Heart Disease Detection

Table of contents

Refresh

1 Introduction

According to the world health organization, Cardiovascular diseases (CVDs) are the leading cause of death globally. In 2019 alone, around 17.9 million people died from CVDs. Of these deaths, **85%** of them were due to heart diseases. There are many factors that play a role in increasing the risk of heart disease. Identifying these factors and their impact is paramount in the field of healthcare. Identifying patients who are at greater risk enables medical professionals to respond quickly and efficiently, saving more lives.

1.1 About This Project

In this project we will delve deep into the causess of heart disease and its relations with other health indicators, drawing insights and exploring the data in order to get a better picture of the leading causes. Finally, we use statistical models in order to automate heart disease detection.

An immediate application on this is early detection of heart disease in hospitals, thus enabling proactive measurements instead of reactive. Therefore, the goal of the project is to draw statistical insights, and construct a real world application for the dataset.

1.2 About the Dataset

The Personal Key Indicators of Heart Disease dataset contains 320K rows and 18 columns. It is a cleaned, smaller version of the 2020 annual CDC (Centers for Disease Control and Prevention) survey data of 400k adults. For each patient (row), it contains the health status of that individual. The data was collected in the form of surveys conducted over the phone. Each year, the CDC calls around 400K U.S residents and asks them about their health status, with the vast majority of questions being yes or no questions. Below is a description of the features collected for each patient:

#	Feature	Description
1	HeartDisease	Respondents that have ever reported having coronary heart disease (CHD) or myocardial infarction (MI)
2	BMI	Body Mass Index (BMI)
3	Smoking	Have you smoked at least 100 cigarettes in your entire life? [Note: 5 packs = 100 cigarettes]
4	AlcoholDrinking	Heavy drinkers (adult men having more than 14 drinks per week and adult women having more than 7 drinks
5	Stroke	(Ever told) (you had) a stroke?
6	PhysicalHealth	Now thinking about your physical health, which includes physical illness and injury, for how many days during
7	MentalHealth	Thinking about your mental health, for how many days during the past 30 days was your mental health not go
8	DiffWalking	Do you have serious difficulty walking or climbing stairs?
9	Sex	Are you male or female?
10	AgeCategory	Fourteen-level age category
11	Race	Imputed race/ethnicity value
12	Diabetic	(Ever told) (you had) diabetes?
13	PhysicalActivity	Adults who reported doing physical activity or exercise during the past 30 days other than their regular job
14	GenHealth	Would you say that in general your health is...
15	SleepTime	On average, how many hours of sleep do you get in a 24-hour period?
16	Asthma	(Ever told) (you had) asthma?
17	KidneyDisease	Not including kidney stones, bladder infection or incontinence, were you ever told you had kidney disease?
18	SkinCancer	(Ever told) (you had) skin cancer?

1.3 About This Notebook

This Notebook is concerned with exploring the data, the relation of heart disease with other features on our data, and the patterns and insights hidden inside the data. We will visualize the the data in various ways in our exploration.

2 Data Overview

```
!pip install pywaffle
```

```
Looking in indexes: https://pypi.org/simple, https://us-python.pkg.dev/colab-wheels/pub
Collecting pywaffle
  Downloading pywaffle-1.1.0-py2.py3-none-any.whl (30 kB)
Collecting fontawesomefree
  Downloading fontawesomefree-6.2.0-py3-none-any.whl (25.1 MB)
     |████████████████████████████████| 25.1 MB 79.6 MB/s
Requirement already satisfied: matplotlib in /usr/local/lib/python3.7/dist-packages (fro
Requirement already satisfied: cycler>=0.10 in /usr/local/lib/python3.7/dist-packages (f
Requirement already satisfied: numpy>=1.11 in /usr/local/lib/python3.7/dist-packages (fr
Requirement already satisfied: pyparsing!=2.0.4,!=2.1.2,!=2.1.6,>=2.0.1 in /usr/local/li
Requirement already satisfied: python-dateutil>=2.1 in /usr/local/lib/python3.7/dist-pac
Requirement already satisfied: kiwisolver>=1.0.1 in /usr/local/lib/python3.7/dist-packag
Requirement already satisfied: typing-extensions in /usr/local/lib/python3.7/dist-packag
Requirement already satisfied: six>=1.5 in /usr/local/lib/python3.7/dist-packages (from
Installing collected packages: fontawesomefree, pywaffle
Successfully installed fontawesomefree-6.2.0 pywaffle-1.1.0
```

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
import plotly.express as px
from pywaffle.waffle import Waffle
import shutil
from os import path


%matplotlib inline
sns.set_style("darkgrid")

colors6 = sns.color_palette(['#1337f5', '#E80000', '#0f1e41', '#fd523e', '#404e5c', '#c9bbaa'
colors2 = sns.color_palette(['#1337f5', '#E80000'], 2)
colors1 = sns.color_palette(['#1337f5'], 1)
```

2.1 Data Reading

```
if path.exists('./heart-disease'):
    shutil.rmtree('./heart-disease')

!git clone https://github.com/lemonpudding-datasets/heart-disease.git

shutil.move("./heart-disease/heart_2020_cleaned.csv", "./heart_2020_cleaned.csv")
shutil.rmtree('./heart-disease')
```

```
Cloning into 'heart-disease'...
remote: Enumerating objects: 3, done.
remote: Counting objects: 100% (3/3), done.
remote: Compressing objects: 100% (3/3), done.
remote: Total 3 (delta 0), reused 0 (delta 0), pack-reused 0
Unpacking objects: 100% (3/3), done.
```

```
df = pd.read_csv("heart_2020_cleaned.csv")
df.head()
```

	HeartDisease	BMI	Smoking	AlcoholDrinking	Stroke	PhysicalHealth	MentalHealth
0	No	16.60	Yes	No	No	3.0	30.0
1	No	20.34	No	No	Yes	0.0	0.0
2	No	26.58	Yes	No	No	20.0	30.0
3	No	24.21	No	No	No	0.0	0.0
4	No	23.71	No	No	No	28.0	0.0

2.2 Dataset Info

```
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
```

2.3 Numerical Description

```
df.describe()
```

	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

2.4 Categorical Description

```
df.describe(include="object")
```

	HeartDisease	Smoking	AlcoholDrinking	Stroke	DiffWalking	Sex	AgeCategory
count	319795	319795	319795	319795	319795	319795	319795
unique	2	2	2	2	2	2	13
top	No	No	No	No	No	Female	65-69

2.5 Initial Data Assessment

Upon initial inspection, the following is observed:

1. We have 319795 samples.
2. We no missing values.
3. The majority of features are categorical.
4. *BMI* is skewed.
5. *PhysicalHealth* and *MentalHealth* are severly skewed due to the number of zeros.
6. *SleepTime* is normally distributed.
7. There is a severe class imbalance (heart disease vs healthy)
8. *Alcohol* drinking is imbalanced.
9. *Sroke* is imbalanced.
10. *DiffWalking* is imbalanced.
11. *Age* is categorical (divided into bins).
12. *Race* is imbalanced, with the majority being white.
13. *Diabetic* is imbalanced.
14. *PhysicalActivity* is imbalanced
15. *Asthma* is imbalanced
16. *KidneyDisease* is imbalanced
17. *SkinCancer* is imbalanced

3 EDA

3.1 Helper Functions

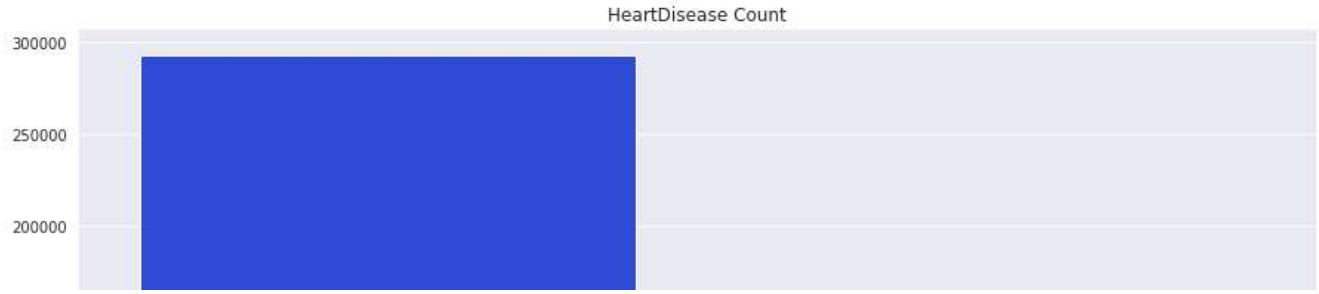
```
def show_relation(col, according_to, type_='dis'):
  plt.figure(figsize=(15,7));

  if type_=='dis':
```

```
      sns.displot(data=df, x=col, hue=according_to, kind='kde', palette=colors2);
    elif type_=='count':
      if according_to != None:
        perc = df.groupby(col)[according_to].value_counts(normalize=True).reset_index(name='Per
        sns.barplot(data=perc, x=col,y='Percentage', hue=according_to, palette=colors6, order=d
      else:
        sns.countplot(data=df, x=col, hue=according_to, palette=colors1, order=df[col].value_co

    if according_to==None:
      plt.title(f'{col}');
    else:
      plt.title(f'{col} according to {according_to}');


def generate_colors(num):
    colors = []
    lst = list('ABCDEF0123456789')

    for i in range(num):
        colors.append('#'+''.join(np.random.choice(lst, 6)))

    return colors
```

3.2 General Questions

3.2.1 What are the percentage of heart disease patients?

```
plt.figure(figsize=(15,7));
plt.title('HeartDisease Count');
sns.countplot(data=df, x='HeartDisease', palette=colors2, order=df['HeartDisease'].value_coun
```

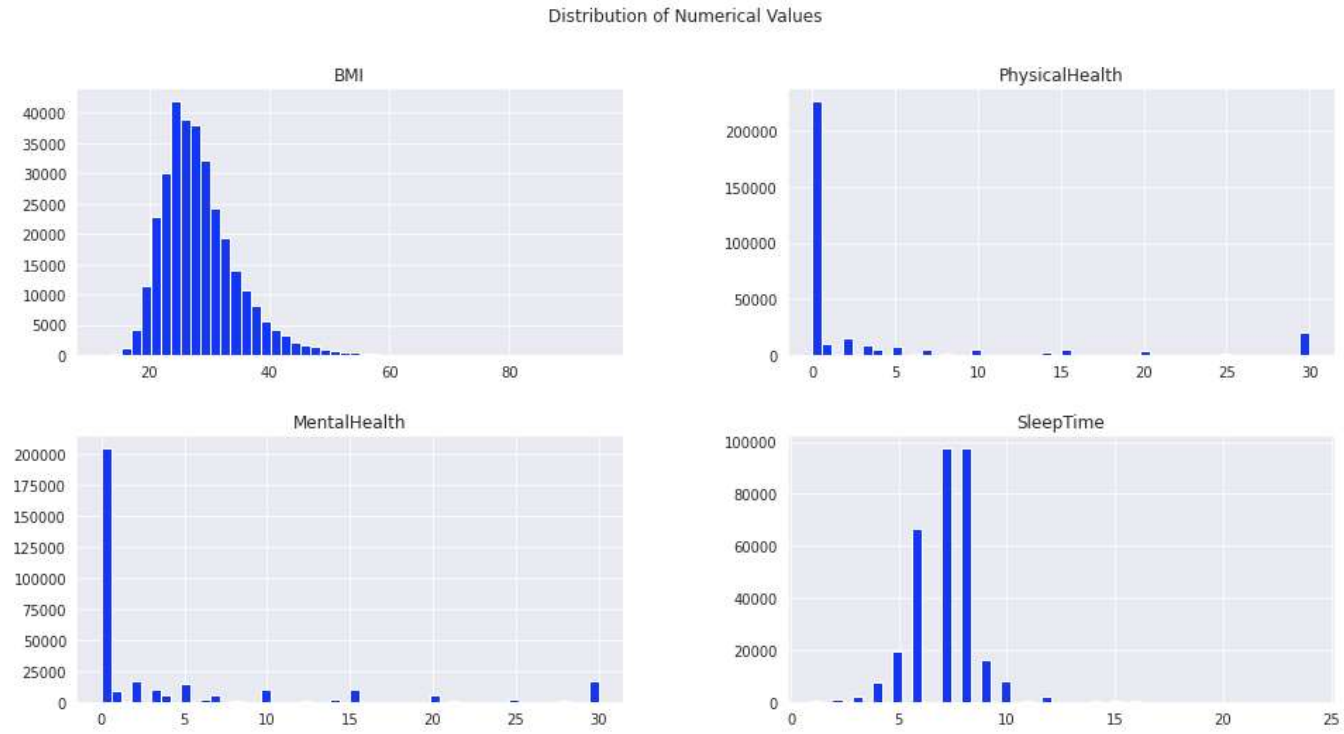



