SAMSUNG

Samsung Innovation Campus

Artificial Intelligence Course 401

Together for Tomorrow!
Enabling People
Education for Future Generations

Personal Key Indicators of Heart Disease

Project presented by

Osama Ali & Mo’men Emad

Data Used

https://www.kaggle.com/datasets/kamilpytlak/personal-key-indicators-of-heart-disease

Facilitator

Haneen El daly

Supervised by

Doaa Mahmoud Abdel-Aty

Agenda

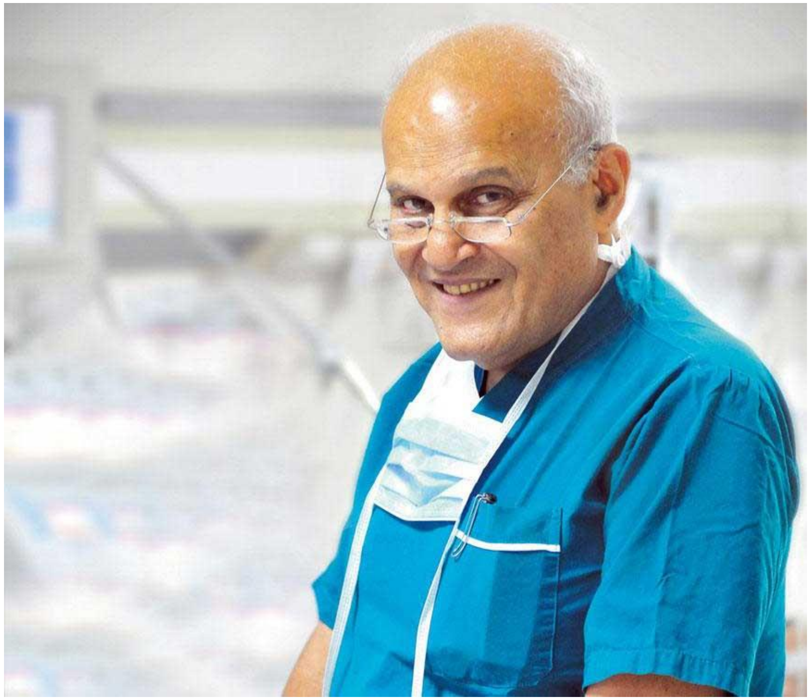
1. Introduction
2. Data description
3. EDA
4. Data preprocessing
5. Modeling
6. Evaluation

Introduction

Heart disease is one of the leading causes of death for people of most races in the US (African Americans, American Indians and Alaska Natives, and white people) According to the Center of disease control and prevention (CDC). About half of all Americans (47%) have at least 1 of 3 key risk factors for heart disease: high blood pressure, high cholesterol, and smoking. Other key indicator include diabetic status, obesity (high BMI), not getting enough physical activity or drinking too much alcohol. Detecting and preventing the factors that have the greatest impact on heart disease is very important in healthcare. Computational developments, in turn, allow the application of machine learning methods to detect "patterns" from the data that can predict a patient's condition.

About

- The dataset is a group of residents in the US
- It’s containing some information about Indicators of Heart Disease
- Our goal is to predict Heart Disease

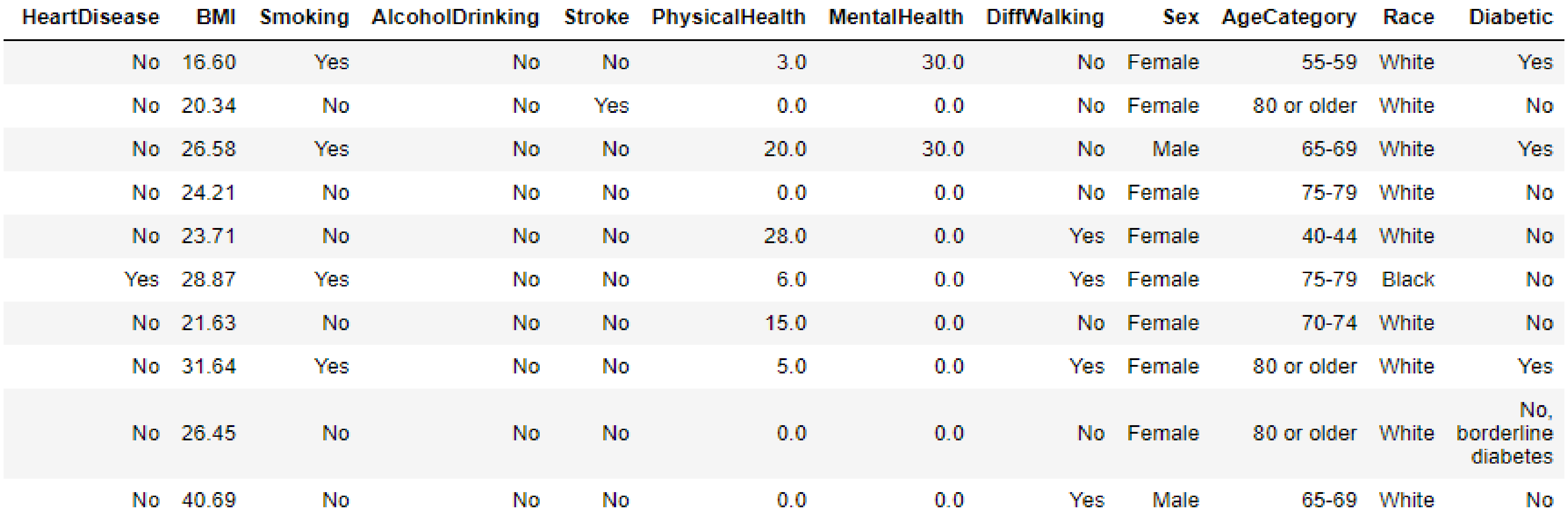


Data Description

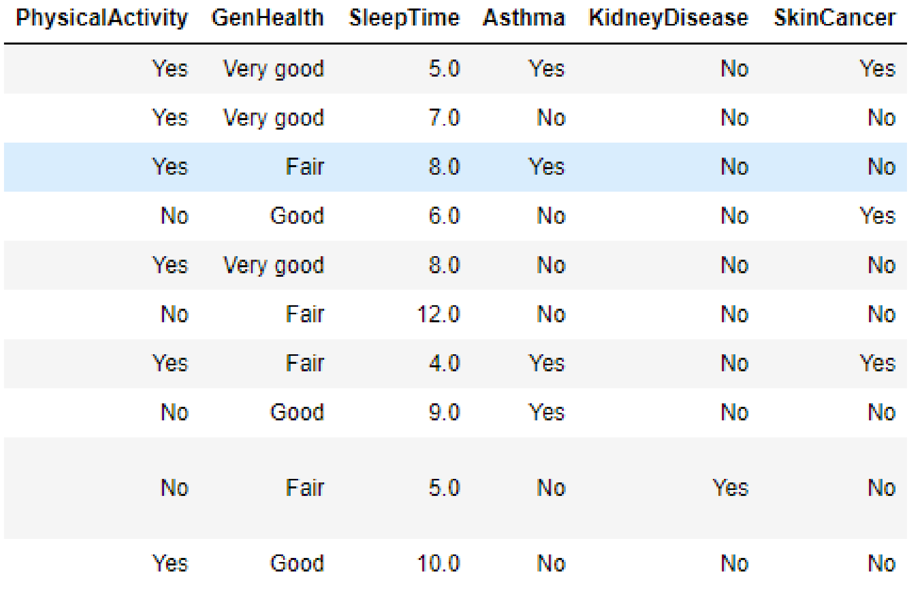
The dataset contains 18 variables (9 Booleans, 5 strings and 4 decimals). In machine learning projects, "HeartDisease" can be used as the explanatory variable but note that the classes are heavily unbalanced.

