```
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
```

[10]:
```
#missing value check percentage
print(heart_data.isnull().sum()/heart_data.shape[0] * 100)
```

```
HeartDisease        0.0
BMI                 0.0
Smoking             0.0
AlcoholDrinking     0.0
Stroke              0.0
PhysicalHealth      0.0
MentalHealth        0.0
DiffWalking         0.0
Sex                 0.0
AgeCategory         0.0
Race                0.0
Diabetic            0.0
PhysicalActivity    0.0
GenHealth           0.0
SleepTime           0.0
Asthma              0.0
KidneyDisease       0.0
SkinCancer          0.0
dtype: float64
```

[11]:
```
#data columns
print(heart_data.columns)
```

```
Index(['HeartDisease', 'BMI', 'Smoking', 'AlcoholDrinking', 'Stroke',
       'PhysicalHealth', 'MentalHealth', 'DiffWalking', 'Sex', 'AgeCategory',
       'Race', 'Diabetic', 'PhysicalActivity', 'GenHealth', 'SleepTime',
       'Asthma', 'KidneyDisease', 'SkinCancer'],
      dtype='object')
```

[12]:
```
#checking duplicate values
heart_data.duplicated().sum()
```

[12]:
```
18078
```

[13]:
```
#dropping duplicates
heart_data.drop_duplicates(inplace=True)
```

[14]:
```
heart_data.shape
```

[14]:
```
(301717, 18)
```

[15]:
```
heart_data.head()
```

[15]:

	HeartDisease	BMI	Smoking	AlcoholDrinking	Stroke	PhysicalHealth	MentalHealth	DiffWalking	Sex	AgeCategory	Race	Diabetic	PhysicalActivity	GenHealth	SleepTime	Asthma	KidneyDisease	SkinCancer
0	No	16.60	Yes	No	No	3	30	No	Female	55-59	White	Yes	Yes	Very good	5	Yes	No	Yes
1	No	20.34	No	No	Yes	0	0	No	Female	80 or older	White	No	Yes	Very good	7	No	No	No
2	No	26.58	Yes	No	No	20	30	No	Male	65-69	White	Yes	Yes	Fair	8	Yes	No	No
3	No	24.21	No	No	No	0	0	No	Female	75-79	White	No	No	Good	6	No	No	Yes
4	No	23.71	No	No	No	28	0	Yes	Female	40-44	White	No	Yes	Very good	8	No	No	No

[16]:
```
#dropping index
heart_data = heart_data.reset_index(drop=True)
heart_data.head()
```

[16]:

	HeartDisease	BMI	Smoking	AlcoholDrinking	Stroke	PhysicalHealth	MentalHealth	DiffWalking	Sex	AgeCategory	Race	Diabetic
0	No	16.60	Yes	No	No	3	30	No	Female	55-59	White	Yes
1	No	20.34	No	No	Yes	0	0	No	Female	80 or older	White	No
2	No	26.58	Yes	No	No	20	30	No	Male	65-69	White	Yes
3	No	24.21	No	No	No	0	0	No	Female	75-79	White	No
4	No	23.71	No	No	No	28	0	Yes	Female	40-44	White	No

```
0            30          No  Female        55-59  White      Yes
1             0          No  Female  80 or older  White       No
2            30          No    Male        65-69  White      Yes
3             0          No  Female        75-79  White       No
4             0         Yes  Female        40-44  White       No

  PhysicalActivity  GenHealth  SleepTime Asthma KidneyDisease SkinCancer
0              Yes  Very good          5    Yes            No        Yes
1              Yes  Very good          7     No            No         No
2              Yes       Fair          8    Yes            No         No
3               No       Good          6     No            No        Yes
4              Yes  Very good          8     No            No         No
```

[39]:
```
#data description
heart_data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 301717 entries, 0 to 301716
Data columns (total 18 columns):
 #   Column            Non-Null Count   Dtype
---  ------            --------------   -----
 0   HeartDisease      301717 non-null  object
 1   BMI               301717 non-null  float64
 2   Smoking           301717 non-null  object
 3   AlcoholDrinking   301717 non-null  object
 4   Stroke            301717 non-null  object
 5   PhysicalHealth    301717 non-null  int64
 6   MentalHealth      301717 non-null  int64
 7   DiffWalking       301717 non-null  object
 8   Sex               301717 non-null  object
 9   AgeCategory       301717 non-null  object
 10  Race              301717 non-null  object
 11  Diabetic          301717 non-null  object
 12  PhysicalActivity  301717 non-null  object
 13  GenHealth         301717 non-null  object
 14  SleepTime         301717 non-null  int64
 15  Asthma            301717 non-null  object
 16  KidneyDisease     301717 non-null  object
 17  SkinCancer        301717 non-null  object
dtypes: float64(1), int64(3), object(14)
memory usage: 41.4+ MB
```

[17]:
```
#categorical and numerical features
col = list(heart_data.columns)
categorical_features = []
numerical_features = []
for i in col:
```

```
    if len(heart_data[i].unique()) > 6:
        numerical_features.append(i)
    else:
        categorical_features.append(i)

print('Categorical Features :',*categorical_features)
print('Numerical Features :',*numerical_features)
```

```
Categorical Features : HeartDisease Smoking AlcoholDrinking Stroke DiffWalking
Sex Race Diabetic PhysicalActivity GenHealth Asthma KidneyDisease SkinCancer
Numerical Features : BMI PhysicalHealth MentalHealth AgeCategory SleepTime
```

[5]:
```
#Describe the numerical columns
heart_data.describe()
```

[5]:

	BMI	PhysicalHealth	MentalHealth	SleepTime
count	319795.000000	319795.00000	319795.000000	319795.000000
mean	28.325399	3.37171	3.898366	7.097075
std	6.356100	7.95085	7.955235	1.436007
min	12.020000	0.00000	0.000000	1.000000
25%	24.030000	0.00000	0.000000	6.000000
50%	27.340000	0.00000	0.000000	7.000000
75%	31.420000	2.00000	3.000000	8.000000
max	94.850000	30.00000	30.000000	24.000000

[7]:
```
print(f"Number of unique values in BMI :{len(heart_data.BMI.value_counts())}")
print(f"Number of unique values in Physical Health :{len(heart_data.
 ↪PhysicalHealth.value_counts())}")
print(f"Number of unique values in Mental Health :{len(heart_data.MentalHealth.
 ↪value_counts())}")
print(f"Number of unique values in Sleep Time :{len(heart_data.SleepTime.
 ↪value_counts())}")
```

```
Number of unique values in BMI :3604
Number of unique values in Physical Health :31
Number of unique values in Mental Health :31
Number of unique values in Sleep Time :24
```