```
# get percentage of attrition then convert to dicrionary
disease_size = (df.groupby('HeartDisease').size()*100 / len(df)).to_dict()

# create figure
fig = plt.figure(
    FigureClass=Waffle, # type = waffle figure
    rows=5, # rows of people
    figsize = (9,3),
    values=disease_size, # data

    # legend labels
    labels=[f"{k} ({round(v / sum(disease_size.values()) * 100, 2)}%)"
            for k, v in disease_size.items()],
    # colors for attrition and no attrition
    colors=(colors2[0], colors2[1]),
    # icons set to person for both attrition and no attriton
    icons = ['heart','heart'],
    # the legend at the bottom, after playing with the
    # locations i centered it at the bottom
    legend={'loc': 'lower center',
            'bbox_to_anchor': (0.5, -0.5),
            'ncol': len(disease_size),
            'framealpha': 0,
            'fontsize': 20
          },

    # size of icons (people)
    icon_size=20,

    # add icon to the legend at the bottom
    icon_legend=True,

    #title of the waffle graph
    title={
        'label': 'Heart Disease Per 100 People',
        'loc': 'center',
        'fontdict': {'fontsize': 20}
          }
)
```

Approximately 9 in 100 people suffer from heart disease in the united states.

3.2.2 What are the distributions of our features?

```
df.hist(figsize=(16, 8), bins=50, color=colors1);
plt.suptitle("Distribution of Numerical Values");
```

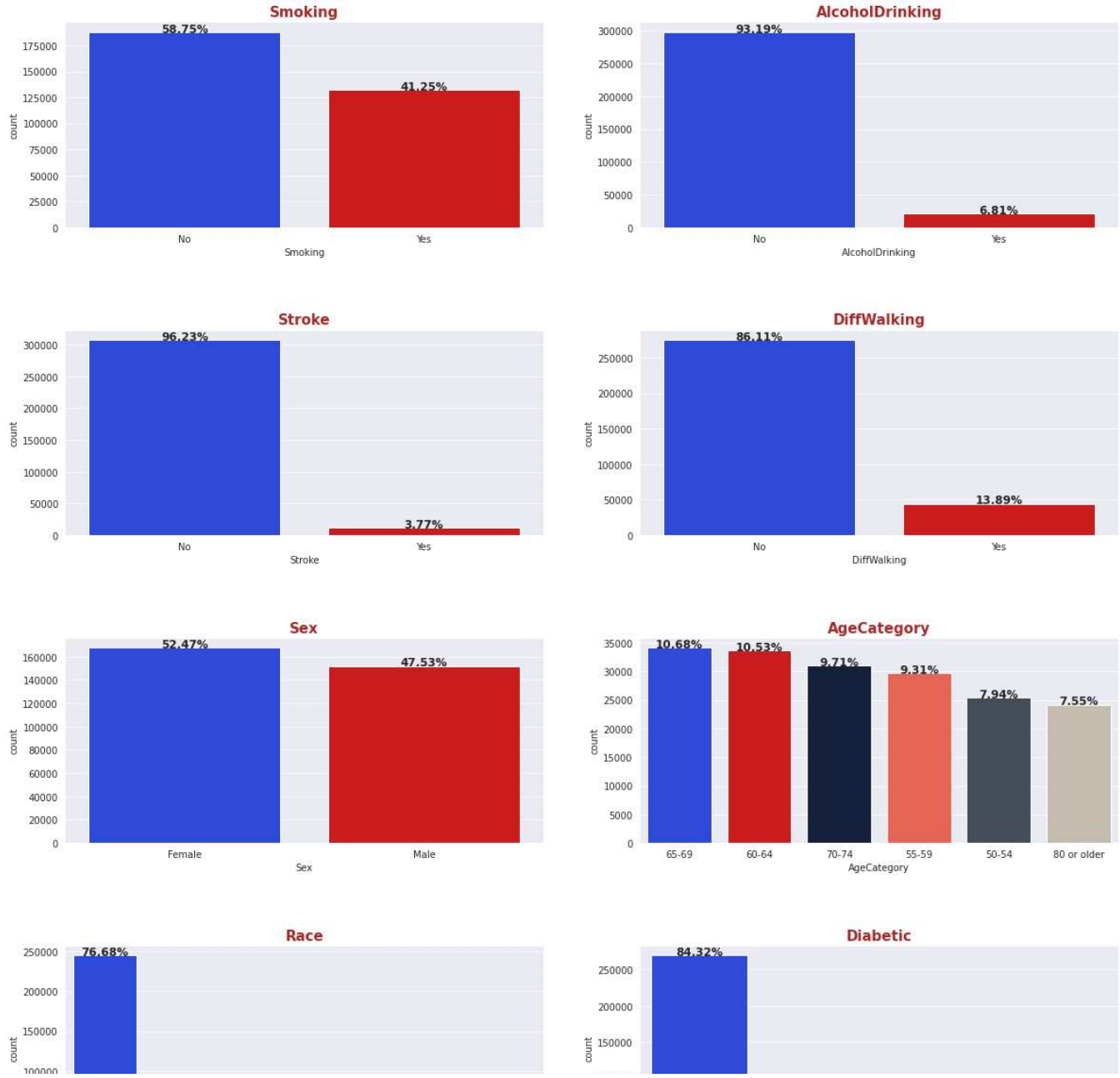



```
obj_cols = df.select_dtypes(include='object').columns[1:]
num_cols = df.select_dtypes(exclude='object').columns
```

```
print(f'Object columns : {obj_cols}', end='\n\n')
print(f'Numberical columns : {num_cols}')
```

```
Object columns : Index(['Smoking', 'AlcoholDrinking', 'Stroke', 'DiffWalking', 'Sex',
       'AgeCategory', 'Race', 'Diabetic', 'PhysicalActivity', 'GenHealth',
       'Asthma', 'KidneyDisease', 'SkinCancer'],
      dtype='object')

Numberical columns : Index(['BMI', 'PhysicalHealth', 'MentalHealth', 'SleepTime'], dtype
```

```
plt.figure(figsize=(20, 40))
for i in range(len(obj_cols)):
  plt.subplot(7, 2, i+1)

  if(df[obj_cols[i]].nunique() < 3):
    ax = sns.countplot(data=df, x=obj_cols[i], palette=colors2, order=df[obj_cols[i]].value_c
  else:
    ax = sns.countplot(data=df, x=obj_cols[i], palette=colors6, order=df[obj_cols[i]].value_c


  plt.title(f'{obj_cols[i]}', fontsize=15, fontweight='bold', color='brown')
  plt.subplots_adjust(hspace=0.5)

  for p in ax.patches:
    height = p.get_height()
    width = p.get_width()
    percent = height/len(df)

    ax.text(x=p.get_x()+width/2, y=height+2, s=format(percent, ".2%"), fontsize=12, ha='cente
```



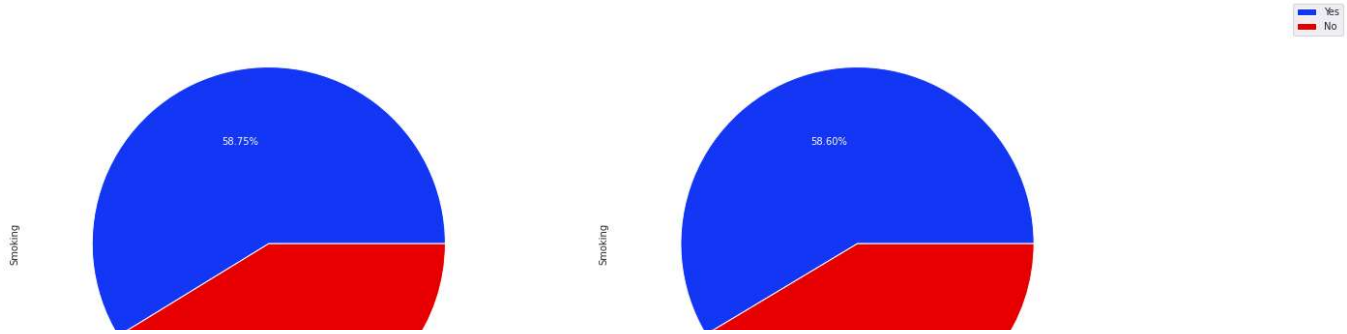
Insights from these plots

- Most of people in our data are white and have no diabetic.
- Most of them had done a physical activity during the past 30 days other than their regular job and in general they have very good health as they said.
- A litle of them who have asthma, kidney disease and skin cancer.
- Most people said that they have generaly very good health. A few of people who said that they have generaly a poor health.

3.2.3 What are the percentages for each categorical variable?


```
for col in obj_cols:
```

```
        fig, ax = plt.subplots(1,2, figsize=(20,20))
        round(df[col].value_counts()/df.shape[0]*100, 2).plot.pie(autopct="%1.2f%%", ax=ax[0], te
        round(df[(df['HeartDisease'] == 'Yes')][col].value_counts()/df.shape[0]*100, 2).plot.pie(
        plt.legend(loc="upper right", bbox_to_anchor=(1, 0, 0.5, 1))
        plt.show();

    plt.tight_layout()
```

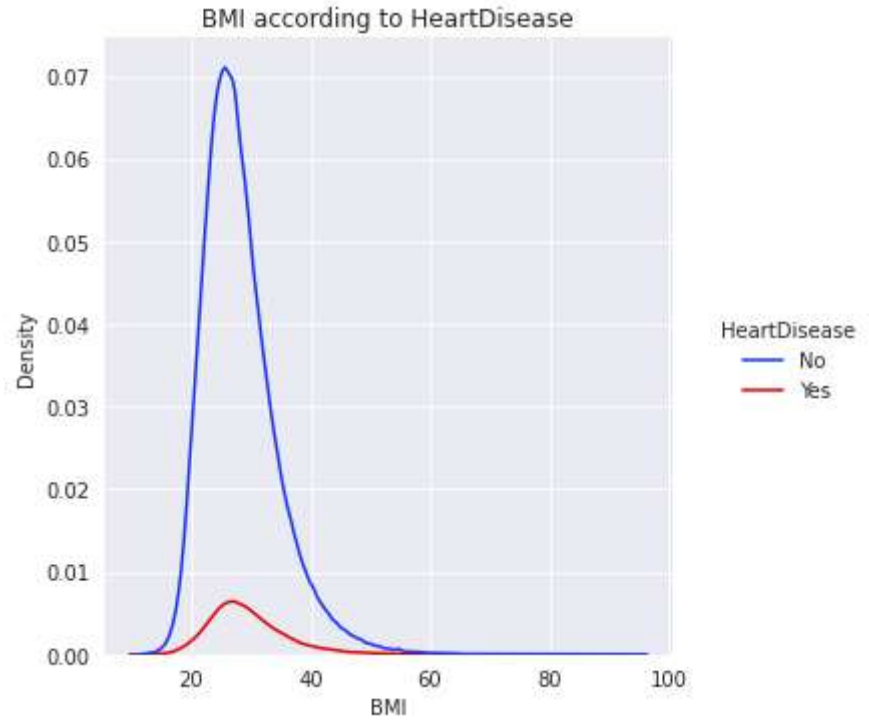


3.3 Demographic and Personal Questions

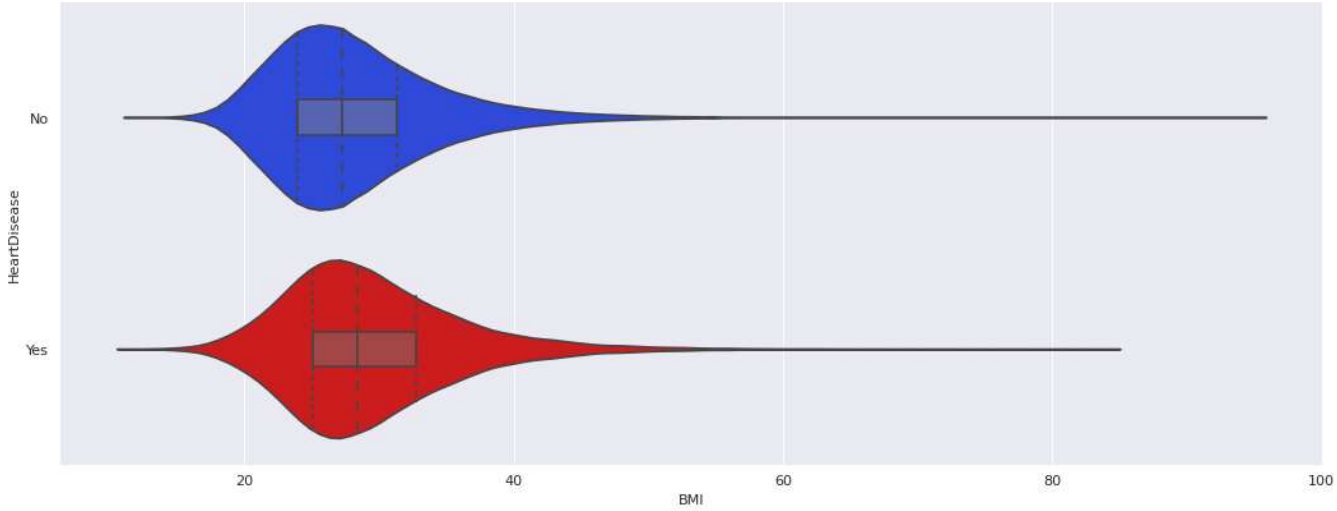
3.3.1 Is the BMI of heart disease patients different?