1. Heart disease: Respondents that have ever reported having coronary heart disease (CHD) or myocardial infarction (MI)
2. Smoking: Have you smoked at least 100 cigarettes in your entire life? [Note: 5 packs = 100 cigarettes]
3. AlcoholDrinking: Heavy drinkers (adult men having more than 14 drinks per week and adult women having more than 7 drinks per week
4. PhysicalHealth: Now thinking about your physical health, which includes physical illness and injury, for how many days during the past 30 days was your physical health not good? (0-30 days)
5. MentalHealth: Thinking about your mental health, for how many days during the past 30 days was your mental health not good? (0-30 days)
6. DiffWalking: Do you have serious difficulty walking or climbing stairs?
7. AgeCategory: Fourteen-level age category
8. Race: Imputed race/ethnicity value
9. PhysicalActivity: Adults who reported doing physical activity or exercise during the past 30 days other than their regular job
10. GenHealth: Would you say that in general your health is...
11. SleepTime:On average, how many hours of sleep do you get in a 24-hour period?
12. KidneyDisease: Not including kidney stones, bladder infection or incontinence, were you ever told you had kidney disease?

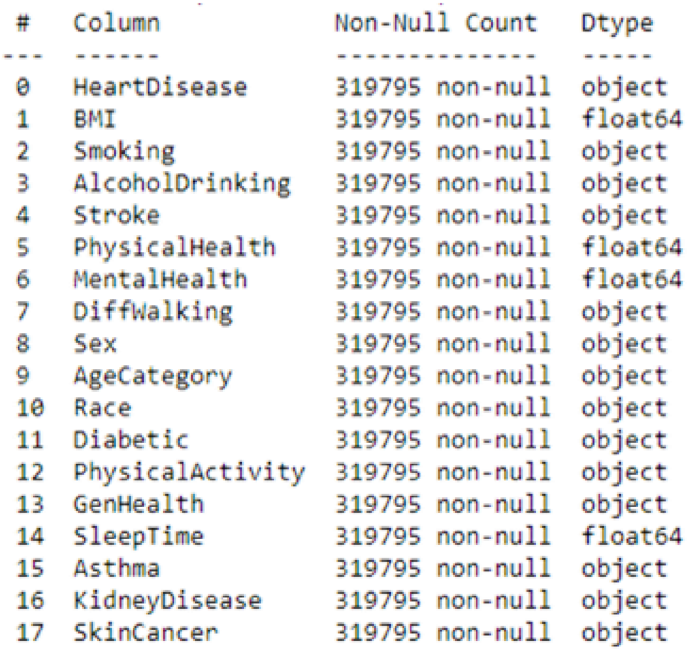
Data Sample



HeartDisease	BMI	Smoking	AlcoholDrinking	Stroke	PhysicalHealth	MentalHealth	DiffWalking	Sex	AgeCategory	Race	Diabetic
No	16.60	Yes	No	No	3.0	30.0	No	Female	55-59	White	Yes
No	20.34	No	No	Yes	0.0	0.0	No	Female	80 or older	White	No
No	26.58	Yes	No	No	20.0	30.0	No	Male	65-69	White	Yes
No	24.21	No	No	No	0.0	0.0	No	Female	75-79	White	No
No	23.71	No	No	No	28.0	0.0	Yes	Female	40-44	White	No
Yes	28.87	Yes	No	No	6.0	0.0	Yes	Female	75-79	Black	No
No	21.63	No	No	No	15.0	0.0	No	Female	70-74	White	No
No	31.64	Yes	No	No	5.0	0.0	Yes	Female	80 or older	White	Yes
No	26.45	No	No	No	0.0	0.0	No	Female	80 or older	White	No, borderline diabetes
No	40.69	No	No	No	0.0	0.0	Yes	Male	65-69	White	No

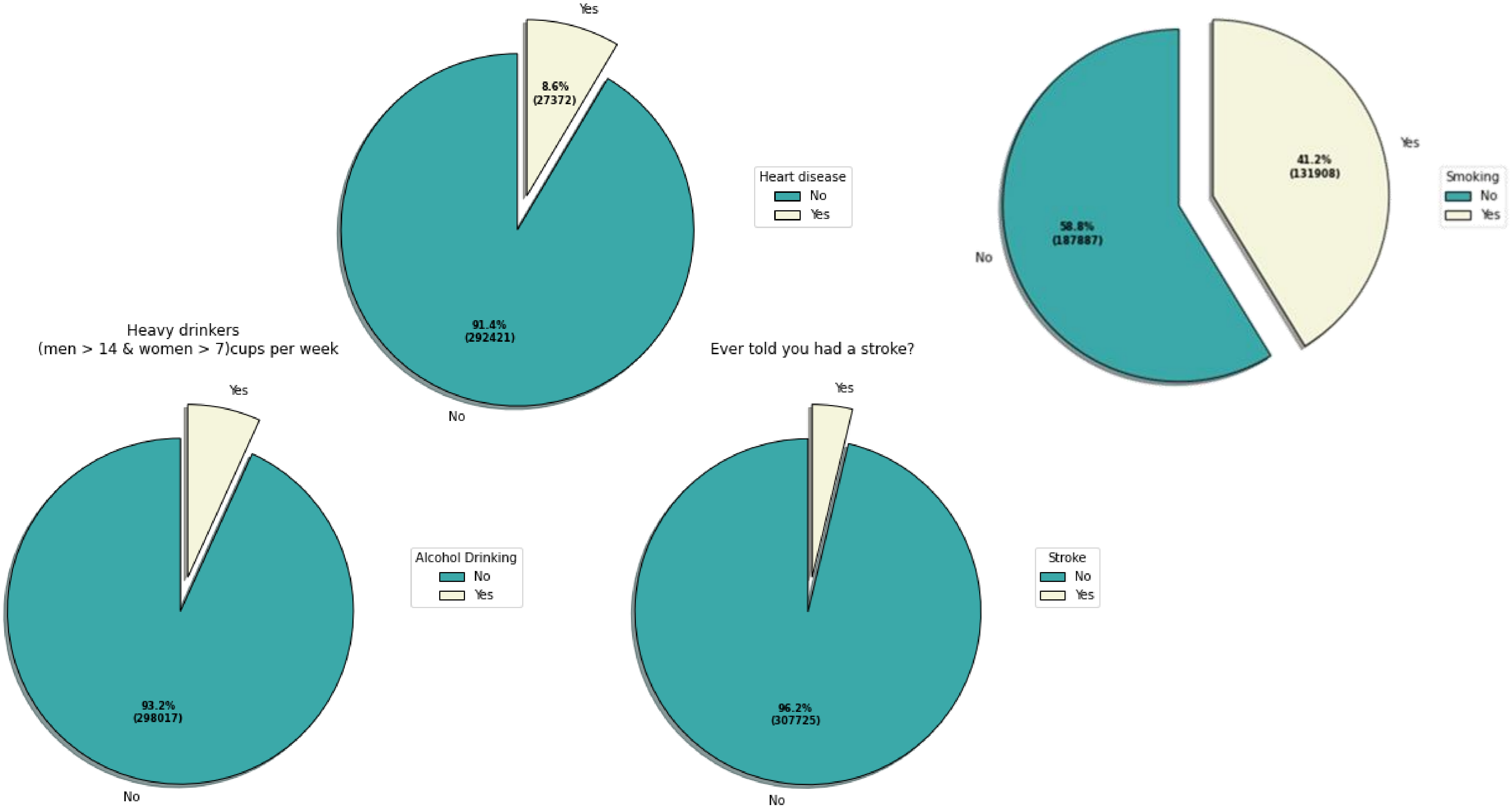


PhysicalActivity	GenHealth	SleepTime	Asthma	KidneyDisease	SkinCancer
Yes	Very good	5.0	Yes	No	Yes
Yes	Very good	7.0	No	No	No
Yes	Fair	8.0	Yes	No	No
No	Good	6.0	No	No	Yes
Yes	Very good	8.0	No	No	No
No	Fair	12.0	No	No	No
Yes	Fair	4.0	Yes	No	Yes
No	Good	9.0	Yes	No	No
No	Fair	5.0	No	Yes	No
Yes	Good	10.0	No	No	No



```
 #   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
```