[20]:
```
#Checking categorical data description
heart_data.describe(include='object')
```

[20]:

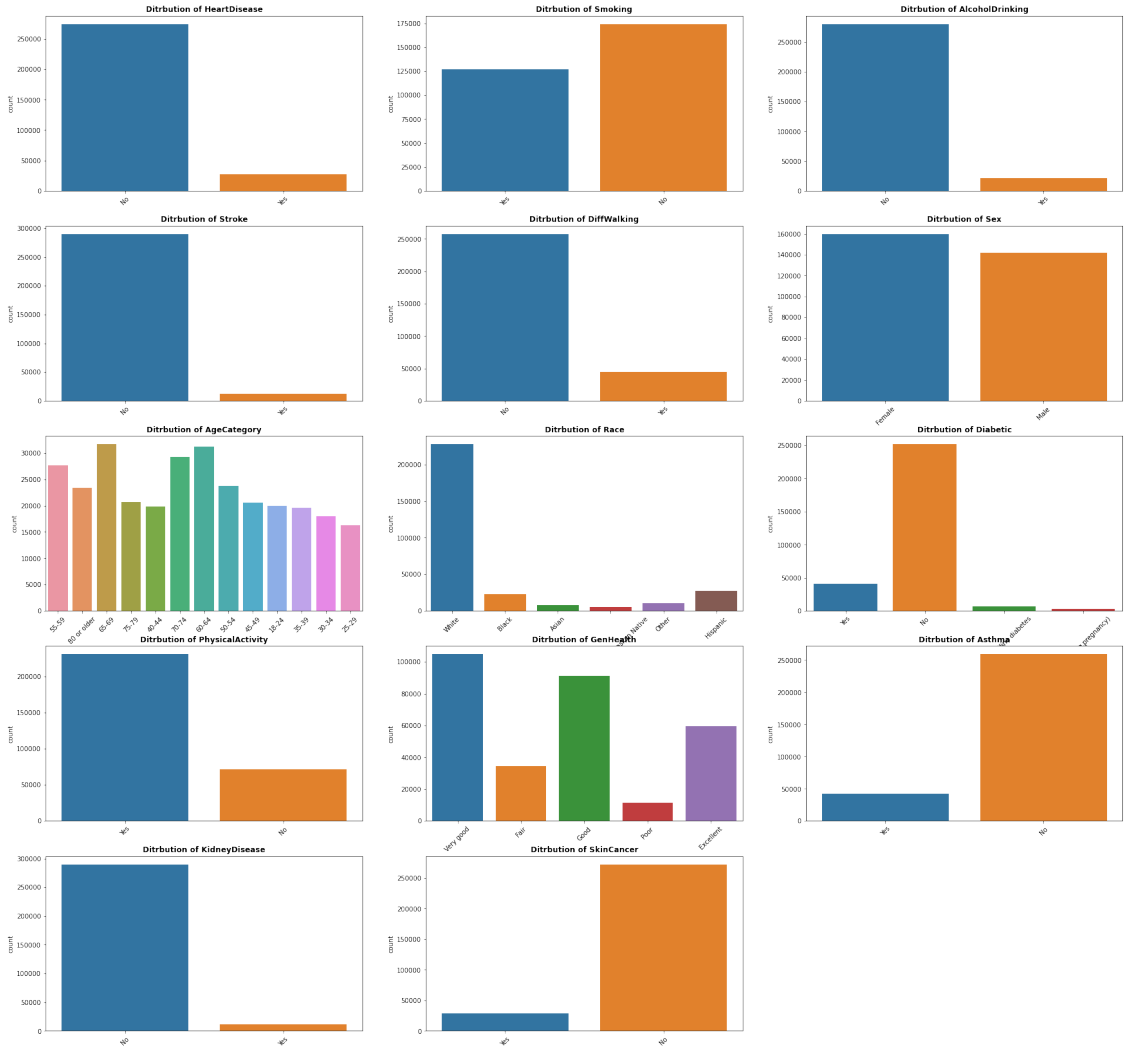
	HeartDisease	Smoking	AlcoholDrinking	Stroke	DiffWalking	Sex	\
count	301717	301717	301717	301717	301717	301717	
unique	2	2	2	2	2	2	
top	No	No	No	No	No	Female	
freq	274456	174312	280136	289653	257362	159671	

```
       AgeCategory    Race Diabetic PhysicalActivity  GenHealth  Asthma  \
count       301717  301717   301717           301717     301717  301717
unique          13       6        4                2          5       2
top          65-69   White       No              Yes  Very good      No
freq         31670  227724   251796           230412     104796  259066

       KidneyDisease SkinCancer
count         301717     301717
unique             2          2
top               No         No
freq          289941     272425
```

3 Visualization and Analysis

4 Univariate Analysis

[52]:
```
#distribution of data (Column wise analysis)
i=1
plt.figure(figsize=(30,35))
for col in heart_data.select_dtypes(include='object').columns:
    plt.subplot(6,3,i)
    plt.xticks(rotation=45)
    sns.countplot(x=col,data=heart_data)
    plt.title(f"Ditrbution of {col}",weight='bold')
    plt.xlabel('')
    i+=1
```



5 Bivariate Analysis

6 Who is more inclined towards getting heart disease - male or female?

[21]:
```
heart_data.groupby(['Sex'])['HeartDisease'].count().reset_index(name ='count')
```

[21]:

	Sex	count
0	Female	159671
1	Male	142046

[22]:
```
Heart_diseases_yes = heart_data.loc[heart_data['HeartDisease'] == 'Yes', :]
Heart_diseases_yes.head()
```

[22]:
```
   HeartDisease    BMI Smoking AlcoholDrinking Stroke  PhysicalHealth  \
5           Yes  28.87     Yes              No     No               6
10          Yes  34.30     Yes              No     No              30
35          Yes  32.98     Yes              No    Yes              10
42          Yes  25.06      No              No     No               0
43          Yes  30.23     Yes              No     No               6

    MentalHealth DiffWalking     Sex  AgeCategory   Race Diabetic  \
5              0         Yes  Female        75-79  Black       No
10             0         Yes    Male        60-64  White      Yes
35             0         Yes    Male        75-79  White      Yes
42             0         Yes  Female  80 or older  White      Yes
43             2         Yes  Female        75-79  White      Yes

   PhysicalActivity GenHealth  SleepTime Asthma KidneyDisease SkinCancer
5                No      Fair         12     No            No         No
10               No      Poor         15    Yes            No         No
35              Yes      Poor          4     No            No        Yes
42               No      Good          7     No            No        Yes
43              Yes      Fair          8     No           Yes         No
```

