DAV -assignment

INDICATORS OF HEART DISEASES
ROLL NO. 22BEI007, 22BEI043, 22BEI051

Problem Statement:

Heart disease is one of the leading causes of death globally, with numerous lifestyle and health factors contributing to its prevalence. The **Personal Key Indicators of Heart Disease** dataset contains 400,000+ records based on a 2020 survey from the CDC, focusing on the key indicators like high blood pressure, smoking, high cholesterol, diabetes, physical inactivity, and others. This dataset aims to identify key factors that influence the likelihood of individuals developing heart disease and enable targeted interventions.

Objective: To analyze the key health and lifestyle indicators associated with heart disease, identify trends or significant correlations between these indicators and the occurrence of heart disease, and suggest actionable insights for prevention strategies.

Sample Research Questions:

1. **Demographic Analysis**:
   - ○ What age group shows the highest incidence of heart disease? ○ How does the distribution of heart disease vary across different races or ethnicities?
   - ○ Is there a significant difference in heart disease prevalence between men and women?
2. **Health & Lifestyle Factors**:
   - ○ How does smoking correlate with heart disease prevalence? ○ Is there a significant relationship between high BMI and heart disease?
   - ○ What impact does physical activity (or the lack thereof) have on heart disease occurrence?
3. **Mental and Physical Health**:
   - ○ Do mental health issues, such as stress or anxiety, increase the likelihood of heart disease?
   - ○ How do physical health issues (e.g., stroke, difficulty walking) relate to heart disease occurrence?
4. **Comorbid Conditions**:

○ What is the correlation between diabetes and heart disease? ○ Does kidney disease increase the likelihood of heart disease? 5. **Preventive Measures**:

- ○ Are individuals who engage in regular physical activity less likely to develop heart disease?
- ○ How does alcohol consumption affect the chances of developing heart disease?

6. **Sleep and Heart Disease**:

- ○ What is the relationship between sleep duration and the likelihood of heart disease?

DATA PREPROCESSING

CODE

1.0.1 Data cleaning

[1]:
```
import numpy as np
import pandas as pd
from matplotlib import pyplot as plt
```

[2]:
```
df   pd.read_csv("heart_2022_with_nans.csv")
```

[3]:
```
df.head(10)
```

[3]:

	State	Sex	GeneralHealth	PhysicalHealthDays	MentalHealthDays	LastCheckupTime	PhysicalActivities
0	Alabama	Female	Very good	0.0	0.0	Within past year (anytime less than 12 months …	No
1	Alabama	Female	Excellent	0.0	0.0	NaN	No
2	Alabama	Female	Very good	2.0	3.0	Within past year (anytime less than 12 months …	Yes
3	Alabama	Female	Excellent	0.0	0.0	Within past year (anytime less than 12 months …	Yes
4	Alabama	Female	Fair	2.0	0.0	Within past year (anytime less than 12 months …	Yes
5	Alabama	Male	Poor	1.0	0.0	Within past year (anytime less than 12 months …	No
6	Alabama	Female	Very good	0.0	0.0	Within past year (anytime less than 12 months …	Yes
7	Alabama	Female	Good	0.0	0.0	Within past year (anytime less than 12 months …	No
8	Alabama	Female	Good	0.0	0.0	Within past year (anytime less than 12 months …	Yes
9	Alabama	Female	Good	1.0	0.0		

```
9 Within past year (anytime less than 12 months …         Yes
   SleepHours RemovedTeeth HadHeartAttack …
   HeightInMeters \
0        8.0      NaN No … NaN
1        6.0      NaN No … 1.60 2        5.0 NaN No …
          1.57 3 7.0 NaN No … 1.65 4 9.0  NaN No … 1.57
5        7.0     NaN Yes … 1.80
6        7.0     NaN No … 1.65 7  8.0 NaN No …
           1.63 8        6.0 NaN No … 1.70
9        7.0         NaN            No …          1.68

 WeightInKilogramsBMI AlcoholDrinkers HIVTesting FluVaxLast12 \
0             NaN     NaN No No Yes
1             68.04 26.57 No No No
2             63.50 25.61 No No No 3 63.50 23.30 No No
              Yes
4 53.98  21.77 Yes No No 5 84.82  26.08 No
No No 6 62.60 22.96 Yes No No
7             73.48 27.81 No No Yes
8             NaN     NaN No Yes No
9             81.65 29.05 Yes NaN Yes

 PneumoVaxEver                                TetanusLast10Tdap \
0     No Yes, received tetanus shot but not sure what
type 1     No No, did not receive any tetanus shot in the
pa…
2           No NaN
3           Yes No, did not receive any tetanus shot in
            the pa…
4           Yes No, did not receive any tetanus shot in
            the pa…
5           Yes No, did not receive any tetanus shot in
            the pa… 6 No No, did not receive any tetanus
            shot in the pa…
7     Yes Yes, received tetanus shot but not sure what
type 8     No Yes, received tetanus shot but not sure
what type
9       Yes No, did not receive any tetanus shot in the pa…

  HighRiskLastYear CovidPos
0     No    No
1     No    No 2  No    Yes
```

```
3              No       No
4              No       No
5              No       No
6              No       No
7              No       No
8              No       No
9              No       No
[10 rows x 40 columns]
```