```
#Body mass index is a value derived from the mass and height of a person.
#The BMI is defined as the body mass divided by the square of the body height
show_relation(num_cols[0], 'HeartDisease');
```

```
<Figure size 1080x504 with 0 Axes>
```




```
plt.figure(figsize=(16, 6), dpi=80)

sns.boxplot(data=df, x='BMI', y='HeartDisease', saturation=0.4,
            width=0.15, boxprops={'zorder': 2},
            showfliers = False, whis=0,  palette=colors2);
sns.violinplot(data=df, x='BMI', y='HeartDisease',inner='quartile', palette=colors2);
```



the both distributions are normal distriubtions and in the same range which is from 12 to 94. The BMI distribution of individuals who suffer from heart disease is slightly shifted towards higher values in comparison to the distribution of those who don't. so we can observe that **BMI didn't affect on heart disease**.

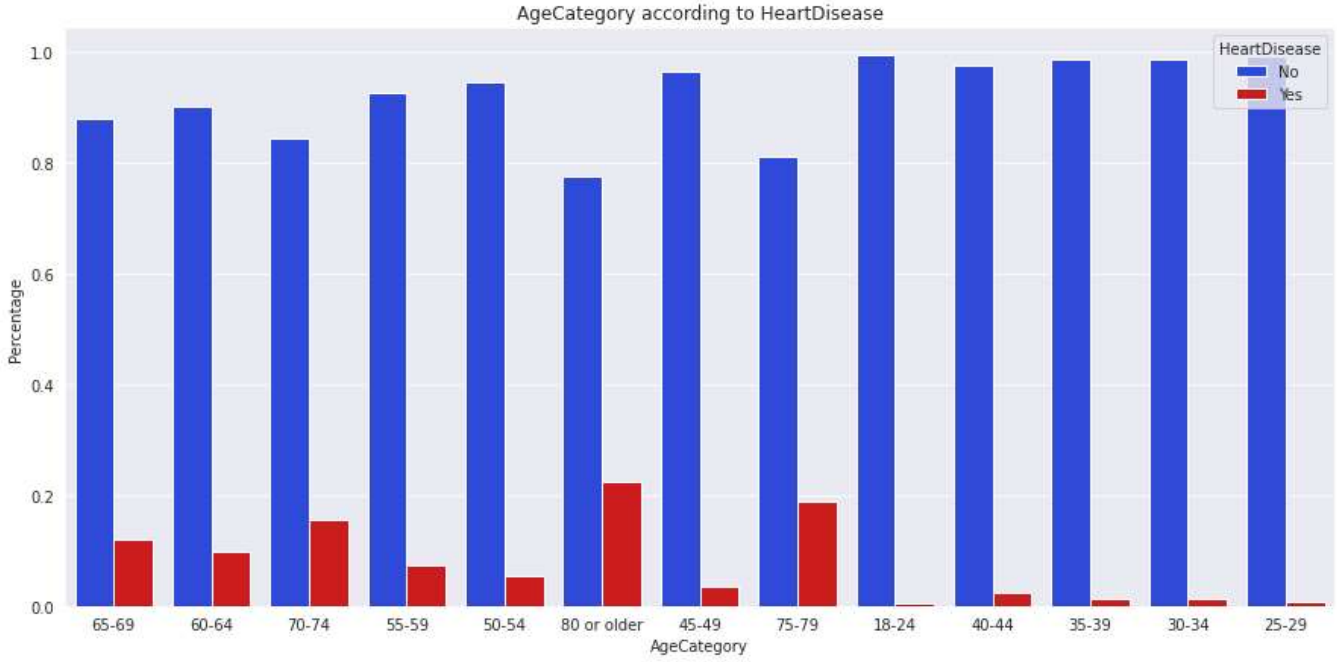
3.3.2 Are heart disease patients more mentally unwell?

```
show_relation(num_cols[2], 'HeartDisease')
```

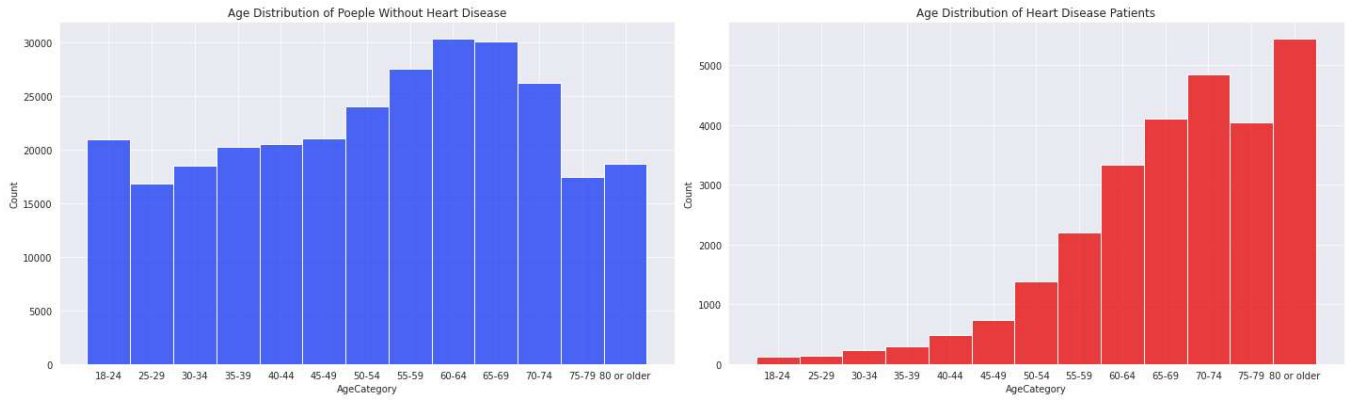
```
<Figure size 1080x504 with 0 Axes>
```

3.3.3 Are older individuals more susceptible to heart disease?

```
show_relation(obj_cols[5], 'HeartDisease', type_='count')
```



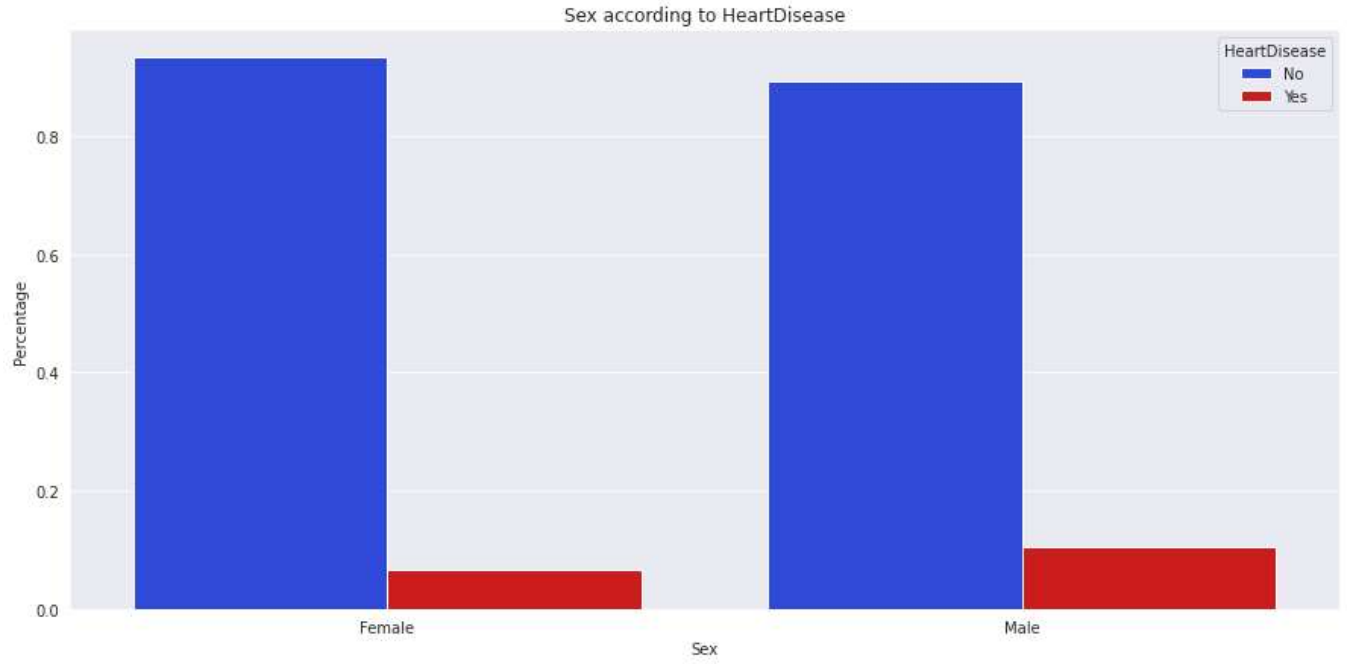