Explore Data



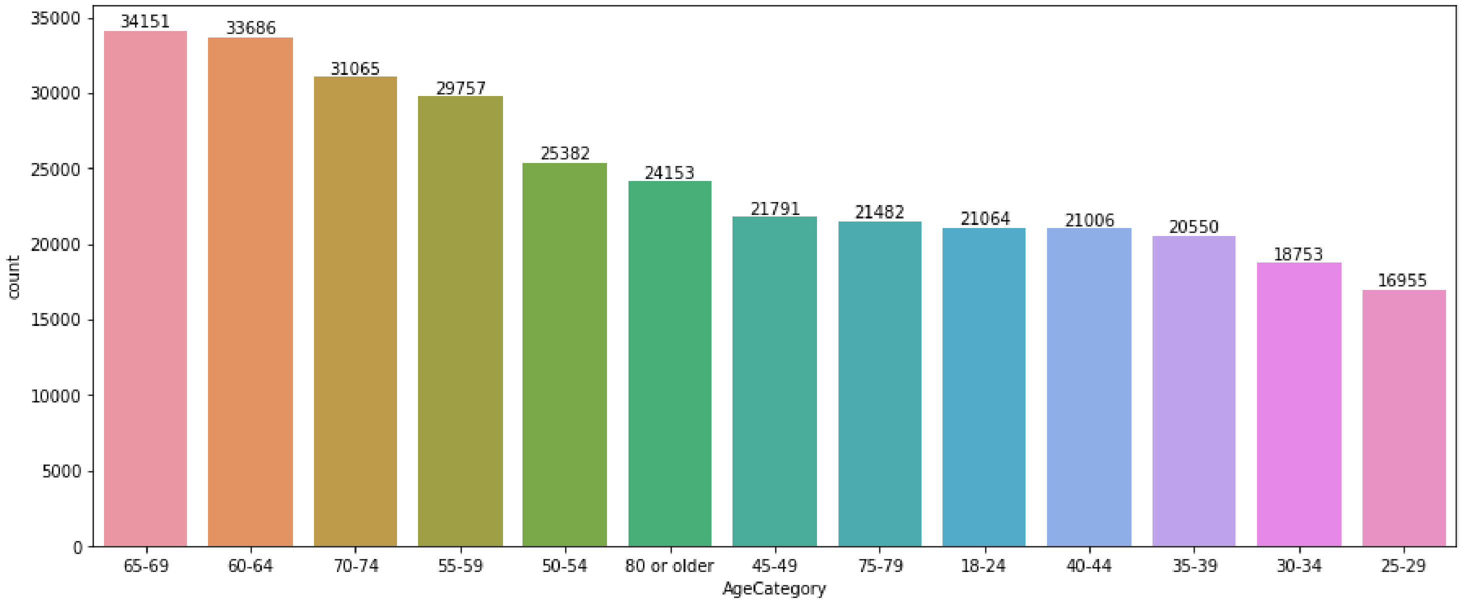
Have ever reported having coronary heart disease (CHD) or myocardial infarction (MI)?

Have you smoked at least 100 cigarettes in your entire life?

Heavy drinkers (men > 14 & women > 7)cups per week

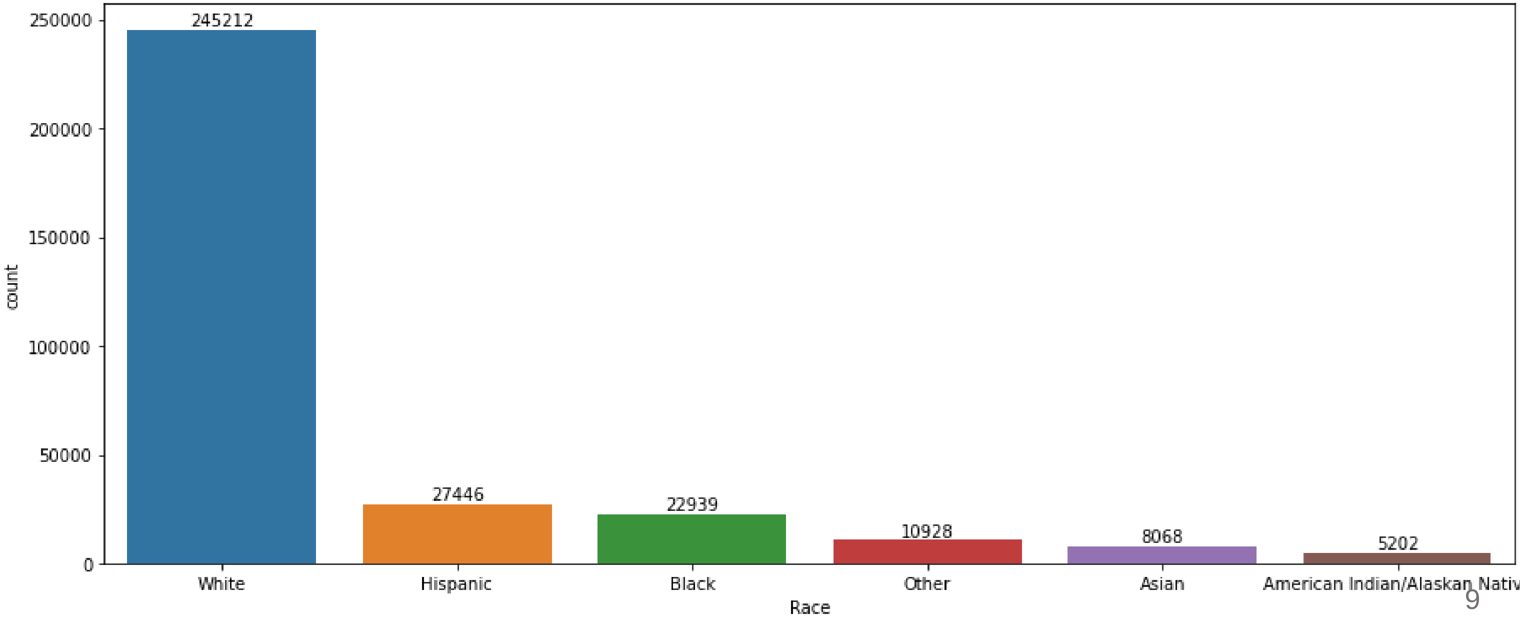
Ever told you had a stroke?