[10]:
```
fig = px.histogram(heart_data, x='Sex', color='HeartDisease',
                   template='plotly_dark', barmode='group',
                   color_discrete_sequence=['#71AEC2', '#D58989'])
fig.update_layout(title='Heart Disease Frequency Gender Wise',
                  xaxis_title='Gender',
                  yaxis_title='Frequency',
                  legend_title='Heart Disease')
fig.show()
```

7 Age distribution vs heartdisease

[11]:
```
fig = px.histogram(heart_data, x='AgeCategory',␣
 ↪category_orders=dict(AgeCategory=["18-24", "25-29", "30-34", "35-39","40-44",
                                                                        ␣
 ↪"45-49","50-54","55-59","60-64","65-69","70-74",
                                                                       ␣
 ↪"75-79","80 or older"]), color='HeartDisease',
                   template='plotly_dark', barmode='group',
                   color_discrete_sequence=['#71AEC2', '#D58989'])
fig.update_layout(title='Heart Disease Frequency for AgeCategory',
                  xaxis_title='AgeCategory',
                  yaxis_title='Frequency',
                  legend_title='Heart Disease')
fig.show()
```

8 Is BMI is one of the factor for heart disease?

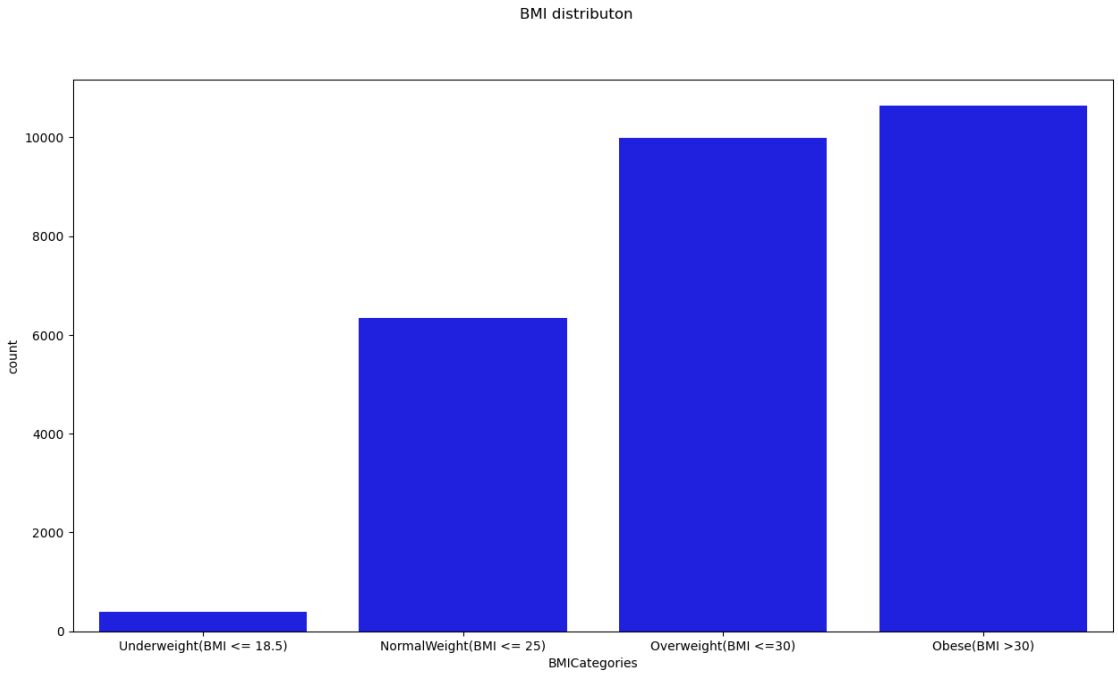
[13]:
```
bins = [0,18.5,25,30, np.inf]
Categories = ['Underweight(BMI <= 18.5)','NormalWeight(BMI <=␣
 ↪25)','Overweight(BMI <=30)', 'Obese(BMI >30)']
heart_data['BMICategories'] = pd.cut(heart_data['BMI'], bins, labels =␣
 ↪Categories)
print(heart_data['BMICategories'])
```

```
0         Underweight(BMI <= 18.5)
1          NormalWeight(BMI <= 25)
2             Overweight(BMI <=30)
3          NormalWeight(BMI <= 25)
4          NormalWeight(BMI <= 25)
                    …
319790        Overweight(BMI <=30)
319791        Overweight(BMI <=30)
319792     NormalWeight(BMI <= 25)
319793               Obese(BMI >30)
319794               Obese(BMI >30)
Name: BMICategories, Length: 319795, dtype: category
Categories (4, object): ['Underweight(BMI <= 18.5)' < 'NormalWeight(BMI <= 25)'
< 'Overweight(BMI <=30)' < 'Obese(BMI >30)']
```

[14]:
```
heart_disease = heart_data.loc[heart_data['HeartDisease']== 'Yes',:]
```

[15]:
```
BMI = heart_disease.groupby('BMICategories')['HeartDisease'].count().
 ↪reset_index(name = 'count')
```

