Dataset Description

Dataset Ref: https://www.cdc.gov/brfss/annual_data/2020/pdf/codebook20_llcp-v2-508.pdf

```
from google.colab import drive
drive.mount('/content/drive')
```

```
Drive already mounted at /content/drive; to attempt to forcibly remount, call drive.mount("/content/drive", force_remoun
```

```
import pandas as pd
df = pd.read_csv('/content/drive/MyDrive/DM Project/Data/heart_2020_cleaned.csv')
```

```
display(df)
```

	HeartDisease	BMI	Smoking	AlcoholDrinking	Stroke	PhysicalHealth	MentalHealth	DiffWalking	Sex	AgeCategor
0	No	16.60	Yes	No	No	3	30	No	Female	55-5
1	No	20.34	No	No	Yes	0	0	No	Female	80 or olde
2	No	26.58	Yes	No	No	20	30	No	Male	65-6
3	No	24.21	No	No	No	0	0	No	Female	75-7
4	No	23.71	No	No	No	28	0	Yes	Female	40-4
...	...	...	...	...	...	...	...	...	...	.
319790	Yes	27.41	Yes	No	No	7	0	Yes	Male	60-6
319791	No	29.84	Yes	No	No	0	0	No	Male	35-3
319792	No	24.24	No	No	No	0	0	No	Female	45-4
319793	No	32.81	No	No	No	0	0	No	Female	25-2
319794	No	46.56	No	No	No	0	0	No	Female	80 or olde

319795 rows × 18 columns

```
df.shape
```

```
(319795, 18)
```

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

```
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
```

```
df['HeartDisease'].value_counts()
```

```
No     292422
Yes     27373
Name: HeartDisease, dtype: int64
```