Explore Data

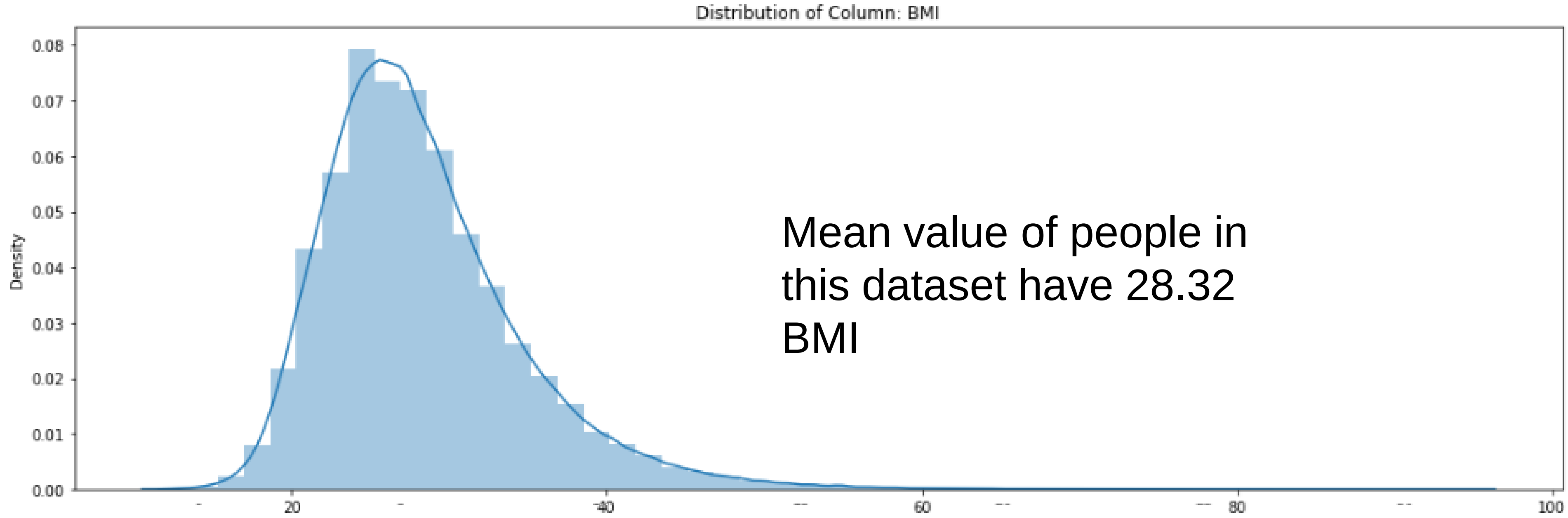


Most of dataset samples are older than 50

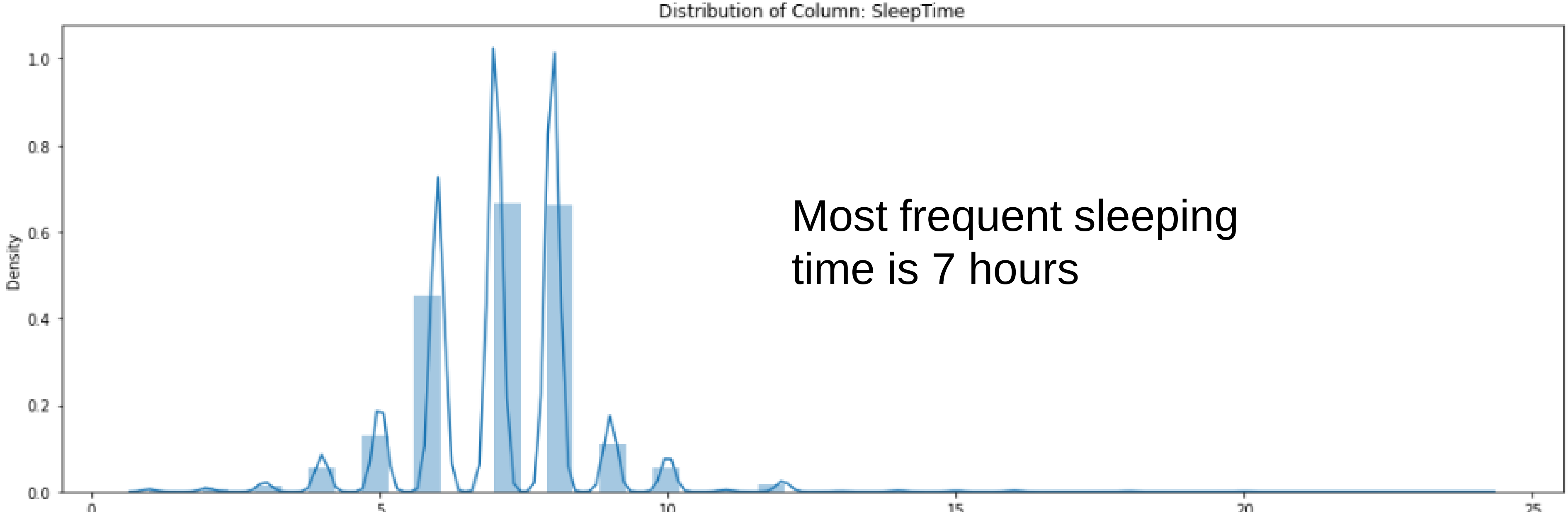
Most of samples are white people



Explore Data

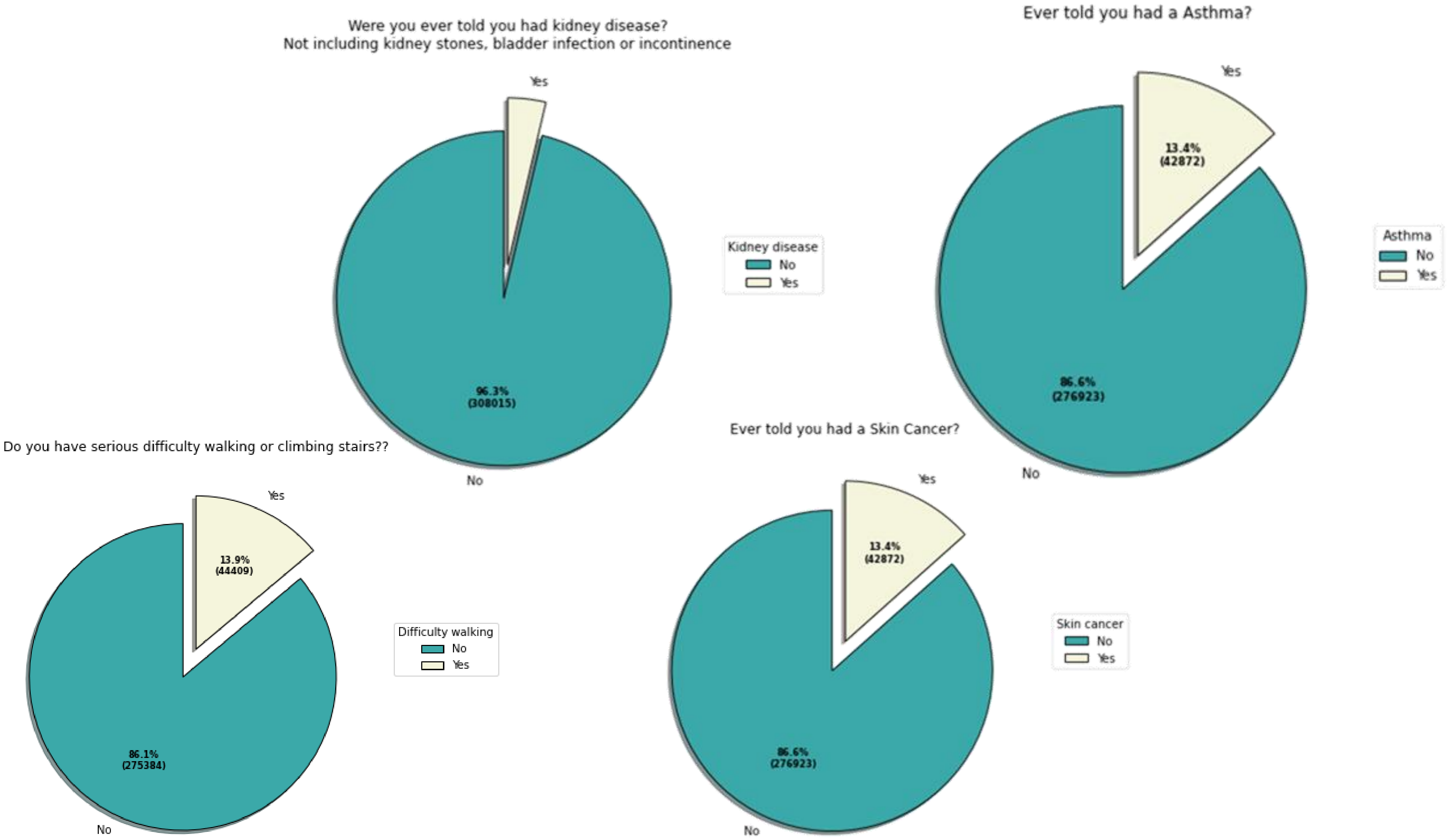


Mean value of people in this dataset have 28.32 BMI



Most frequent sleeping time is 7 hours

Explore Data



Were you ever told you had kidney disease?
Not including kidney stones, bladder infection or incontinence