Checking for null

[4]:
```
df.isnull().sum()
```

[4]:
```
State                          0
Sex                            0
GeneralHealth               1198
PhysicalHealthDays         10927
MentalHealthDays            9067
LastCheckupTime             8308
PhysicalActivities          1093
SleepHours                  5453
RemovedTeeth               11360
HadHeartAttack              3065
HadAngina                   4405
HadStroke                   1557
HadAsthma                   1773
HadSkinCancer               3143
HadCOPD                     2219
HadDepressiveDisorder       2812
HadKidneyDisease            1926
HadArthritis                2633
HadDiabetes                 1087
DeafOrHardOfHearing        20647
BlindOrVisionDifficulty    21564
DifficultyConcentrating    24240
DifficultyWalking          24012
DifficultyDressingBathing  23915
DifficultyErrands          25656
SmokerStatus               35462
ECigaretteUsage            35660
ChestScan                  56046
RaceEthnicityCategory      14057
AgeCategory                 9079
HeightInMeters             28652
WeightInKilograms          42078
BMI                        48806
AlcoholDrinkers            46574
HIVTesting                 66127
FluVaxLast12               47121
```

```
PneumoVaxEver            77040
TetanusLast10Tdap        82516
HighRiskLastYear         50623
CovidPos    50764 dtype: int64
```

Removing Nulls

[5]:
```
print("Initial shape of dataset - ",
df.shape,'\n') df =
df.dropna().reset_index(drop=True) print("New shape
after removing NaN's - ",df.shape,'\n')
df.isnull().sum()
```

```
Initial shape of dataset - (445132, 40)

New shape after removing NaN's - (246022, 40)
```

[5]:
```
State                    0
Sex                      0
GeneralHealth            0
PhysicalHealthDays       0
MentalHealthDays         0
LastCheckupTime          0
PhysicalActivities       0
SleepHours               0
RemovedTeeth             0
HadHeartAttack           0
HadAngina                0
HadStroke                0
HadAsthma                0
HadSkinCancer            0
HadCOPD                  0
HadDepressiveDisorder    0
HadKidneyDisease         0
HadArthritis             0
HadDiabetes              0
DeafOrHardOfHearing      0
BlindOrVisionDifficulty  0
DifficultyConcentrating  0
DifficultyWalking        0
```

```
DifficultyDressingBathing 0
 DifficultyErrands         0
 SmokerStatus              0
 ECigaretteUsage 0
 ChestScan                 0
 RaceEthnicityCategory     0
 AgeCategory               0
 HeightInMeters            0
 WeightInKilograms 0
 BMI                       0
 AlcoholDrinkers           0
 HIVTesting                0
 FluVaxLast12              0
 PneumoVaxEver             0
 TetanusLast10Tdap         0
 HighRiskLastYear          0
 CovidPos dtype:           0
 int64
```

checking for duplicates

[6]:
```
df duplicated() sum()
```

[6]:
```
np.int64(9)
```

Removing duplicates

[7]:
```
print('Shape before removing duplicates ',
df.shape,'\n') df.drop_duplicates(inplace=True)
print('Shape after removing duplicates ', df.shape,'\n')
```

```
Shape before removing duplicates (246022, 40)

Shape after removing duplicates (246013, 40)
```