```
fig, (ax1, ax2) = plt.subplots(ncols=2, figsize=(20, 6))

sns.histplot(data=df.loc[df.HeartDisease == 'No'].sort_values("AgeCategory"), x='AgeCategory'
                  color=colors1, ax=ax1);
ax1.set_title("Age Distribution of Poeple Without Heart Disease")

sns.histplot(data=df.loc[df.HeartDisease == 'Yes'].sort_values("AgeCategory"), x='AgeCategory
                  color=colors2[1], ax=ax2);
ax2.set_title("Age Distribution of Heart Disease Patients")


fig.tight_layout()
```



We can see that age plays a big factor in heart disease, as the amount of heart disease patients increases with age. The most susceptible people to the heart disease are people who are greater than 70 years old.

3.3.4 Are males more likely to suffer from heart disease?

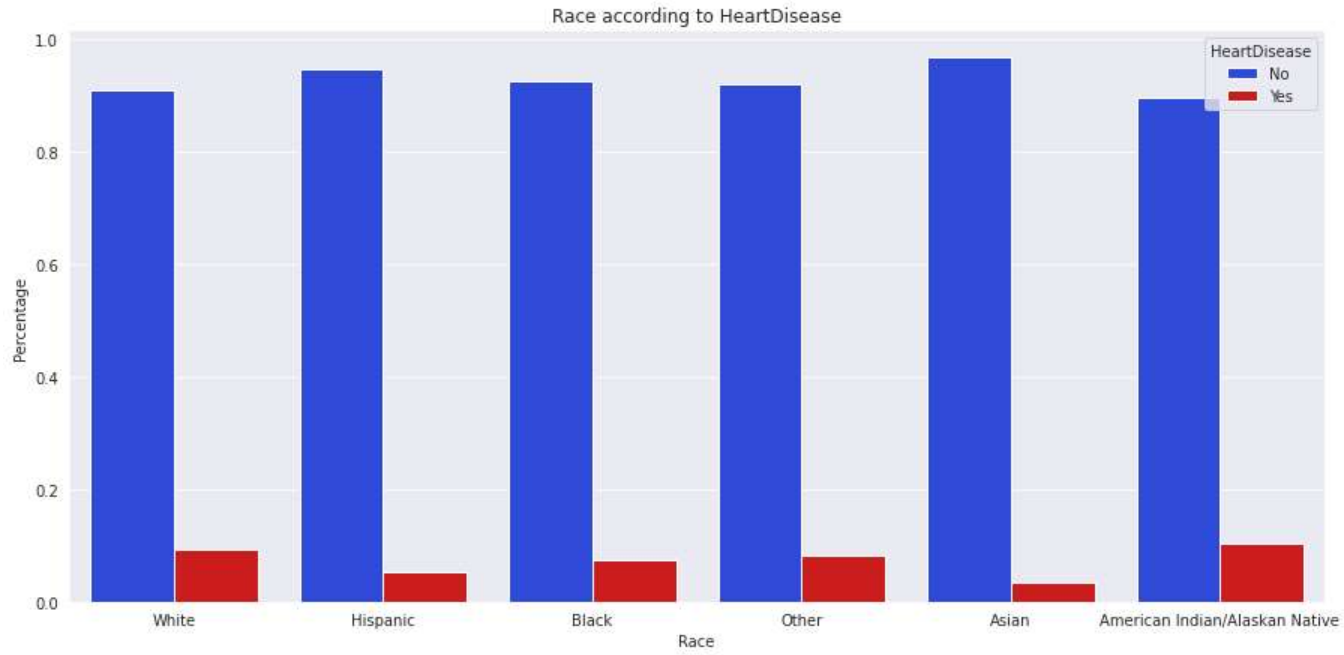
```
show_relation(obj_cols[4], 'HeartDisease', type_='count')
```



> In our data, Males are more susceptible to the heart disease.

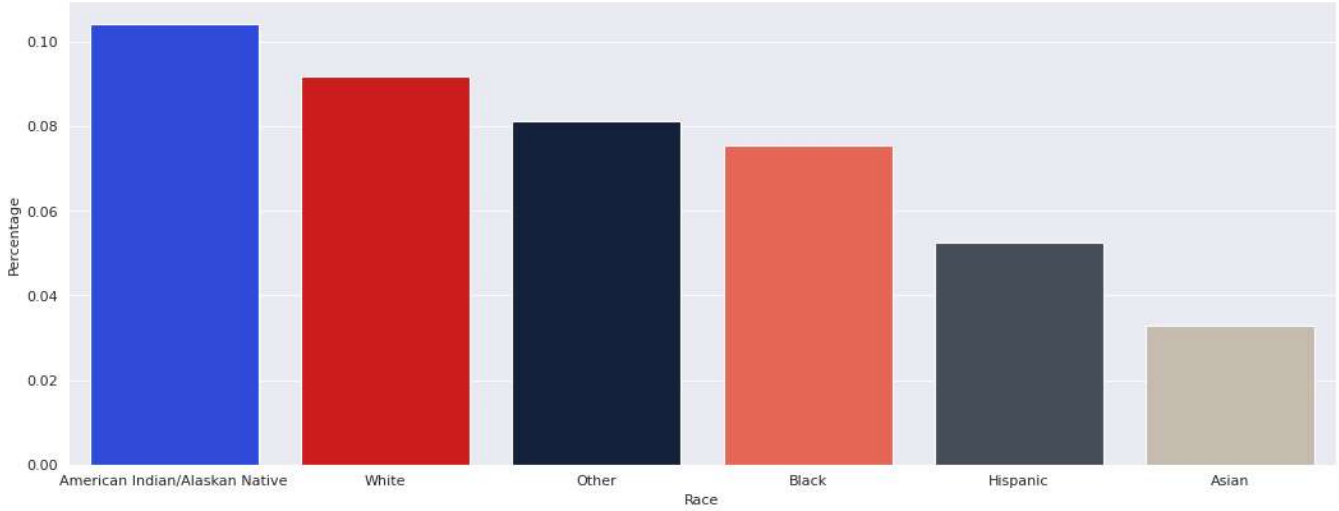
3.3.5 What is the precentage of heart disease among races?

```
show_relation(obj_cols[6], 'HeartDisease', type_='count')
```



> In this plot, we can see the most people has the heart disease is american indian/alaskan native people. **But we can't considerate the race at all. we know that this is have no correlation or affect.**

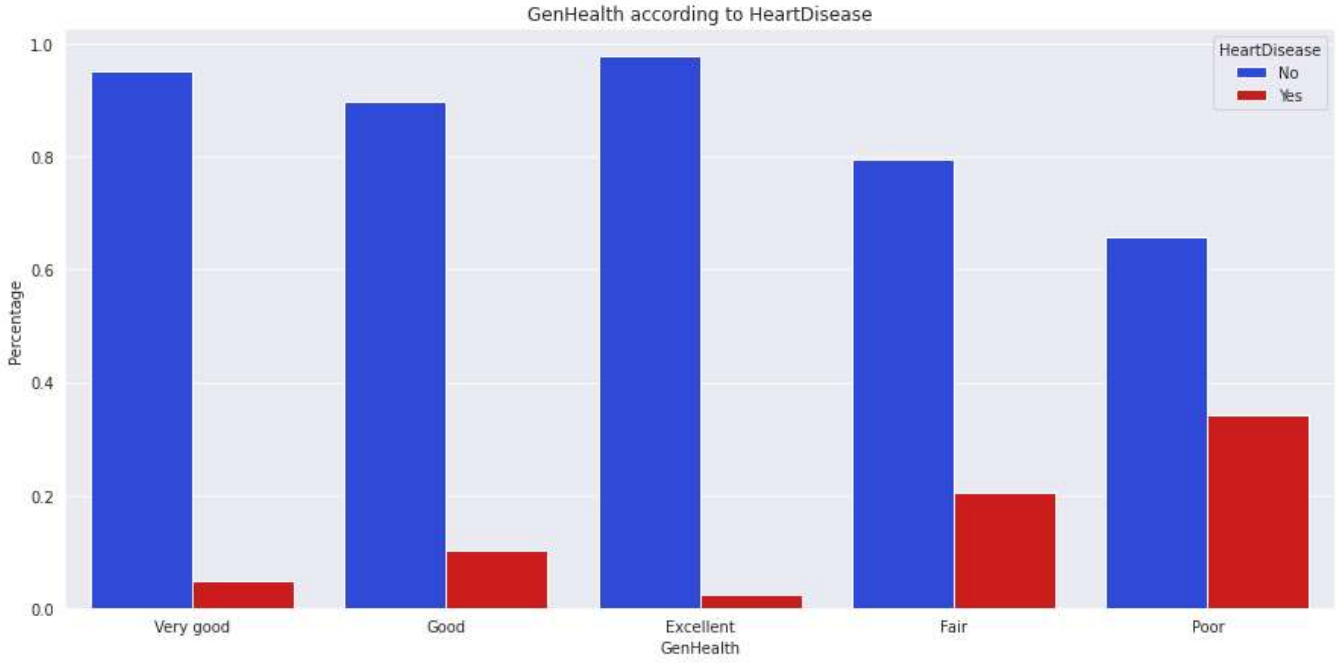
```
plt.figure(figsize=(16, 6), dpi=80)
x = df.groupby('Race').HeartDisease.value_counts(normalize=True).reset_index(name='Percentage
x = x.loc[x.HeartDisease == 'Yes'].sort_values('Percentage', ascending=False)

sns.barplot(data=x, x='Race', y='Percentage', palette=colors6);
```



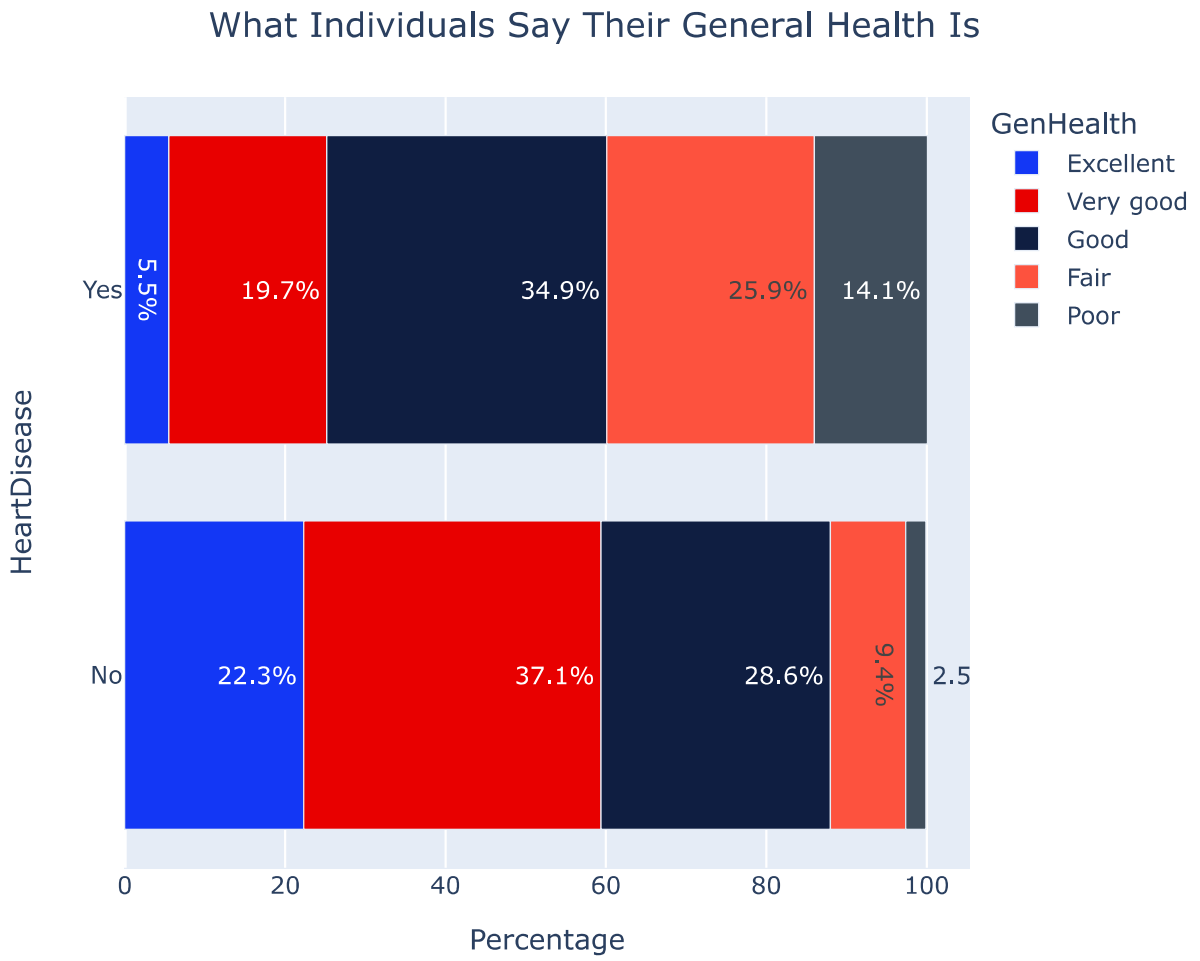
The pecentage of heart disease is highest (> 10%) among Native americans, followed by whites (~9%). The least percentage of heart disease (~3%) is among asians.

3.3.6 What do people who suffer from heart disease perceive their general health?