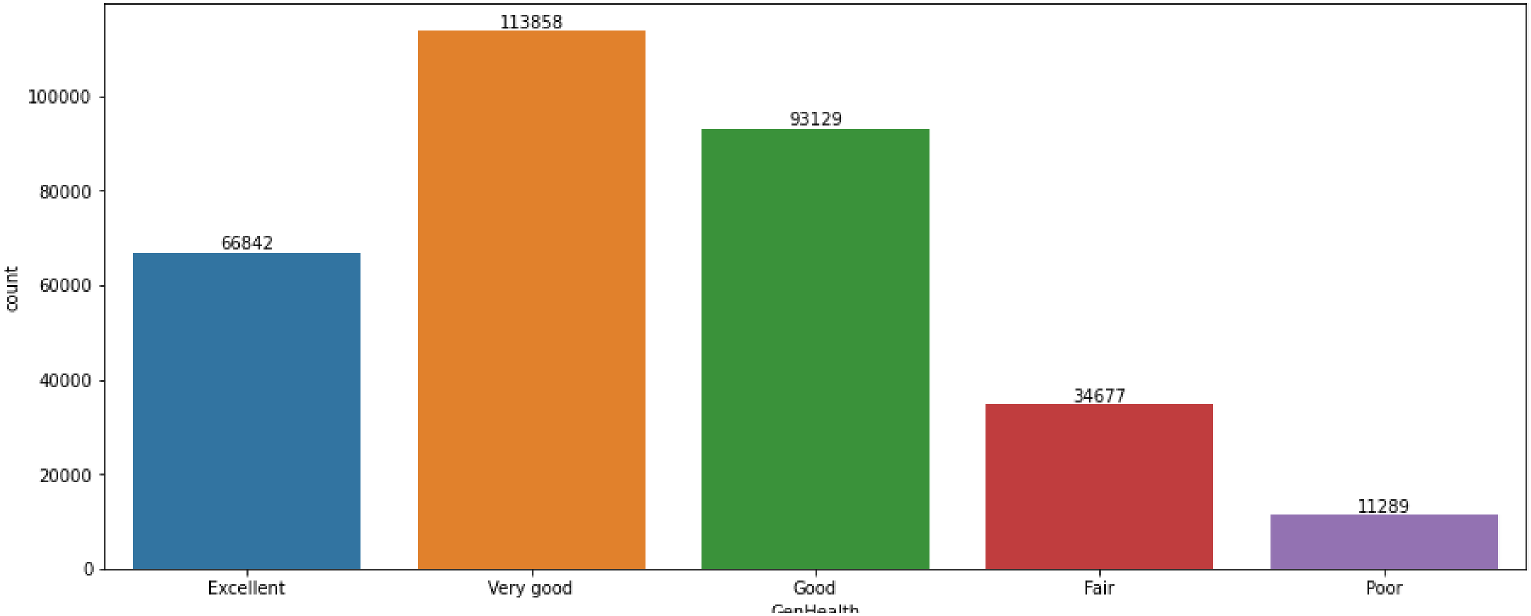
Ever told you had a Asthma?

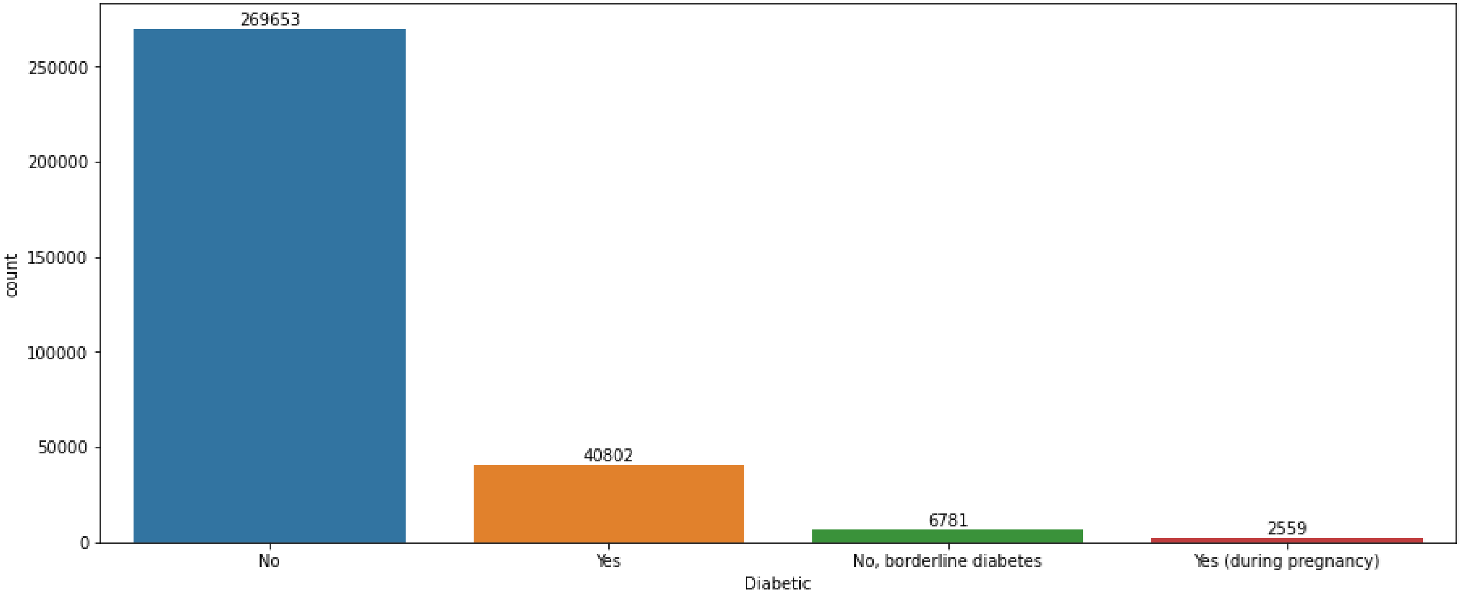
Do you have serious difficulty walking or climbing stairs??

Ever told you had a Skin Cancer?

Explore Data

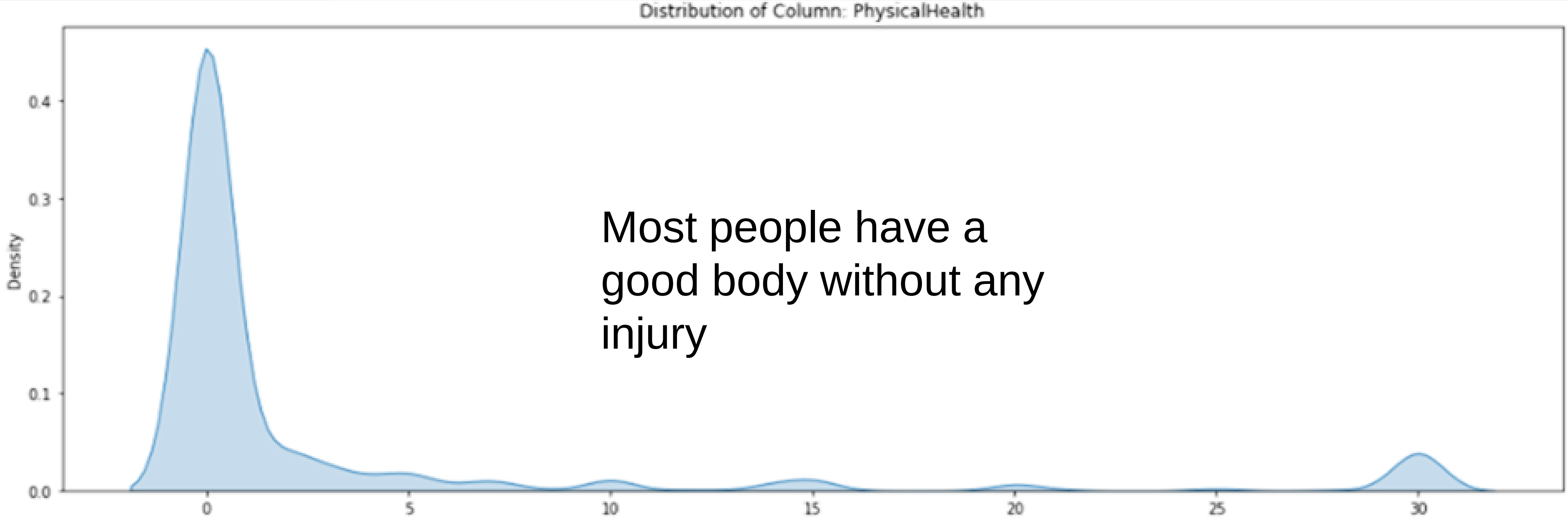
Most people in this dataset sample have above fair health