1.0.2 Exploring dataset

[8]:
```
df info()
```

```
<class
 'pandas.core.frame.DataFrame'>
 Index: 246013 entries, 0 to
 246021 Data columns (total 40
 columns):
  # Column                   Non-Null Count Dtype
--- ------                   -------------- -----
 0   State  246013 non-null object
```

```
1   Sex    246013 non-null object
2   GeneralHealth     246013 non-null
                          object
3   PhysicalHealthDays     246013     non-null
    float64
4   MentalHealthDays  246013     non-null
                          float64
5   LastCheckupTime   246013 non-null object
6   PhysicalActivities     246013 non-null
    object
7   SleepHours  246013     non-null
                          float64
8   RemovedTeeth      246013 non-null object
9   HadHeartAttack    246013 non-null object
10  HadAngina    246013 non-null object
11  HadStroke    246013 non-null object
12  HadAsthma    246013 non-null object
13  HadSkinCancer     246013 non-null object
14  HadCOPD     246013 non-null object
15  HadDepressiveDisorder 246013 non-null object
16  HadKidneyDisease  246013 non-null object
17  HadArthritis      246013     non-null object
18  HadDiabetes 246013     non-null object
19  DeafOrHardOfHearing    246013     non-null
    object
20  BlindOrVisionDifficulty 246013    non-null
    object
21  DifficultyConcentrating 246013    non-null
    object
22  DifficultyWalking      246013     non-null
    object
23  DifficultyDressingBathing 246013 non-null
    object
 24 DifficultyErrands     246013 non-
                          null object
 25 SmokerStatus          246013 non-
                          null object
 26 ECigaretteUsage       246013 non-
                          null object
 27 ChestScan             246013 non-
                          null object
 28 RaceEthnicityCategory 246013 non-
                          null object
```

```
 29 AgeCategory           246013 non-
                          null object
 30 HeightInMeters        246013    non-null
                          float64
 31 WeightInKilograms     246013    non-null
                          float64
 32 BMI                   246013    non-null
                          float64
 33 AlcoholDrinkers       246013 non-null
                          object
 34 HIVTesting            246013 non-null
                          object
 35 FluVaxLast12          246013 non-null
                          object
 36 PneumoVaxEver         246013 non-null
                          object
 37 TetanusLast10Tdap     246013 non-null
                          object
 38 HighRiskLastYear      246013 non-null
                          object
 39   CovidPos   dtypes: 246013 non-
null float64(6),      object(34)
object memory usage: 77.0+ MB
```

Tranforming large strings into short ones for easy visualization

1. LastCheckupTime

[9]:
```
print(df['LastCheckupTime'].head
(5)) a = df['LastCheckupTime'].unique()
new_checkup_time = ['Recently','1Year','2Year','>2Year']
df['LastCheckupTime'] =
df['LastCheckupTime'].replace(a,new_checkup_time)
df['LastCheckupTime'].head(5)
```

```
0     Within past year (anytime less than 12 months …
1     Within past year (anytime less than 12 months …
2     Within past year (anytime less than 12 months …
3     Within past year (anytime less than 12 months …
4     Within past year (anytime less than 12 months …
Name: LastCheckupTime, dtype: object
```

[9]:
```
0 Recently 1
Recently
2   Recently
3   Recently
4   Recently
Name: LastCheckupTime, dtype: object
```

2. SmokerStatus

```
[10]: print(df['SmokerStatus'].head(5)) a = df['SmokerStatus'].unique()
new_smoker_status =
['Never','Somedays','Former','Everyday'] df['SmokerStatus'] =
     df['SmokerStatus'].replace(a,new_smoker_status)
df['SmokerStatus'].head(5)
```

```
0    Former smoker
1    Former smoker
2    Former smoker
3    Never smoked
4    Never smoked
Name: SmokerStatus, dtype: object
[10]: 0 Never 1
Never
2      Never
3      Somedays
4      Somedays
Name: SmokerStatus, dtype: object
```