Understanding the data

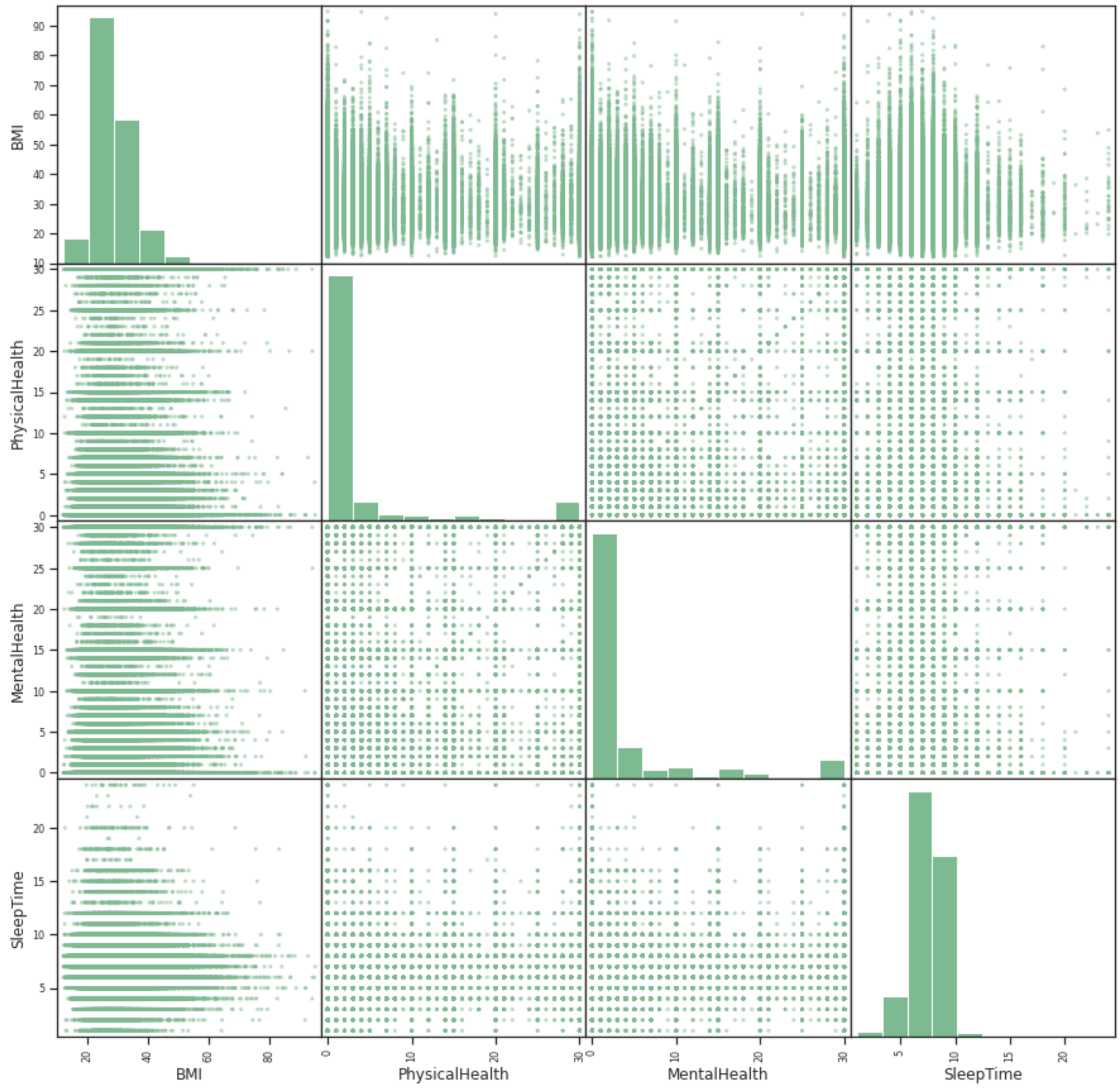
```
col_df = list(df.columns.values)
for column in col_df:
    print(column, ':', str(df[column].unique()))
```

```
HeartDisease : ['No' 'Yes']
BMI : [16.6  20.34 26.58 ... 62.42 51.46 46.56]
Smoking : ['Yes' 'No']
AlcoholDrinking : ['No' 'Yes']
Stroke : ['No' 'Yes']
PhysicalHealth : [ 3  0 20 28  6 15  5 30  7  1  2 21  4 10 14 18  8 25 16 29 27 17 24 12
 23 26 22 19  9 13 11]
MentalHealth : [30  0  2  5 15  8  4  3 10 14 20  1  7 24  9 28 16 12  6 25 17 18 21 29
 22 13 23 27 26 11 19]
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']
SleepTime : [ 5  7  8  6 12  4  9 10 15  3  2  1 16 18 14 20 11 13 17 24 19 21 22 23]
Asthma : ['Yes' 'No']
KidneyDisease : ['No' 'Yes']
SkinCancer : ['Yes' 'No']
```

Exploratory Data Analysis(EDA)

```
from matplotlib import pyplot as plt
import seaborn as sns
from pandas.plotting import scatter_matrix
```