```
show_relation(obj_cols[9], 'HeartDisease', type_='count')
```



```
x = df.groupby('HeartDisease').GenHealth.value_counts(normalize=True).reset_index(name='Perce
x = x.sort_values(by='GenHealth', key=lambda x: x.map({'Excellent': 0,
                                                       'Very good': 1,
                                                       'Good': 2,
                                                       'Fair': 3,
                                                       'Poor': 4} ))

x.Percentage = round(x.Percentage * 100, 1)

fig = px.bar(data_frame=x, x='Percentage', y='HeartDisease', color='GenHealth',
       text=x.Percentage.map(lambda x: str(x) + '%'),
       color_discrete_sequence=['#1337f5', '#E80000', '#0f1e41', '#fd523e', '#404e5c', '#c9bb
fig.update_layout(title="What Individuals Say Their General Health Is", title_x=0.5)

fig.show()
```



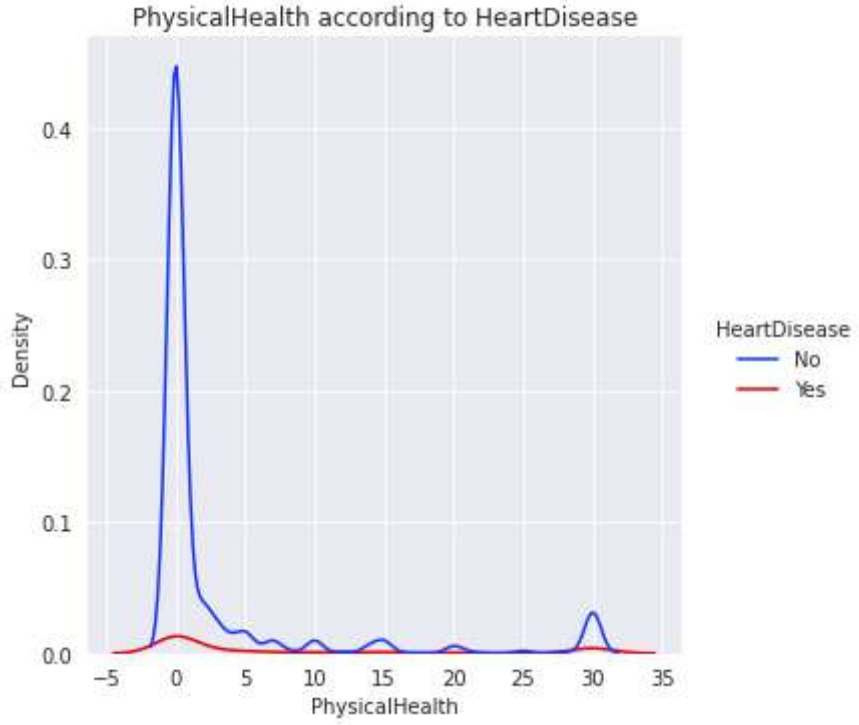
We can see that most heart disease patients believe their health is good, while most people who do not suffer heart diseases say their health is very good. Furhtermore, A lot more people who suffer from heart disease say they have poor or fair health compared to those who don't.

3.4 Routine Related Questions

3.4.1 Are heart disease patients less physically healthy or active?

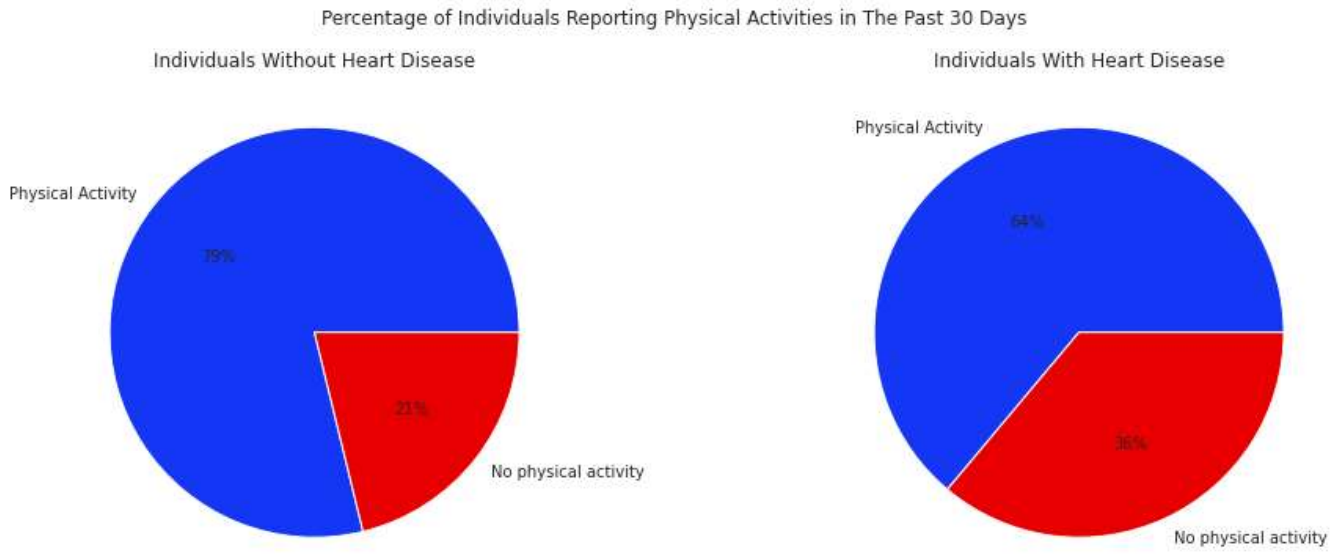
```
show_relation(num_cols[1], 'HeartDisease')
```

```
<Figure size 1080x504 with 0 Axes>
```




```
show_relation(obj_cols[8], 'HeartDisease', type_='count')
```


```
x = df.groupby('HeartDisease').PhysicalActivity.value_counts().reset_index(name='Count').Coun

fig, (ax1, ax2) = plt.subplots(ncols=2, figsize=(16, 6))
fig.suptitle("Percentage of Individuals Reporting Physical Activities in The Past 30 Days")

ax1.pie([x[0], x[1]], labels = ['Physical Activity', 'No physical activity'], colors=colors2,

ax1.set_title("Individuals Without Heart Disease");

ax2.pie([x[2], x[3]], labels = ['Physical Activity', 'No physical activity'], colors=colors2,
ax2.set_title("Individuals With Heart Disease");
```

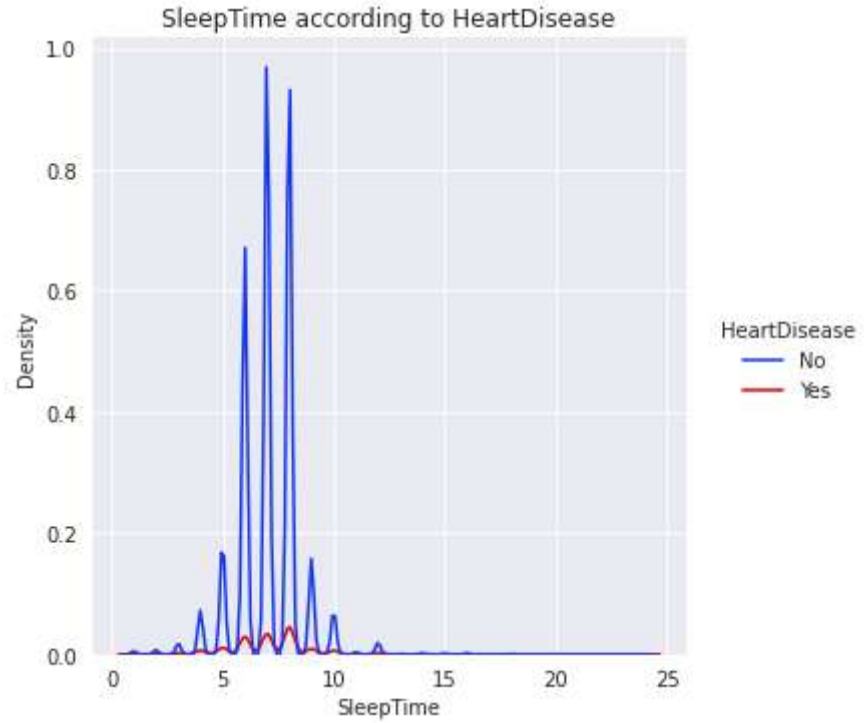


> Adults who reported doing physical activity or exercise during the past 30 days other than their regular job have a heart disease compared to those who didn't make any physical activity. A larger percentage of heart disease patients are not physically active.

3.4.2 Is the distribution of sleep time among heart disease patients different?

```
show_relation(num_cols[3], 'HeartDisease')
```

```
<Figure size 1080x504 with 0 Axes>
```




```
relative = df.groupby('HeartDisease').SleepTime.value_counts(normalize=True).reset_index(name

plt.figure(figsize=(16, 6), dpi=80)
ax = sns.barplot(data=relative, x='SleepTime', y='Percentage', hue='HeartDisease', palette=co

ax.set_title("Percentage of Sleep Times by Heart Disease");
```

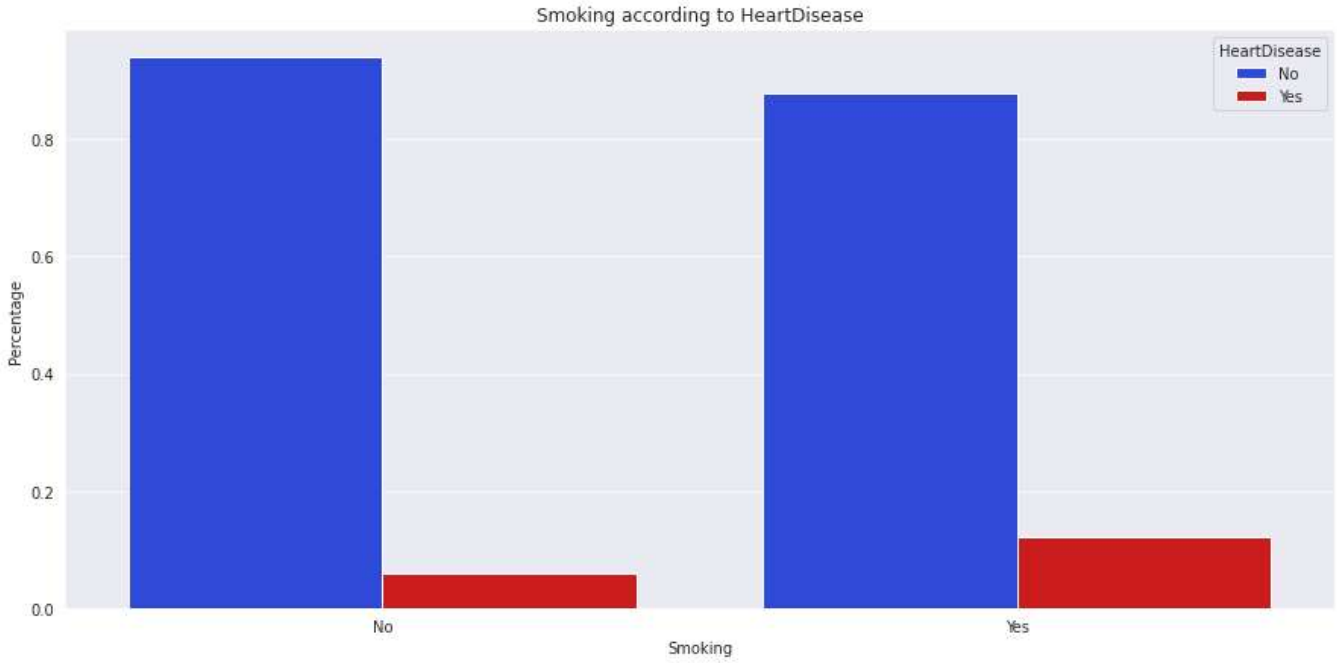
Abnormal sleeep duration is more prevalent in heart disease patients. Even though heart disease patients make 8.5% of the sample, they have higher percentages of sleep less than 6 hours or more than 9 hours, which is considered abnormal.

> As we can see `BMI`, `MentalHealth` and `SleepTime` according to heart disease have similar distributions. which means these columns didn't affect heart disease.

3.5 Substance Related Visualizations

3.5.1 Do people with heart disease smoke more?

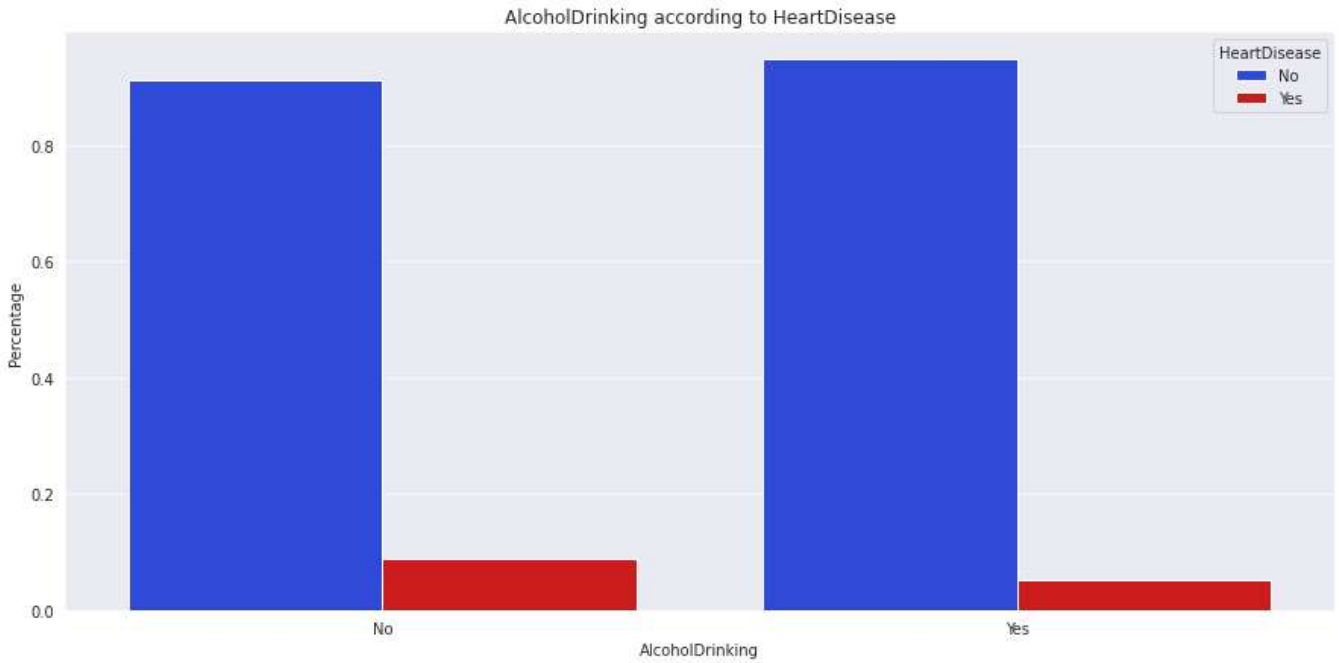
```
show_relation(obj_cols[0], 'HeartDisease', type_='count')
```



We can observe that the people who are smoking are more susceptible to the heart disease.

3.5.2 Do people with heart disease consume more alcohol?