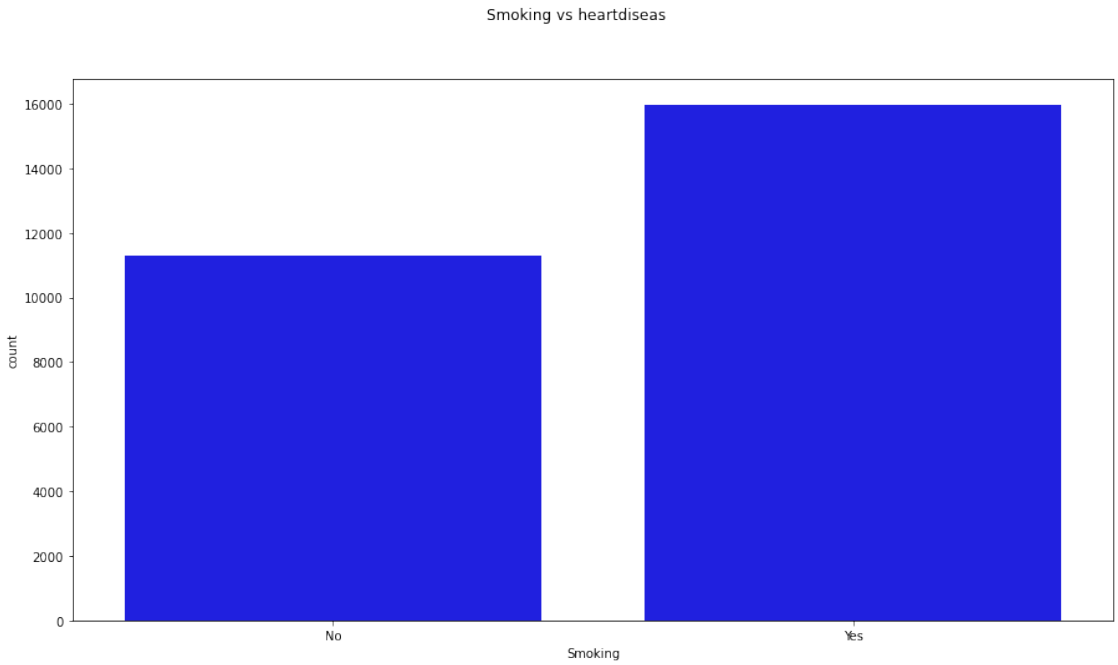
[16]:
```
figure = plt.figure(figsize = (15,8))
figure.suptitle('BMI distributon')
f1 = sns.barplot(x=BMI['BMICategories'], y =BMI['count'], color = 'Blue' )
```



9 Is smoking one of the reasons for heart disease?

[67]:
```
smk = heart_disease.groupby('Smoking')['HeartDisease'].count().reset_index(name␣
↪= 'count')
```

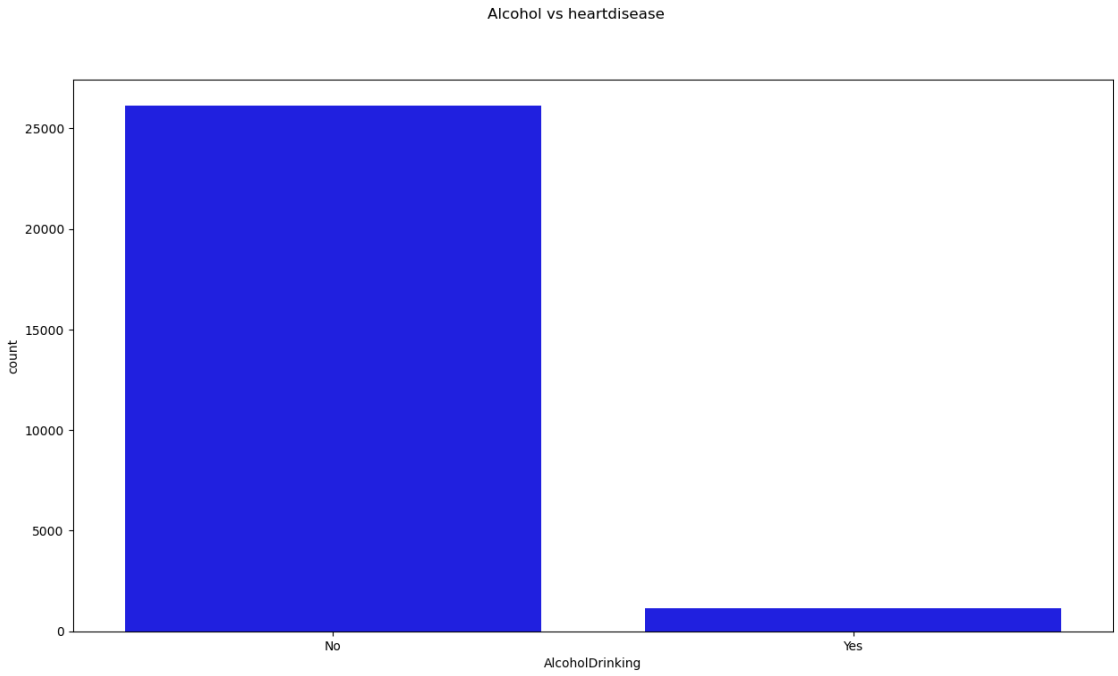
[68]:
```
figure = plt.figure(figsize = (15,8))
figure.suptitle('Smoking vs heartdiseas')
f2 = sns.barplot(x=smk['Smoking'], y =smk['count'], color = 'Blue' )
```



10 Is alcohol one of the reason of heart disease?

[30]:
```
alc = heart_disease.groupby('AlcoholDrinking')['HeartDisease'].count().
↪reset_index(name = 'count')
```

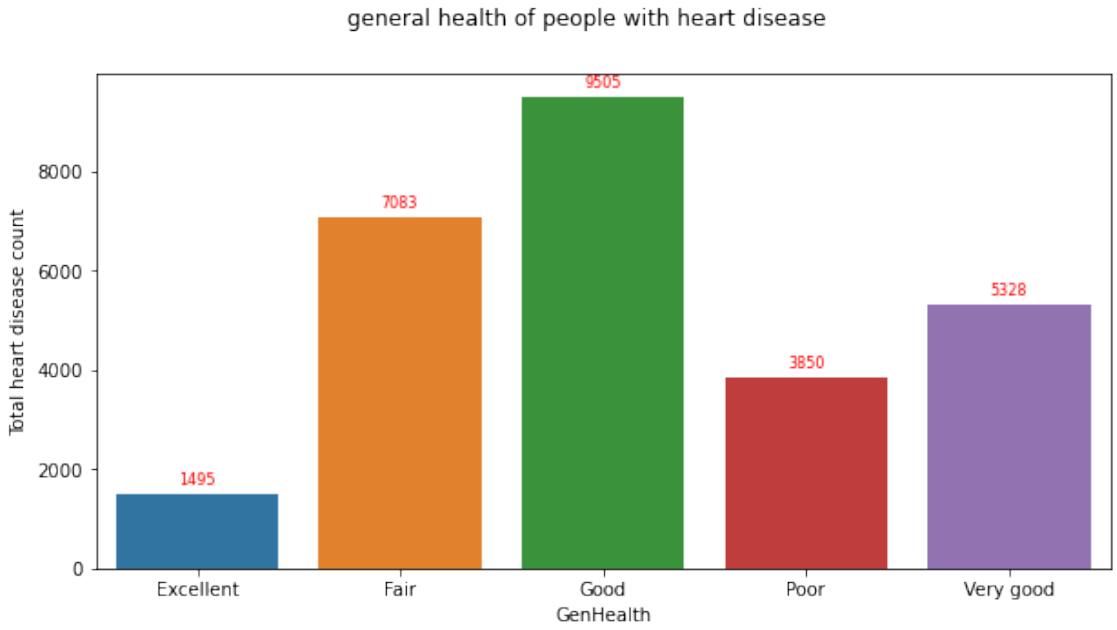
[31]:
```
figure = plt.figure(figsize = (15,8))
figure.suptitle('Alcohol vs heartdisease')
f3 = sns.barplot(x=alc['AlcoholDrinking'], y =alc['count'], color = 'Blue' )
```