3. ECigaretteUsage

```
[11]:
print(df['ECigaretteUsage'].head
(5)) a = df['ECigaretteUsage'].unique()
new_Esmoker_status = ['Never','Somedays','Former','Everyday']
df['ECigaretteUsage'] =
df['ECigaretteUsage'].replace(a,new_smoker_status)
df['ECigaretteUsage'].head(5)
```

```
0    Never used e-cigarettes in my entire life
1    Never used e-cigarettes in my entire life
2    Never used e-cigarettes in my entire life
3    Never used e-cigarettes in my entire life
4    Never used e-cigarettes in my entire life Name:
ECigaretteUsage, dtype: object
[11]: 0 Never 1
Never
2   Never
3   Never
4   Never
Name: ECigaretteUsage, dtype: object
```

4. RaceEthnicityCategory

[12]:
```
print(df['RaceEthnicityCategory'].hea
d(5)) a =
df['RaceEthnicityCategory'].unique()
new_Race = ['White','Black','Other','Multiracial','Hispanic']
df['RaceEthnicityCategory'] =
df['RaceEthnicityCategory'].replace(a,new_Race)
df['RaceEthnicityCategory'].head(5)
```

```
0    White only, Non-Hispanic
1    White only, Non-Hispanic
2    White only, Non-Hispanic
3    White only, Non-Hispanic
4    White only, Non-Hispanic
Name: RaceEthnicityCategory, dtype: object
```

[12]:
```
0 White 1
White
2   White
3   White
4   White
Name: RaceEthnicityCategory, dtype: object
```

5. AgeCategory

[13]:
```
print(df['AgeCategory'].hea
d(5))           a           =
df['AgeCategory'].unique() new_Age
=␣
 ↪['Old','Old','Old','Old','Adult','Adult','Adult','Adult','Adult','Adult','Young
,'Adult','Yo df['AgeCategory'] = df['AgeCategory'].replace(a,new_Age)
df['AgeCategory'].head(5)
```

```
0      Age 65 to 69
1      Age 70 to 74
2      Age 75 to 79
3      Age 80 or older
4      Age 80 or older
Name: AgeCategory, dtype: object
```

[13]:
```
0 Old 1
Old
2   Old
3   Old
4   Old
```

```
Name: AgeCategory, dtype: object
```

6. TetanusLast10Tdap

[14]:
```
print(df['TetanusLast10Tdap'].hea
d(5))              a              =
df['TetanusLast10Tdap'].unique()
new_Tetanus                       =
['Yes','Yes','No','No']
df['TetanusLast10Tdap']           =
df['TetanusLast10Tdap'].replace(a
,new_Tetanus) df['TetanusLast10Tdap'].head(5)
```

```
0                                 Yes, received Tdap
1                                 Yes, received tetanus shot
                                  but not sure what type
2                                 No, did not receive any
                                  tetanus shot in the pa…
3                                 No, did not receive any
                                  tetanus shot in the pa…
4                                 No, did not receive any
                                  tetanus shot in the pa…
Name: TetanusLast10Tdap, dtype: object
```

[14]:
```
0 Yes 1
Yes
2    No
3    No
4    No
Name: TetanusLast10Tdap, dtype: object
```

7. RemovedTeeth

[15]:
```
print(df['RemovedTeeth'].head(5)) a =
df['RemovedTeeth'].unique() new_Teeth =
['None','6orMore','1To5','All'] df['RemovedTeeth'] =
df['RemovedTeeth'].replace(a,new_Teeth)
df['RemovedTeeth'].head(5)
```

```
0                 None of them
1                 None of them
2                 6 or more, but not all
3                 None of them
4                 1 to 5
Name: RemovedTeeth, dtype: object
```

[15]:
```
0      None
1      None
2      6orMore
3      None
4      1To5
Name: RemovedTeeth, dtype: object
```