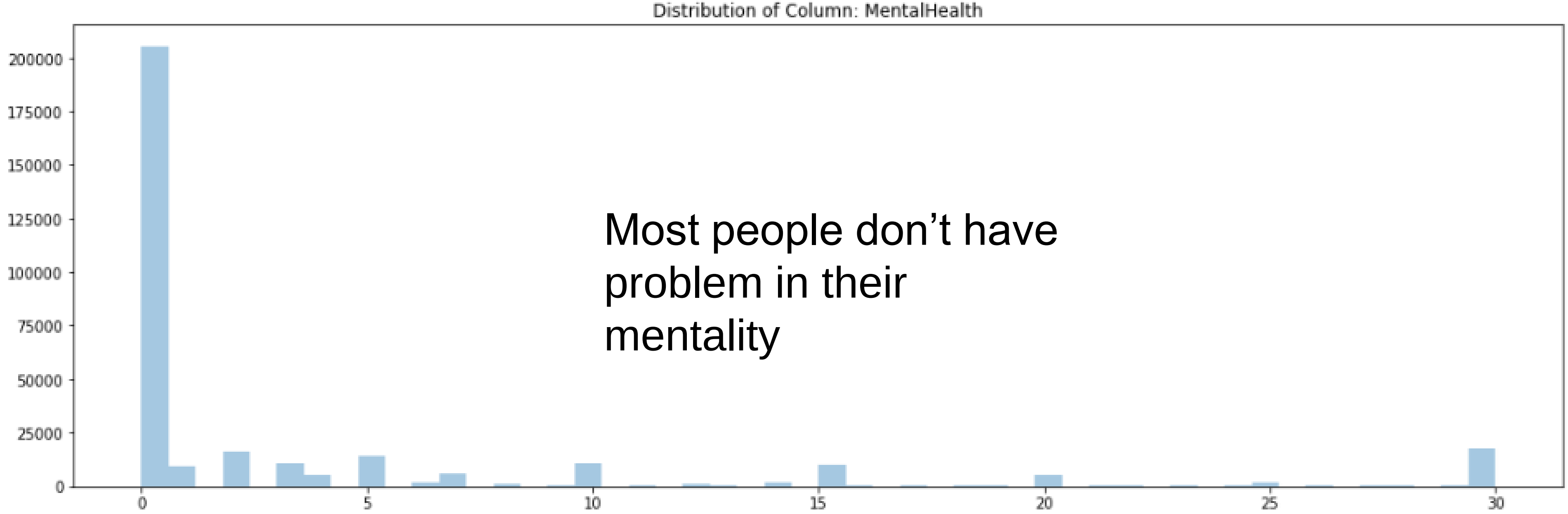


Most people don’t have diabetic

Explore Data

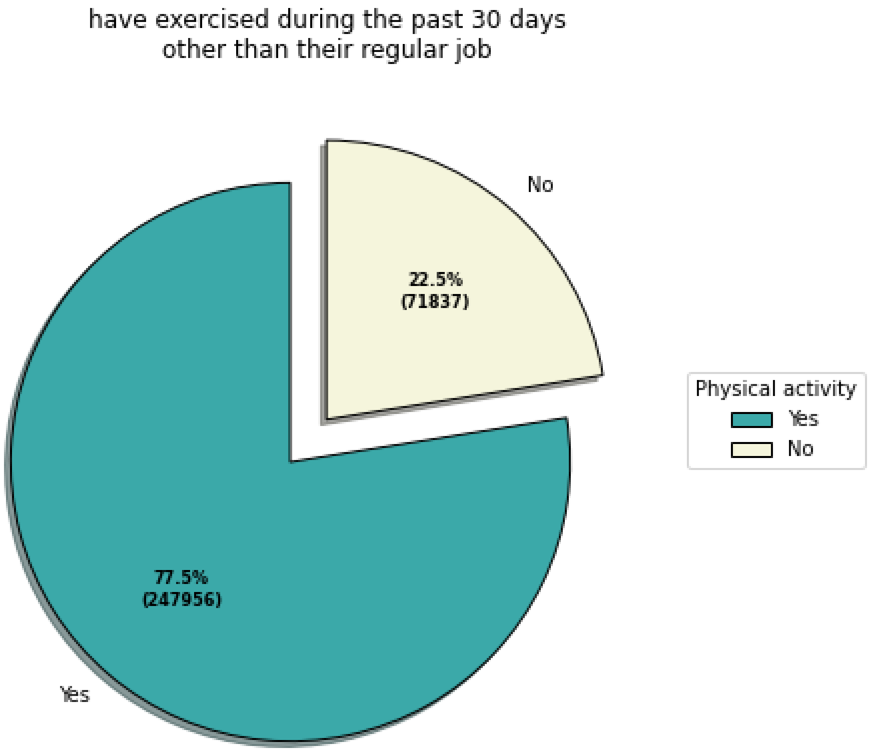


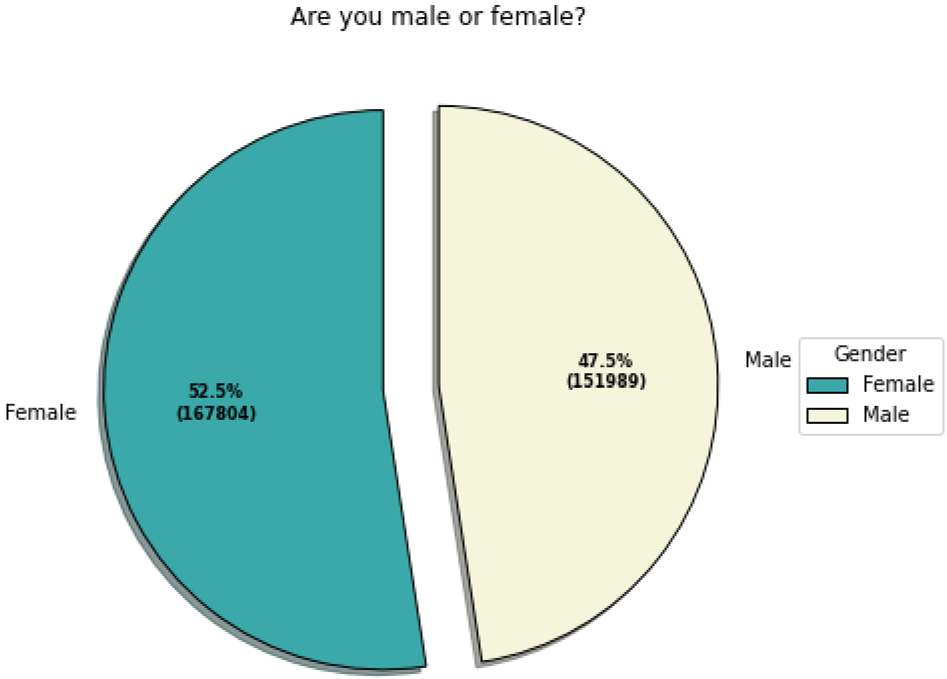
Most people have a good body without any injury