11 General health of people having heart disease?

[71]:
```
gen_health = heart_disease.groupby('GenHealth')['HeartDisease'].count().
  ↪reset_index(name = 'count')
```

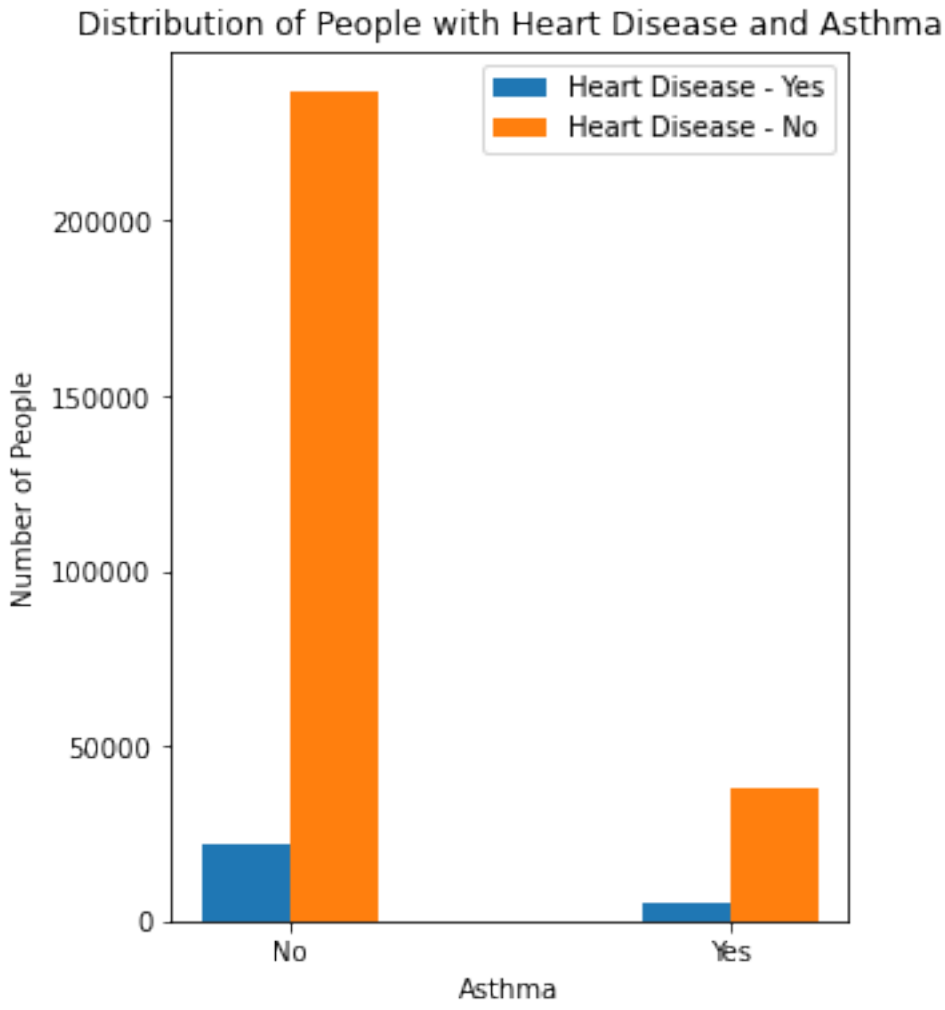
[76]:
```
fig = plt.figure(figsize = (10,5))
fig.suptitle("general health of people with heart disease")
ax= sns.barplot(x=gen_health['GenHealth'], y=gen_health['count'])
ax.set_ylabel('Total heart disease count')
ax.set_xticklabels(ax.get_xticklabels())

for containers in ax.containers:
    ax.bar_label(containers,label_type='edge', padding=3, size=8,color= 'red')
```