8. HadDiabetes

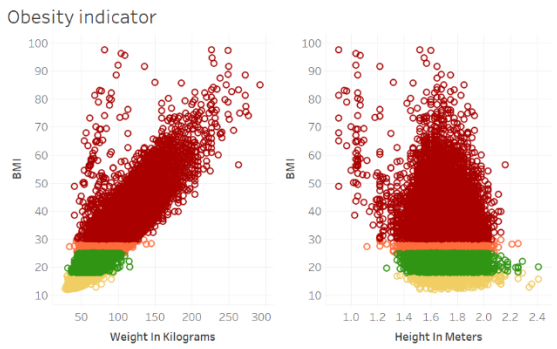
[16]:
```
print(df['HadDiabetes ].head(5))
a   df['HadDiabetes'].unique()
a
new_Teeth   ['No','Yes','Yes','No']
df['HadDiabetes']   df['HadDiabetes'].replace(a,new_Teeth)
df['HadDiabetes'].head(5)
```

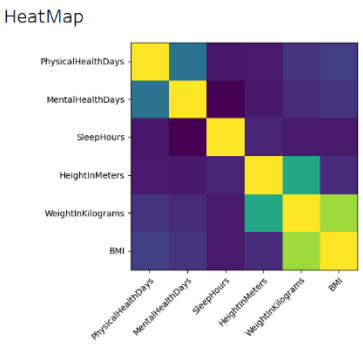
```
0     No
1     Yes
2     No
3     No
4     No
Name: HadDiabetes, dtype: object
```

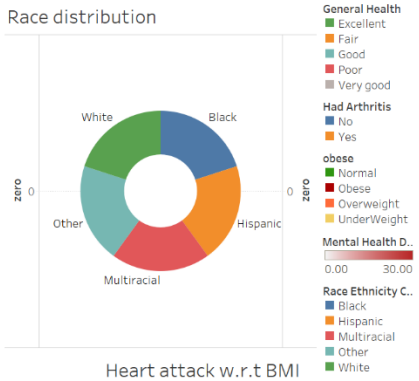
[16]:
```
0 No
1     Yes
2     No
3     No
4     No
Name: HadDiabetes, dtype: object
```

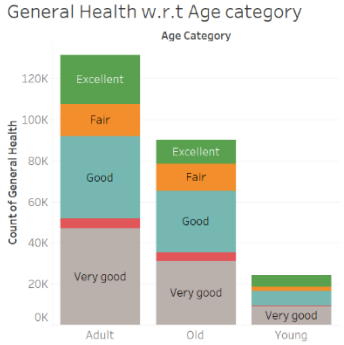
[17]:
```
df to_csv( heart_2022_cleaned.csv )
```