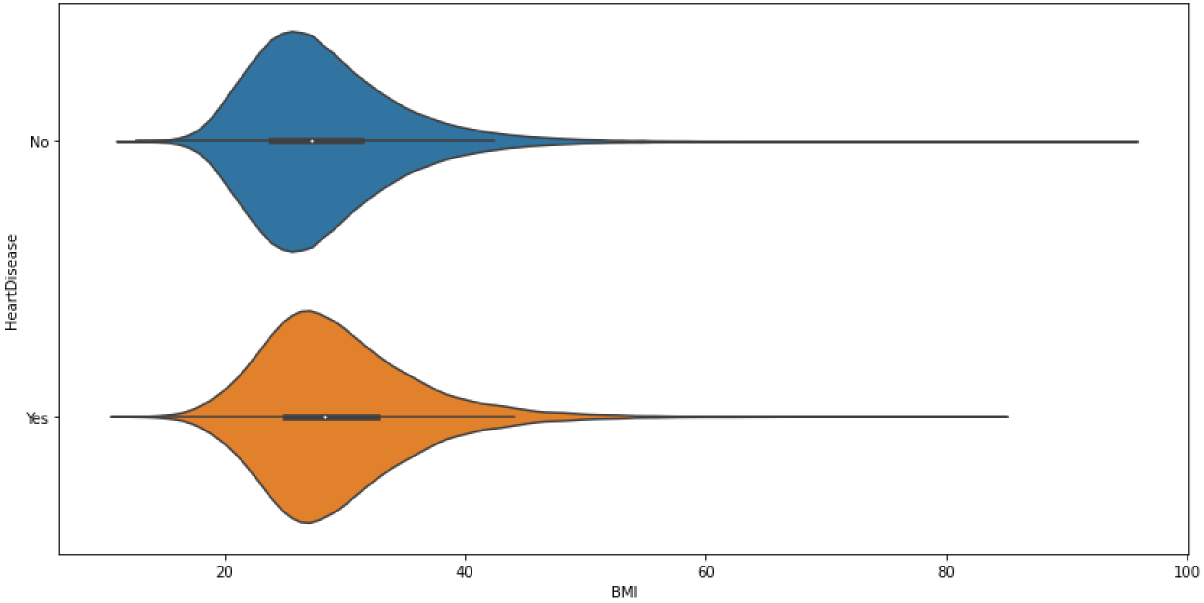
Most people don’t have problem in their mentality

Explore Data


have exercised during the past 30 days other than their regular job


Are you male or female?

Feature Vs Goal

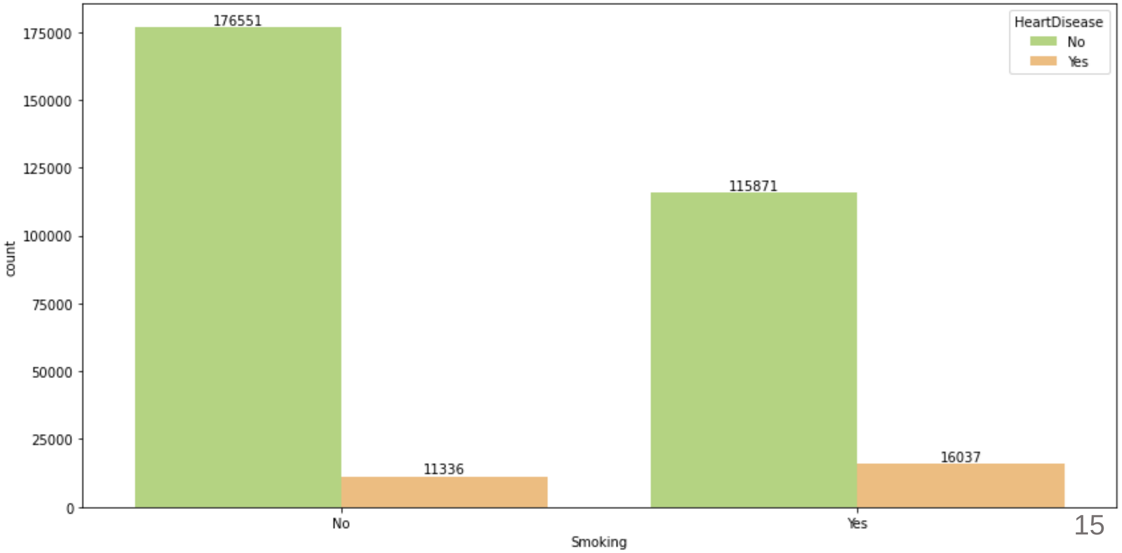


Heart Disease Vs BMI

BMI is not main factor in heart disease

Heart Disease Vs Smoking

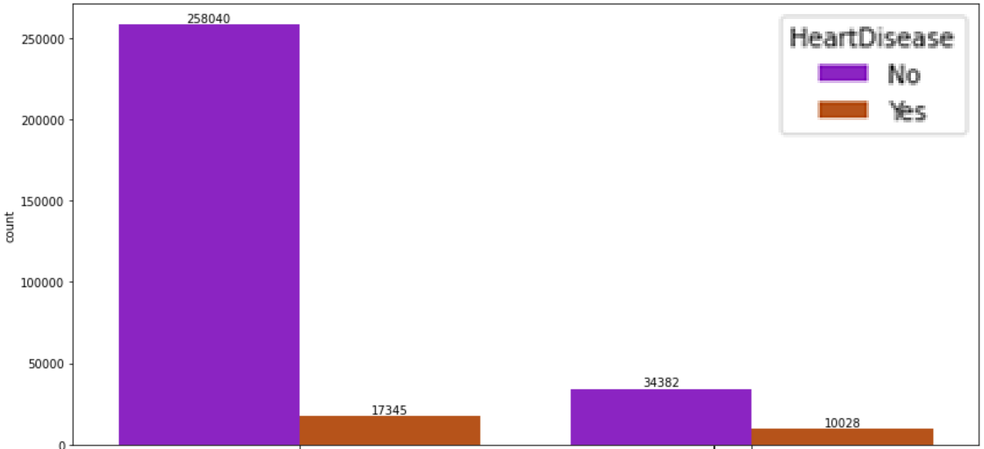
Smoking is one of factors in heart disease