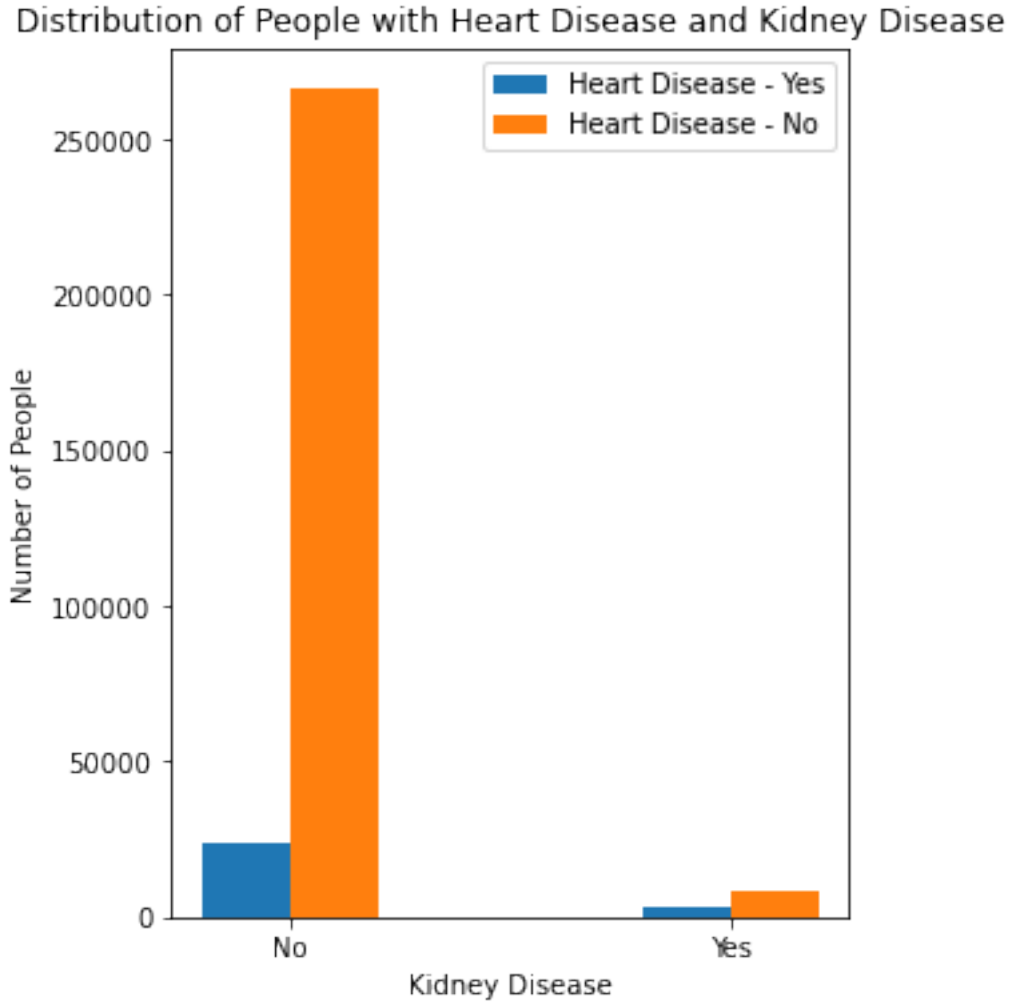


12 Is asthma and kidney diseases is related to heart disease?

[77]:
```
asthma_data = heart_data.groupby(['HeartDisease', 'Asthma']).size().
 ↪unstack(fill_value=0)
kidney_data = heart_data.groupby(['HeartDisease', 'KidneyDisease']).size().
 ↪unstack(fill_value=0)

# Create a bar chart
plt.figure(figsize=(10, 6))
plt.subplot(1,2,1)
x_labels = asthma_data.columns
x = range(len(x_labels))
width = 0.2

plt.bar(x, asthma_data.loc['Yes'], width, label='Heart Disease - Yes')
plt.bar([i + width for i in x], asthma_data.loc['No'], width, label='Heart␣
 ↪Disease - No')

plt.xlabel('Asthma')
plt.ylabel('Number of People')
plt.title('Distribution of People with Heart Disease and Asthma')
plt.xticks([i + width/2 for i in x], x_labels)
plt.legend()
plt.show()
```

```
plt.figure(figsize=(10, 6))
plt.subplot(1,2,2)
x_labels = kidney_data.columns
x = range(len(x_labels))
width = 0.2

plt.bar(x, kidney_data.loc['Yes'], width, label='Heart Disease - Yes')
plt.bar([i + width for i in x], kidney_data.loc['No'], width, label='Heart␣
 ↪Disease - No')

plt.xlabel('Kidney Disease')
plt.ylabel('Number of People')
plt.title('Distribution of People with Heart Disease and Kidney Disease')
plt.xticks([i + width/2 for i in x], x_labels)
plt.legend()
plt.show()
```

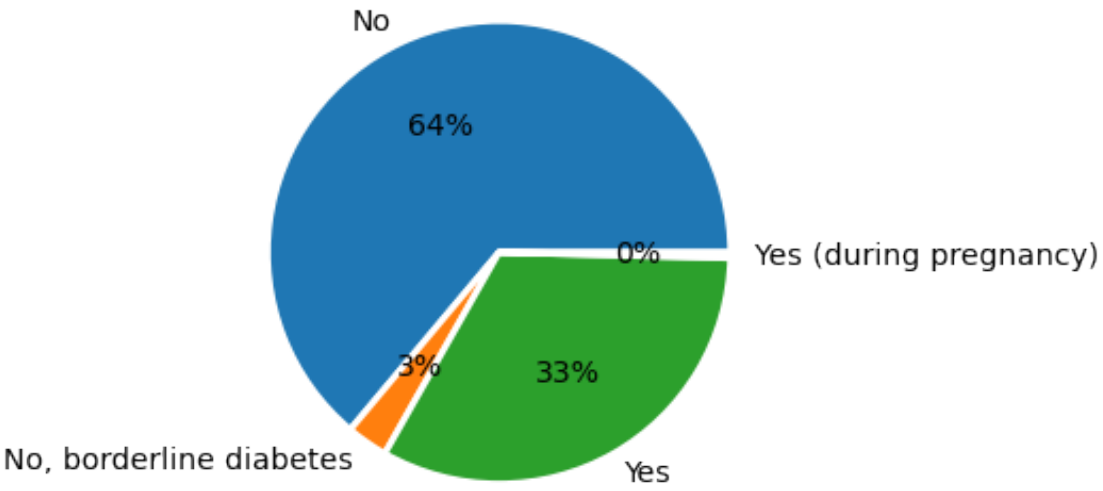




13 Is diabetes is realated to heart disease?

[78]:
```
diab = heart_disease.groupby('Diabetic')['HeartDisease'].count().
↪reset_index(name = 'count')
```

[79]:
```
fig5 = plt.figure(figsize=(10,5))
plt.pie(diab['count'], labels = diab['Diabetic'], autopct='%1.0f%%',textprops =␣
↪{'fontsize': 14}, wedgeprops = {'linewidth': 3,'edgecolor': 'white'})
plt.show()
plt.pie
```



[79]:
```
<function matplotlib.pyplot.pie(x, explode=None, labels=None, colors=None,
autopct=None, pctdistance=0.6, shadow=False, labeldistance=1.1, startangle=0,
radius=1, counterclock=True, wedgeprops=None, textprops=None, center=(0, 0),
frame=False, rotatelabels=False, *, normalize=True, data=None)>
```