```
show_relation(obj_cols[1], 'HeartDisease', type_='count')
```

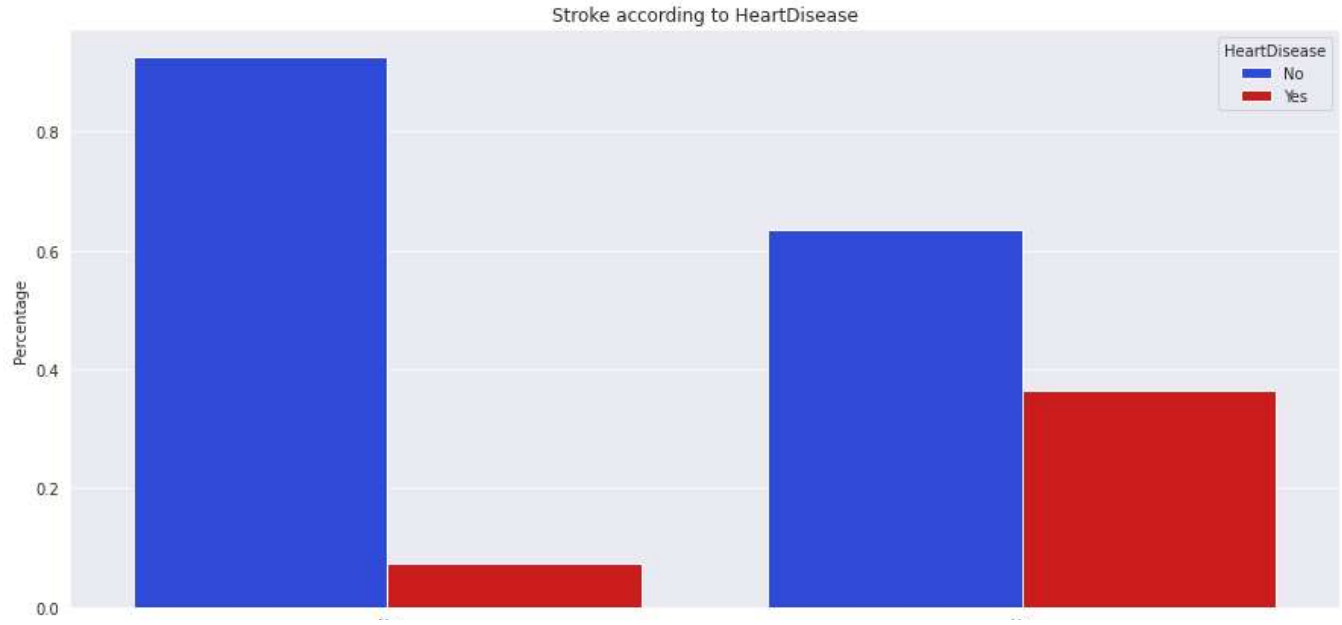


> People who are not drinking alchol, some of them have a heat disease

3.6 Other Diseases and Heart Disease

3.6.1 Does having a stroke affect the chances of heart disease?

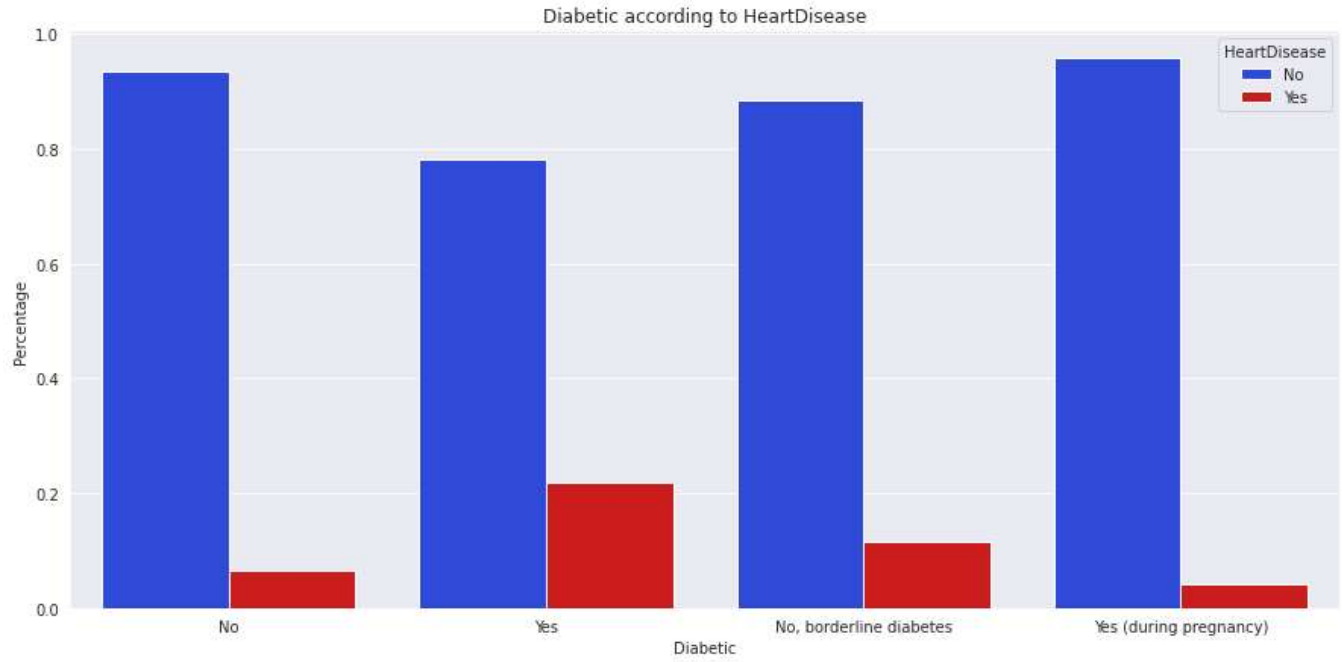
```
show_relation(obj_cols[2], 'HeartDisease', type_='count')
plt.savefig("stroke.png")
```



> Stroke is highly correlated with heart disease.

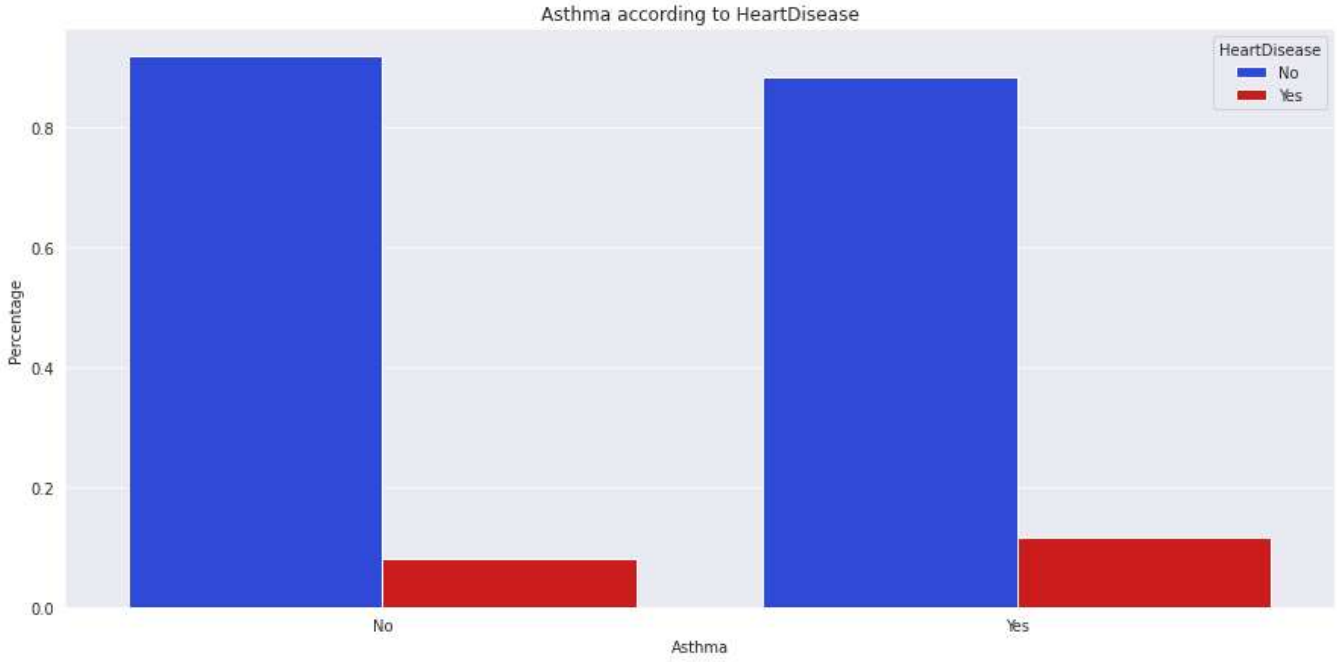
3.6.2 Does being diabetic increase the chances of heart disease?