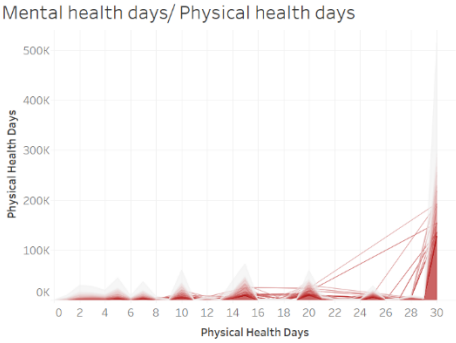
IMPLEMENTATION

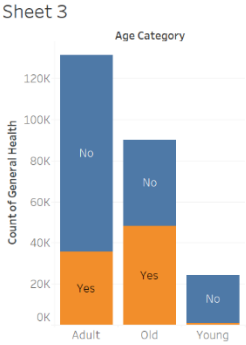


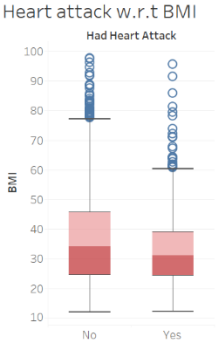


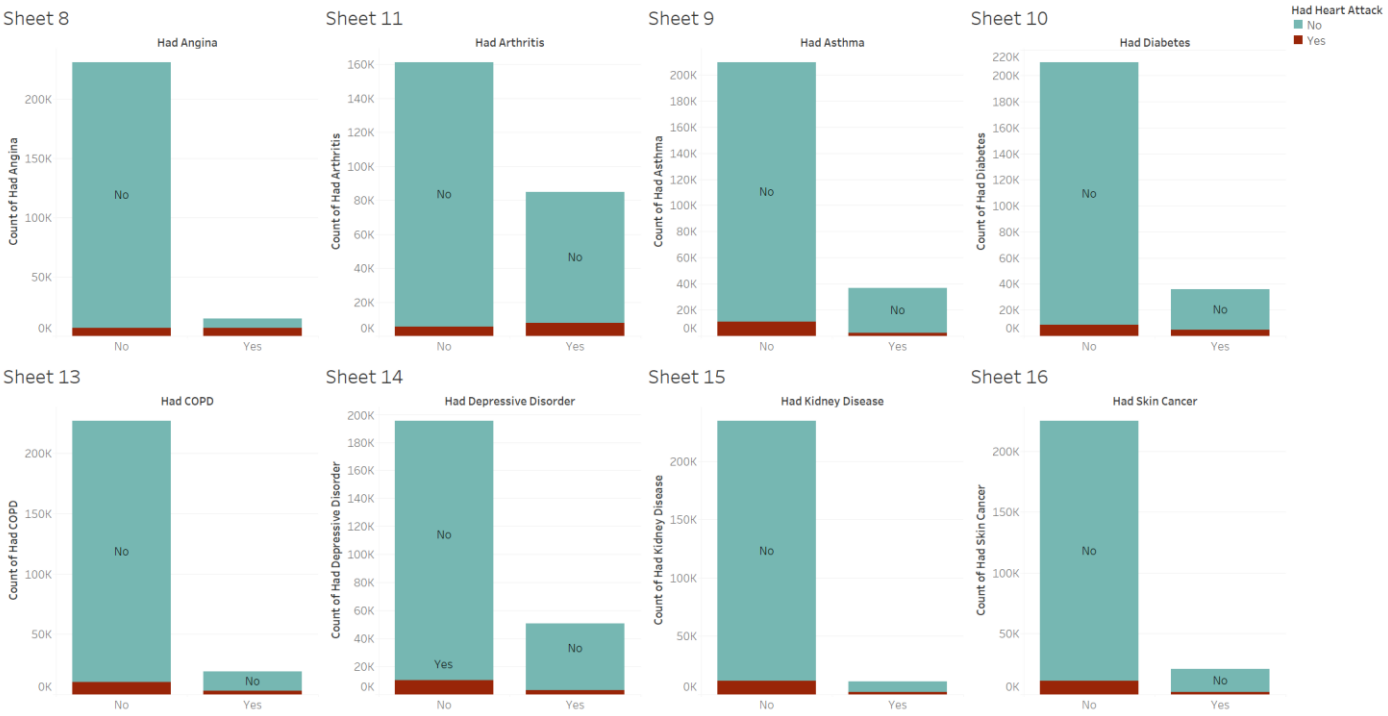











Sheet 8
Had Angina
Count of Had Angina
200K
150K
100K
50K
0K
No
No
Yes
Sheet 11
Had Arthritis
Count of Had Arthritis
160K
140K
120K
100K
80K
60K
40K
20K
0K
No
No
No
Yes
Sheet 9
Had Asthma
Count of Had Asthma
200K
180K
160K
140K
120K
100K
80K
60K
40K
20K
0K
No
No
No
Yes
Sheet 10
Had Diabetes
Count of Had Diabetes
220K
200K
180K
160K
140K
120K
100K
80K
60K
40K
20K
0K
No
No
No
Yes
Had Heart Attack
No
Yes
Sheet 13
Had COPD
Count of Had COPD
200K
150K
100K
50K
0K
No
No
No
Yes
Sheet 14
Had Depressive Disorder
Count of Had Depressive Disorder
200K
180K
160K
140K
120K
100K
80K
60K
40K
20K
0K
No
Yes
No
No
Yes
Sheet 15
Had Kidney Disease
Count of Had Kidney Disease
200K
150K
100K
50K
0K
No
No
No
Yes
Sheet 16
Had Skin Cancer
Count of Had Skin Cancer
200K
150K
100K
50K
0K
No
No
No
Yes