14 Is diffwalking have any relationship with heart disease?

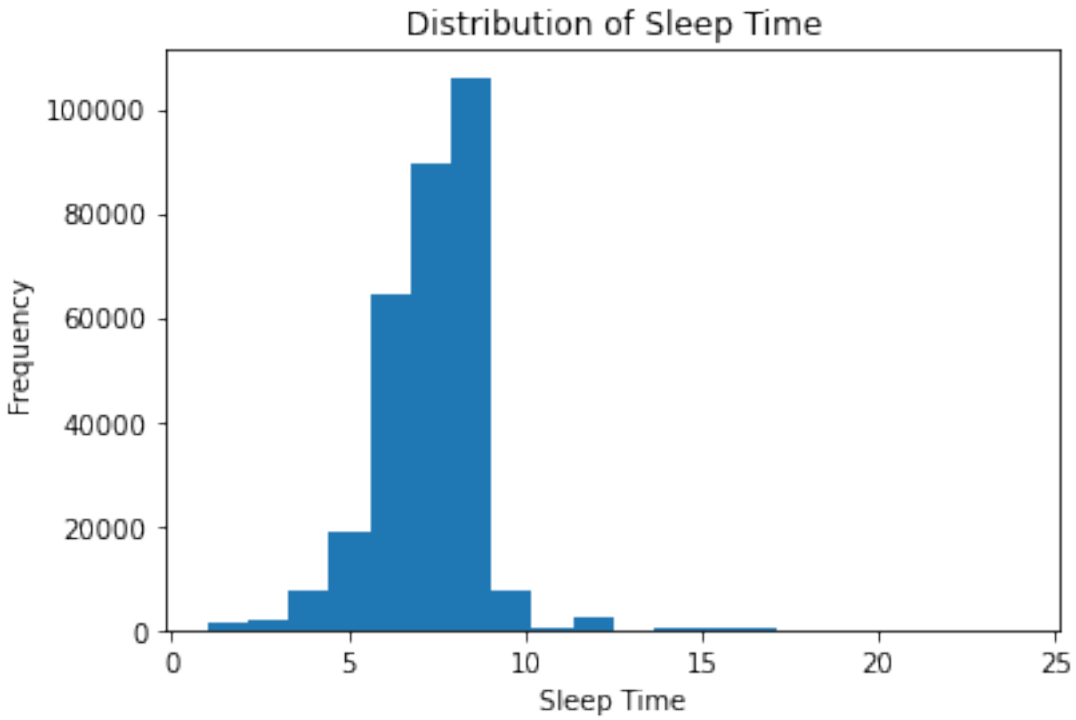
[17]:
```
fig = px.histogram(heart_data, x='DiffWalking', color='HeartDisease',
                   template='plotly_dark', barmode='group',
                   color_discrete_sequence=['#71AEC2', '#D58989'])
fig.update_layout(title='Heart Disease Frequency for DiffWalking',
                  xaxis_title='DiffWalking',
                  yaxis_title='Frequency',
                  legend_title='Heart Disease')
fig.show()
```

15 heart disease vs race

[18]:
```
fig = px.histogram(heart_data, x='Race', color='HeartDisease',
                   template='plotly_dark', barmode='group',
                   color_discrete_sequence=['#71AEC2', '#D58989'])
fig.update_layout(title='Heart Disease Frequency for Race',
                  xaxis_title='Race',
                  yaxis_title='Frequency',
                  legend_title='Heart Disease')
```

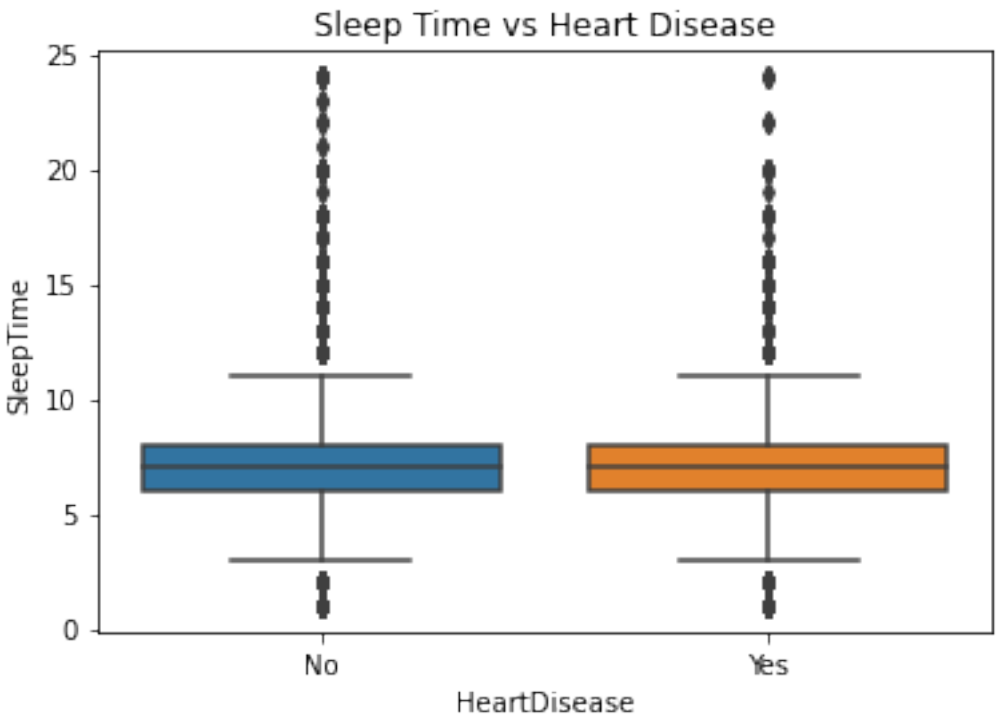
```
fig.show()
```

[82]:
```
plt.hist(heart_data['SleepTime'], bins=20)
plt.title('Distribution of Sleep Time')
plt.xlabel('Sleep Time')
plt.ylabel('Frequency')
plt.show()
```