```
show_relation(obj_cols[7], 'HeartDisease', type_='count')
plt.savefig("diabetes.png")
```



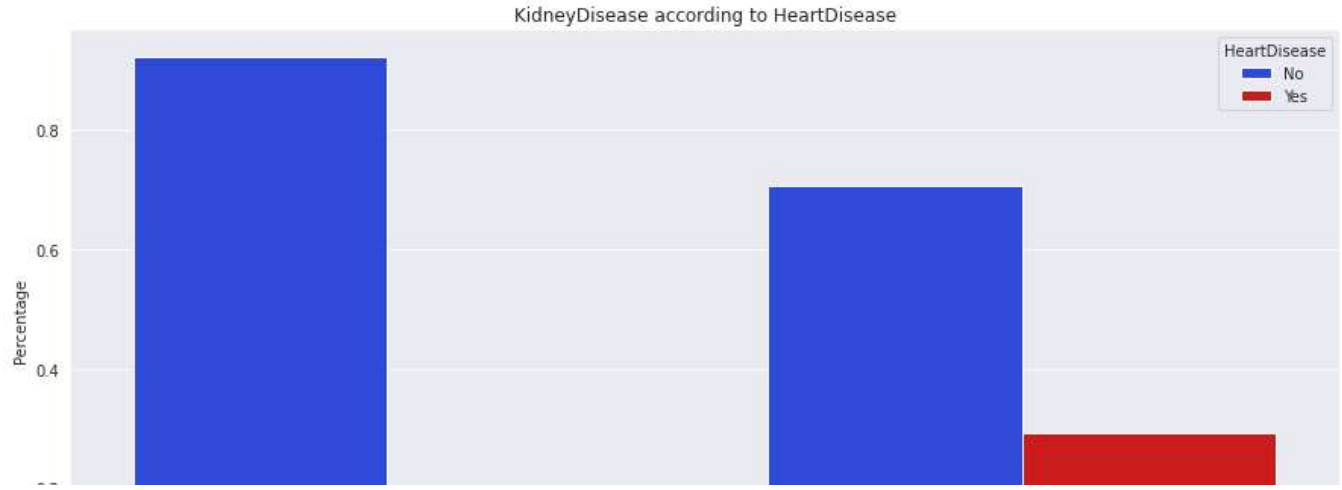
3.6.3 Do asthmatic people suffer more from heart diseases?

```
show_relation(obj_cols[10], 'HeartDisease', type_='count')
plt.savefig("Asthma.png")
```

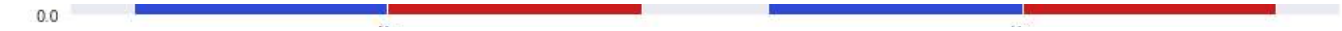


3.6.4 Does kidney disease coincide with heart disease?

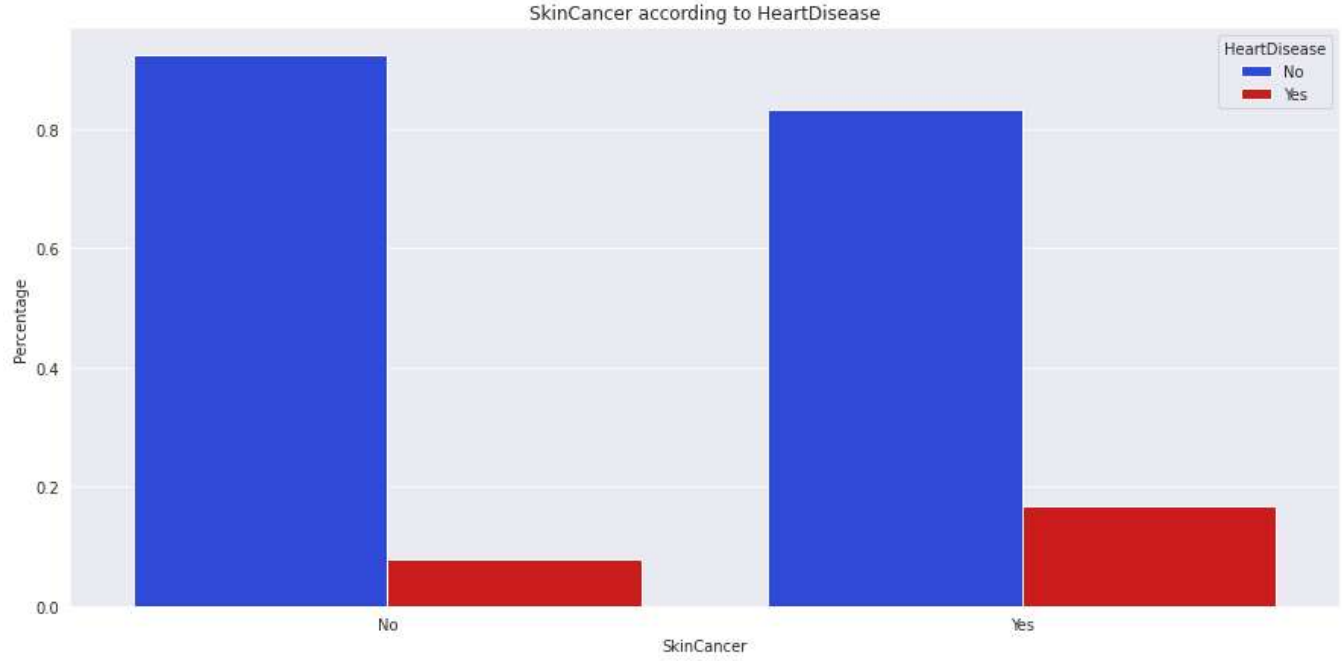
```
show_relation(obj_cols[11], 'HeartDisease', type_='count')
plt.savefig("Kidney.png")
```



3.6.5 Do people who suffer from skin cancer also suffer from heart disease?



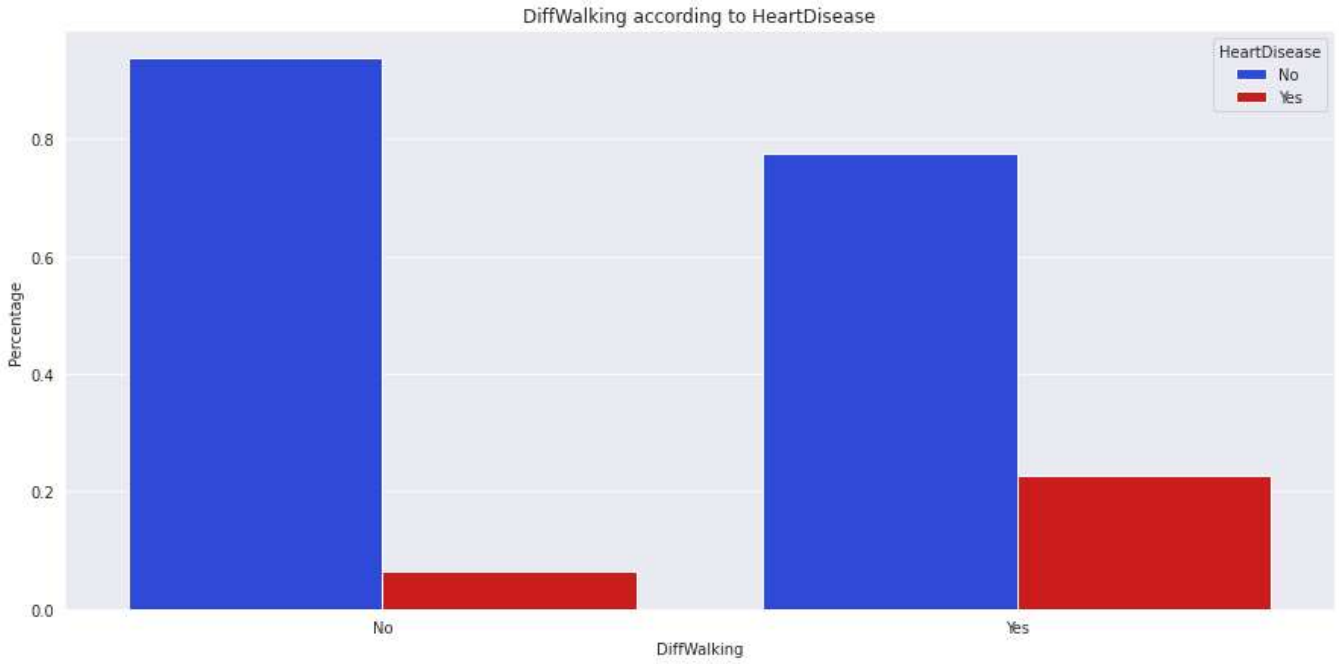

```
show_relation(obj_cols[12], 'HeartDisease', type_='count')
plt.savefig("Skin.png")
```



3.7 Special Circumstances and Heart Disease

3.7.1 Does having difficulty walking affect heart disease?

```
show_relation(obj_cols[3], 'HeartDisease', type_='count')
plt.savefig("difficult.png")
```



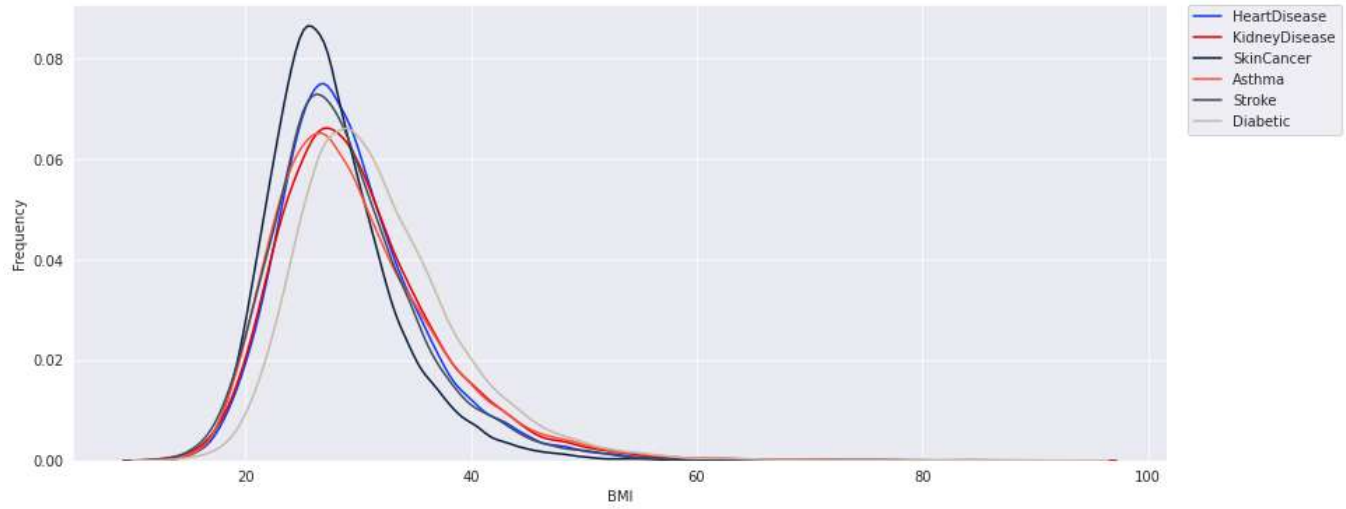
> the number of people who have no difficulty in walking and have a heart disease are bigger than the number of people who have difficulty in walking

3.8 Other Questions

3.8.1 Does BMI differ across diseases?

```
fig, ax = plt.subplots(figsize = (14,6))
sns.kdeplot(df[df["HeartDisease"]=='Yes']["BMI"], alpha=1,shade = False, color=colors6[0], la
sns.kdeplot(df[df["KidneyDisease"]=='Yes']["BMI"], alpha=1,shade = False, color=colors6[1], l
sns.kdeplot(df[df["SkinCancer"]=='Yes']["BMI"], alpha=1,shade = False, color=colors6[2], labe
sns.kdeplot(df[df["Asthma"]=='Yes']["BMI"], alpha=1,shade = False, color=colors6[3], label="A
sns.kdeplot(df[df["Stroke"]=='Yes']["BMI"], alpha=1,shade = False, color=colors6[4], label="S
sns.kdeplot(df[df["Diabetic"]=='Yes']["BMI"], alpha=1,shade = False, color=colors6[5], label=
```

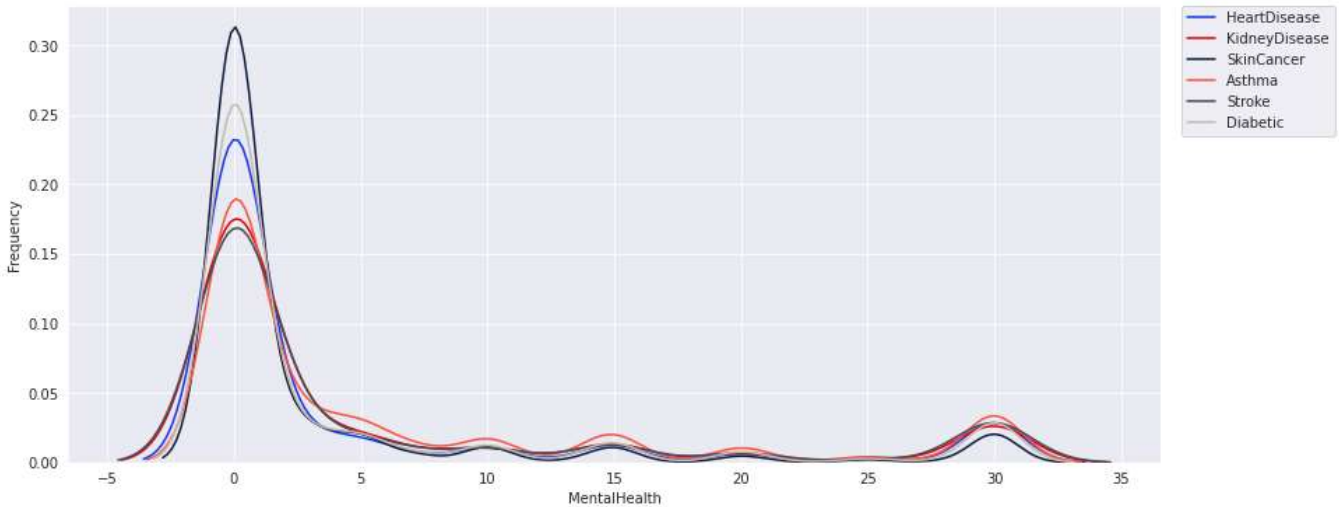
```
ax.set_xlabel("BMI")
ax.set_ylabel("Frequency")
ax.legend(bbox_to_anchor=(1.02, 1), loc=2, borderaxespad=0.)
plt.show()
```



3.8.2 Do different diseases impact mental health differently?