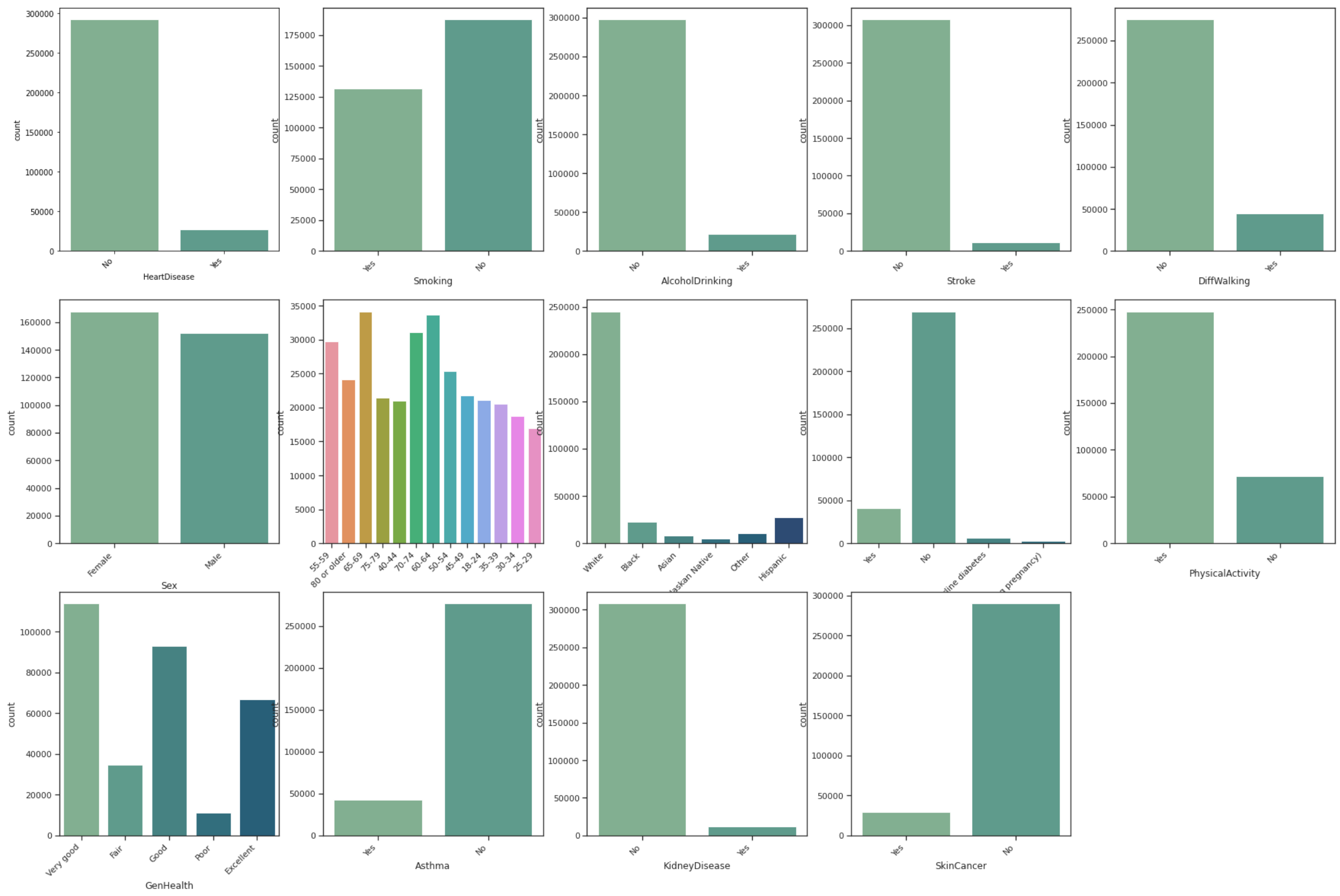
```
%matplotlib inline
scatter_mat = scatter_matrix(df, figsize=(15, 15))
```



```
features = df.select_dtypes(include=[object])
features.columns
```

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

```
def feature_div():
  plt.figure(figsize = (30,20))
  i = 1
  for feature in features:
      plt.subplot(3,5,i)
      sns.set(palette='crest')
      sns.set_style("ticks")
      ax = sns.countplot(x = feature, data = df)#, hue = 'pastel')#, color='#221C35')
      #Set the x-tick labels with list of string labels
      ax.set_xticklabels(ax.get_xticklabels(), rotation=45, ha="right")
      i +=1
feature_div()
```



Start coding or generate with AI.

Start coding or generate with AI.

Start coding or generate with AI.

References:

- https://www.cdc.gov/brfss/annual_data/2020/pdf/codebook20_llcp-v2-508.pdf
- https://digital.library.txstate.edu/bitstream/handle/10877/8132/GRITSENKO-THESIS-2019.pdf?isAllowed=y&sequence=1