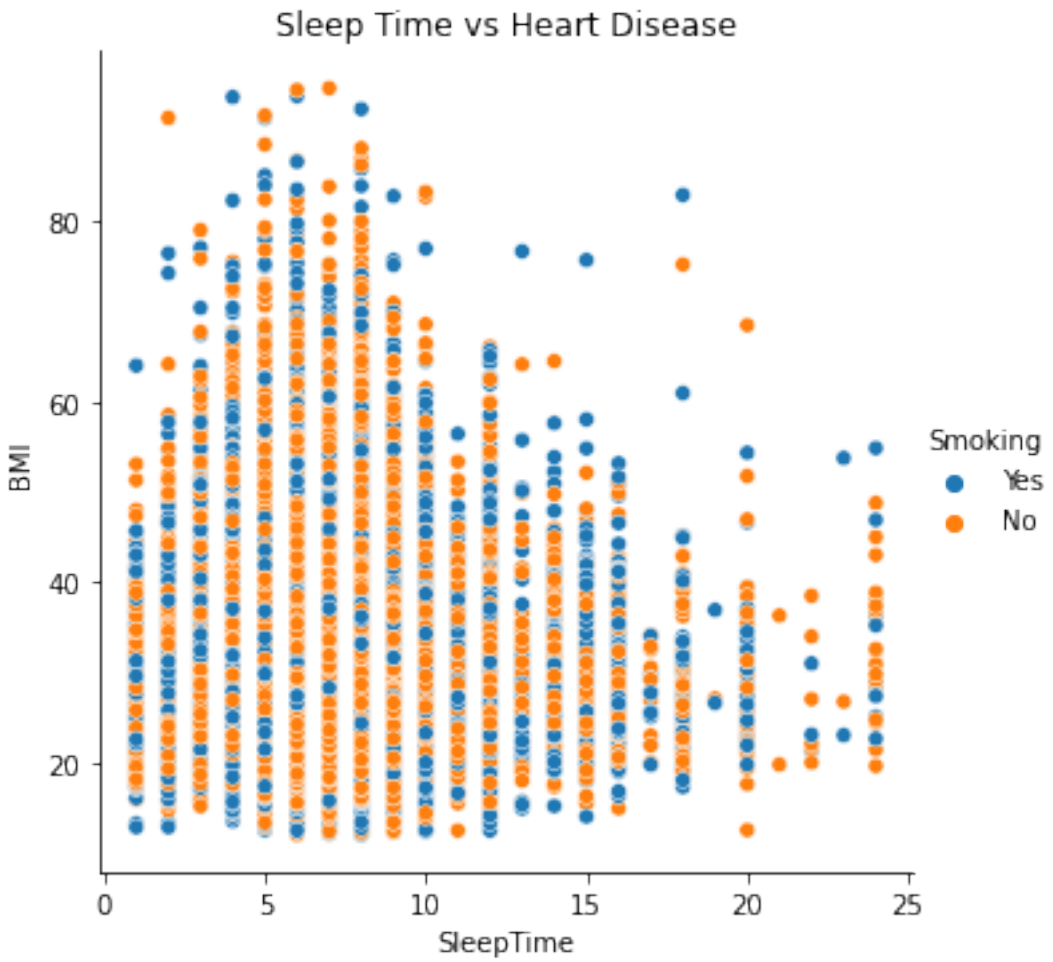


16 Sleep pattern vs heartdisease

[83]:
```
sns.boxplot(x='HeartDisease', y='SleepTime', data=heart_data)
plt.title('Sleep Time vs Heart Disease')
plt.show()
```



[84]:
```
sns.relplot(x='SleepTime', y='BMI', hue='Smoking', data=heart_data)
plt.title('Sleep Time vs Heart Disease')
plt.show()
```



17 stroke vs heartdisease

[85]:
```
stroke = heart_disease.groupby('Stroke')['HeartDisease'].count().
↪reset_index(name = 'count')
```

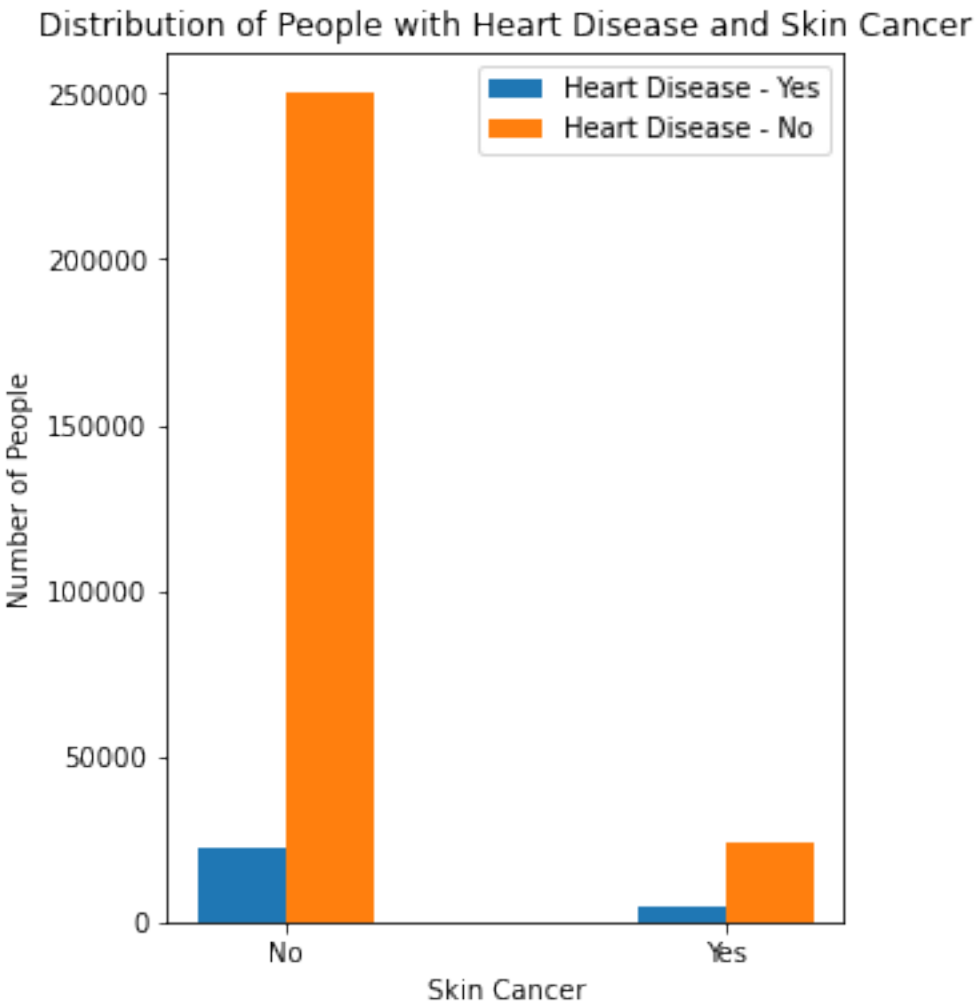
[86]:
```
print(stroke)
```

```
  Stroke  count
0     No  22872
1    Yes   4389
```

18 Distribution of people with Heart Disease vs Skin Cancer

[87]:
```
heart = heart_data.groupby(['HeartDisease', 'SkinCancer']).size().
↪unstack(fill_value=0)
```

[88]:
```
# Create a bar chart
plt.figure(figsize=(10, 6))
plt.subplot(1,2,1)
x_labels = heart.columns
x = range(len(x_labels))
width = 0.2

plt.bar(x, heart.loc['Yes'], width, label='Heart Disease - Yes')
plt.bar([i + width for i in x], heart.loc['No'], width, label='Heart Disease -␣
↪No')

plt.xlabel('Skin Cancer')
plt.ylabel('Number of People')
plt.title('Distribution of People with Heart Disease and Skin Cancer')
plt.xticks([i + width/2 for i in x], x_labels)
plt.legend()
plt.show()
```


Distribution of People with Heart Disease and Skin Cancer
Heart Disease - Yes
Heart Disease - No
250000
200000
150000
100000
50000
0
Number of People
No
Yes
Skin Cancer