```
fig, ax = plt.subplots(figsize = (14,6))
sns.kdeplot(df[df["HeartDisease"]=='Yes']["MentalHealth"], alpha=1,shade = False, color=color
sns.kdeplot(df[df["KidneyDisease"]=='Yes']["MentalHealth"], alpha=1,shade = False, color=colo
sns.kdeplot(df[df["SkinCancer"]=='Yes']["MentalHealth"], alpha=1,shade = False, color=colors6
sns.kdeplot(df[df["Asthma"]=='Yes']["MentalHealth"], alpha=1,shade = False, color=colors6[3],
sns.kdeplot(df[df["Stroke"]=='Yes']["MentalHealth"], alpha=1,shade = False, color=colors6[4],
sns.kdeplot(df[df["Diabetic"]=='Yes']["MentalHealth"], alpha=1,shade = False, color=colors6[5
```

```
ax.set_xlabel("MentalHealth")
ax.set_ylabel("Frequency")
ax.legend(bbox_to_anchor=(1.02, 1), loc=2, borderaxespad=0.)
plt.show()
```

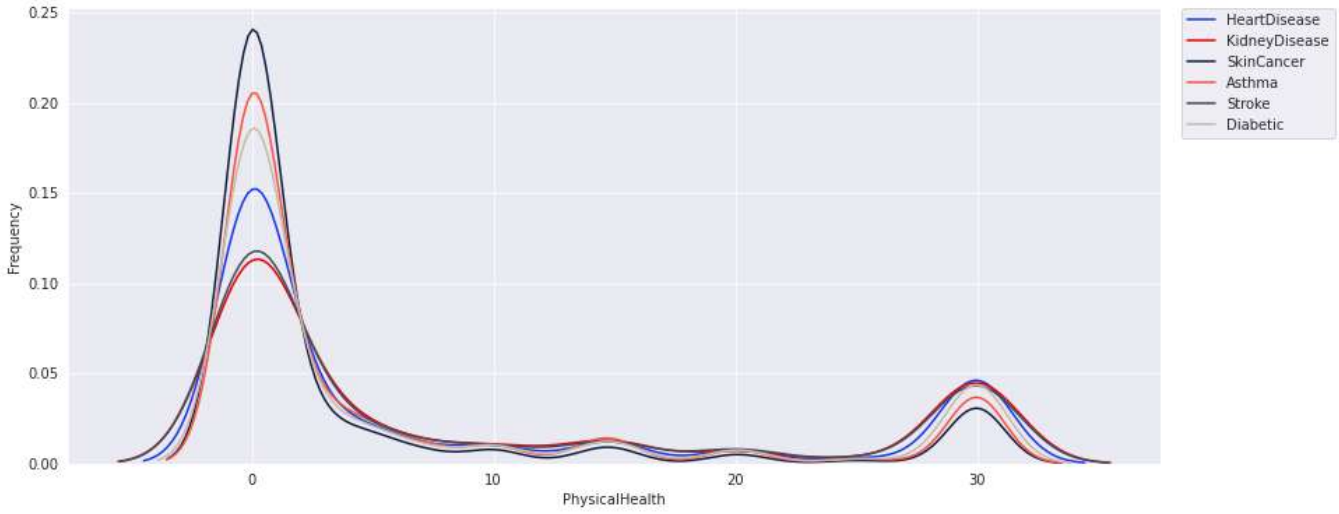


3.8.3 What is the effect of different diseases on sleep times?

```
fig, ax = plt.subplots(figsize = (14,6))
sns.kdeplot(df[df["HeartDisease"]=='Yes']["SleepTime"], alpha=1,shade = False, color=colors6[
sns.kdeplot(df[df["KidneyDisease"]=='Yes']["SleepTime"], alpha=1,shade = False, color=colors6
sns.kdeplot(df[df["SkinCancer"]=='Yes']["SleepTime"], alpha=1,shade = False, color=colors6[2]
sns.kdeplot(df[df["Asthma"]=='Yes']["SleepTime"], alpha=1,shade = False, color=colors6[3], la
sns.kdeplot(df[df["Stroke"]=='Yes']["SleepTime"], alpha=1,shade = False, color=colors6[4], la
sns.kdeplot(df[df["Diabetic"]=='Yes']["SleepTime"], alpha=1,shade = False, color=colors6[5],


ax.set_xlabel("SleepTime")
ax.set_ylabel("Frequency")
ax.legend(bbox_to_anchor=(1.02, 1), loc=2, borderaxespad=0.)
plt.show()
```

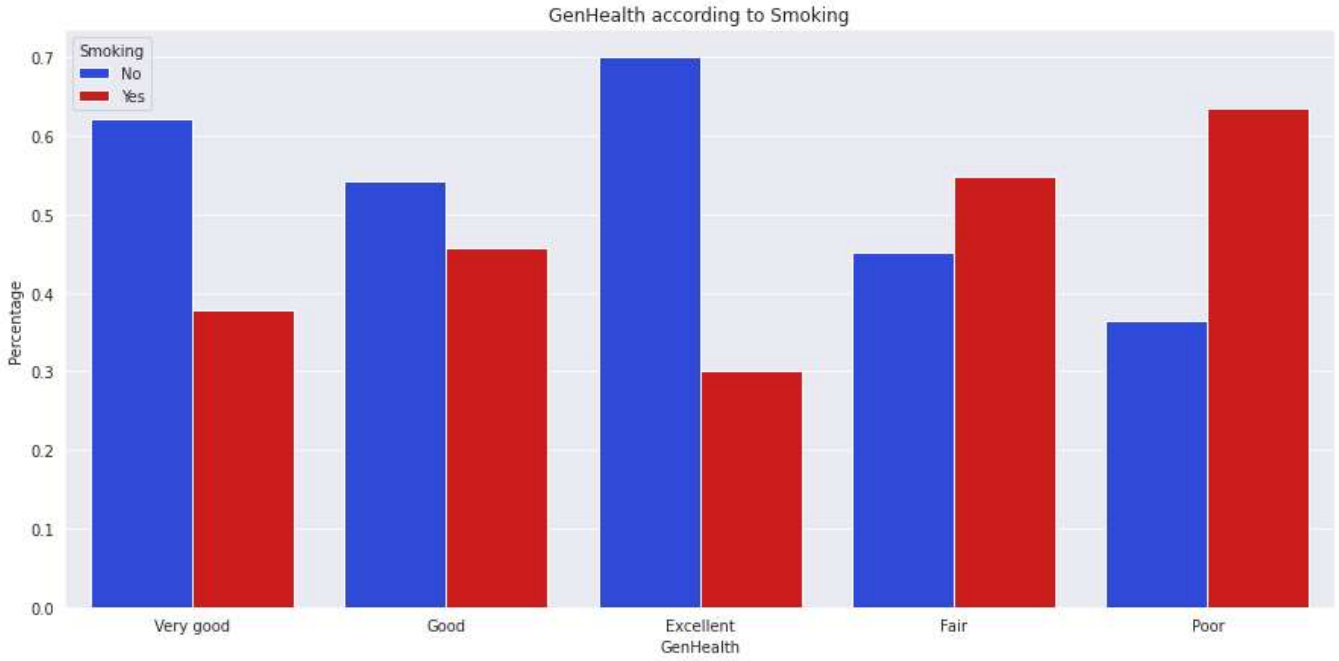
3.8.4 How different is the physical health across different diseases?

```
fig, ax = plt.subplots(figsize = (14,6))
sns.kdeplot(df[df["HeartDisease"]=='Yes']["PhysicalHealth"], alpha=1,shade = False, color=col
sns.kdeplot(df[df["KidneyDisease"]=='Yes']["PhysicalHealth"], alpha=1,shade = False, color=co
sns.kdeplot(df[df["SkinCancer"]=='Yes']["PhysicalHealth"], alpha=1,shade = False, color=color
sns.kdeplot(df[df["Asthma"]=='Yes']["PhysicalHealth"], alpha=1,shade = False, color=colors6[3
sns.kdeplot(df[df["Stroke"]=='Yes']["PhysicalHealth"], alpha=1,shade = False, color=colors6[4
sns.kdeplot(df[df["Diabetic"]=='Yes']["PhysicalHealth"], alpha=1,shade = False, color=colors6


ax.set_xlabel("PhysicalHealth")
ax.set_ylabel("Frequency")
ax.legend(bbox_to_anchor=(1.02, 1), loc=2, borderaxespad=0.)
plt.show()
```



3.8.5 Are smokers satisfied with their health?

```
show_relation('GenHealth', 'Smoking', 'count')
```



> The strange thing is the people who are smoking said that they have good and very heath

3.9 Insights Summary

Insights and takeaways drawn from data exploration:

1. In our sample, around 8 among 100 individuals suffer from heart disease.
2. The BMI of heart disease patients is slightly higher than that of healthy individuals.
3. The older the individual, the more susceptible they are to heart disease.
4. ~10% of males suffer from heart disease, while only ~7% of females do.
5. The pecentage of heart disease is highest (> 10%) among Native americans, followed by whites (~9%). The least percentage of heart disease (~3%) is among asians.
6. A lot more people who suffer from heart disease say they have poor or fair health compared to those who don't.
7. 79% of healthy individuals have been physically active in the past 30 days, compared to 64% in heart disease patients.
8. Abnormal sleeep duration is more prevalent in heart disease patients. Even though heart disease patients make 8.5% of the sample, they have higher percentages of sleep less than 6 hours or more than 9 hours, which is considered abnormal.

9. ~12% of people who smoke suffer from heart disease. In contrast, ~5% of non-smokers suffer from heart disease.
10. Surprisingly, people who drink alcohol have a lower percentage of heart disease (~4%) than those who do not (~9%).
11. Having a stroke is highly correlated with heart disease. People who have had a stroke before have a heart disease percentage of around 48%. On the other hand, people who did not suffer a stroke had a significantly lower percentage of heart disease (~8%).
12. Diabetic people are at higher risk of heart disease (~25%).
13. Asthmatic people are at a slightly higher risk of heart disease.
14. Those who have suffered from kidney disease are at a sginificantly higher risk of heart disease. With a percentage of ~30% compared to ~9% in healthy people.
15. People who suffered from skin cancer are at a moderately higher risk of heart disease (~18% vs ~9%).
16. Difficilty of walking is present in ~18% of heart disease patients vs ~7% in healthy individuals.
17. The BMI distribution differs slightly in patients of different diseases. With diabetic people having the highest BMI mode, and stroke victims having the lowest BMI mode.
18. Mental health, sleep duration, and physical health are similar among people who suffer from different dieseases.
19. ~64% of people who say they have poor health are smokers. While people who say they have excellent health are 30% smokers.

4 What is next?

In the second part, we will clean our data and fix most of the issues in our data prior to modeling. Then, we will try different models on our data and compare their results.

ACCESS PART 2 FROM HERE

Colab paid products - Cancel contracts here