Feature Vs Goal

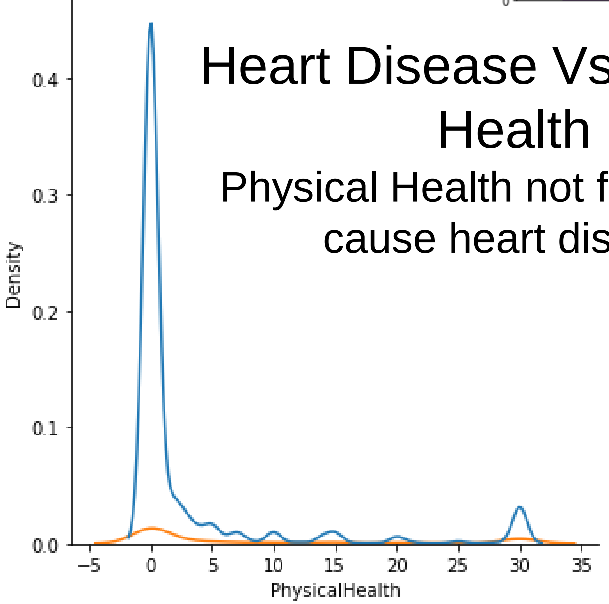
Heart Disease Vs difficult walking

difficult walking is one of factors in heart disease



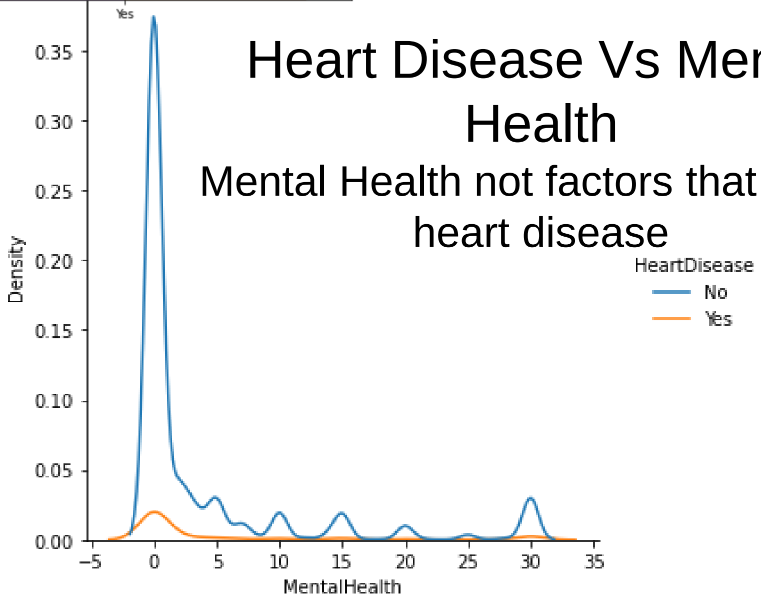
Heart Disease Vs Physical Health

Physical Health not factors that cause heart disease

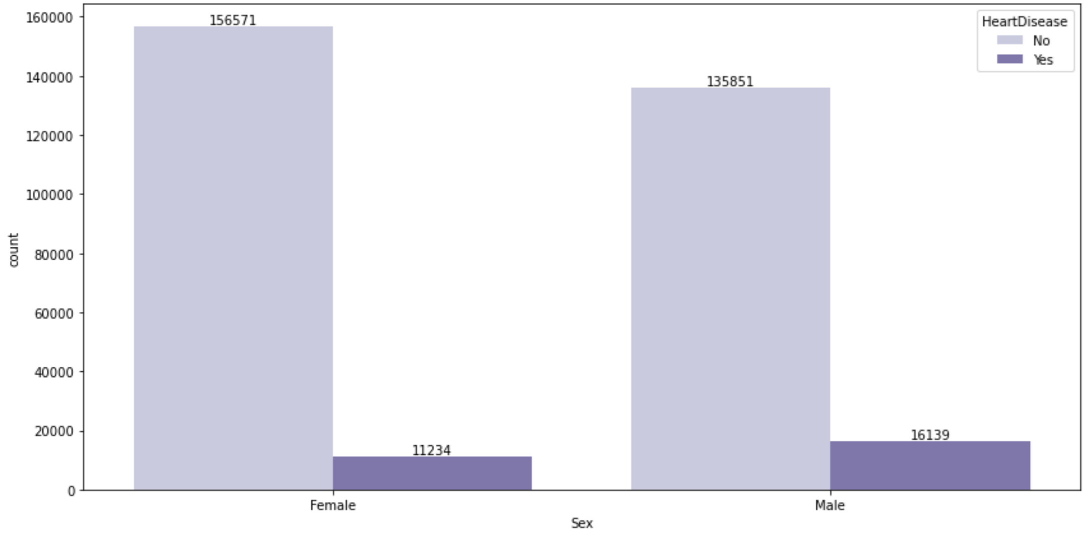


Heart Disease Vs Mental Health

Mental Health not factors that cause heart disease



Feature Vs Goal

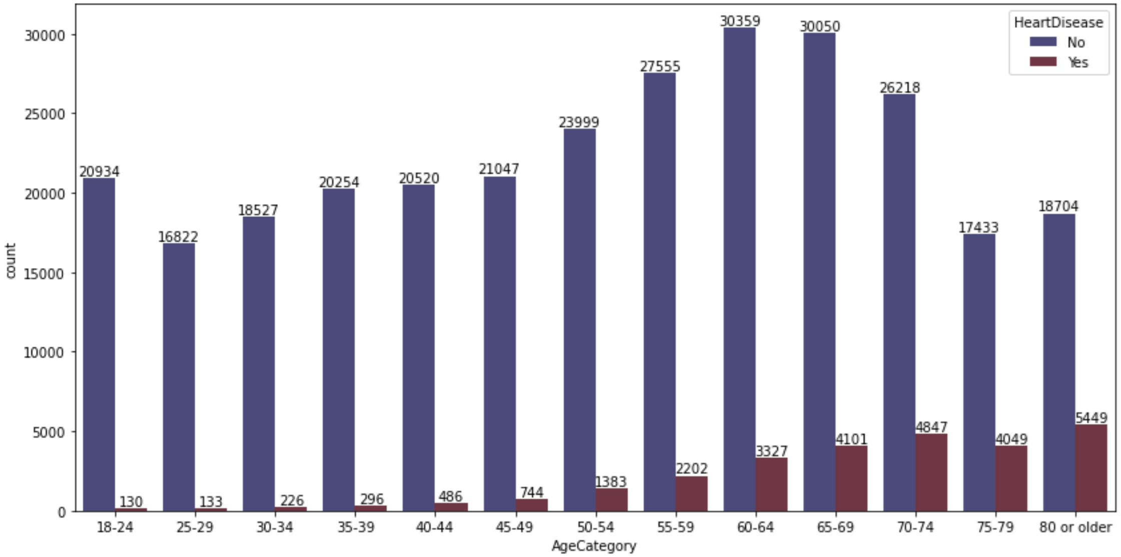


Heart Disease Vs Gender

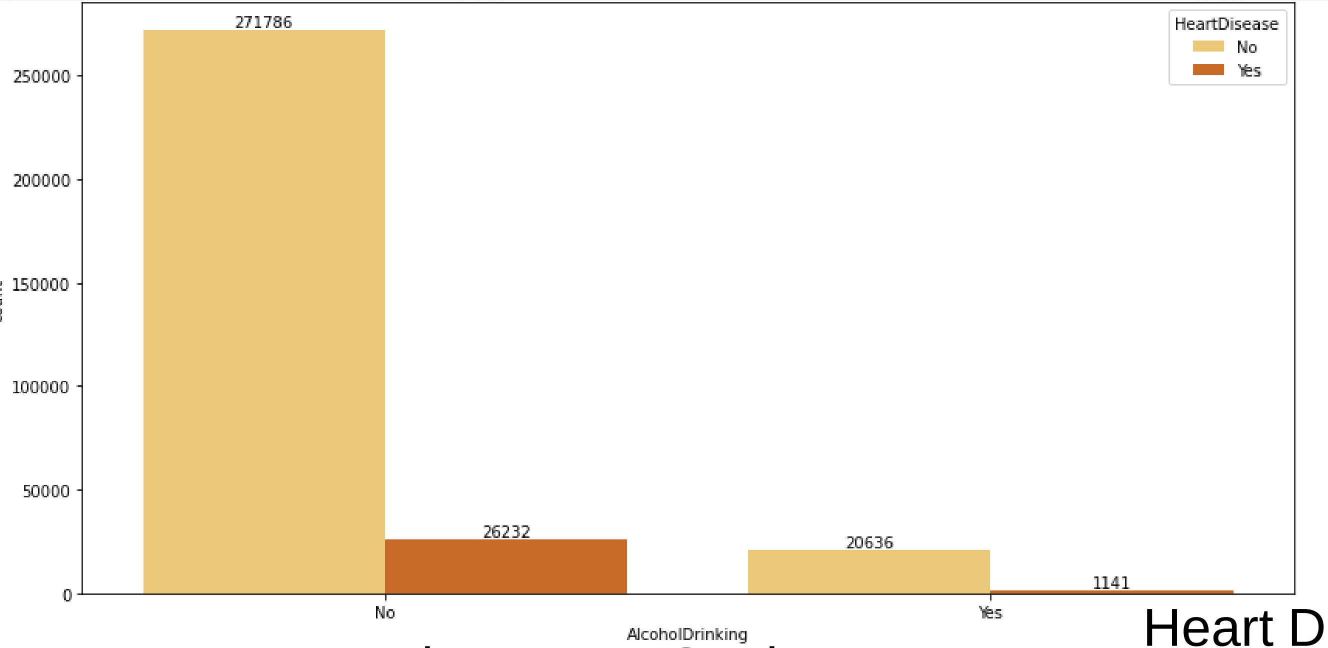
Males are the most people have heart disease

Heart Disease Vs age

People after 60 might have heart disease



Feature Vs Goal

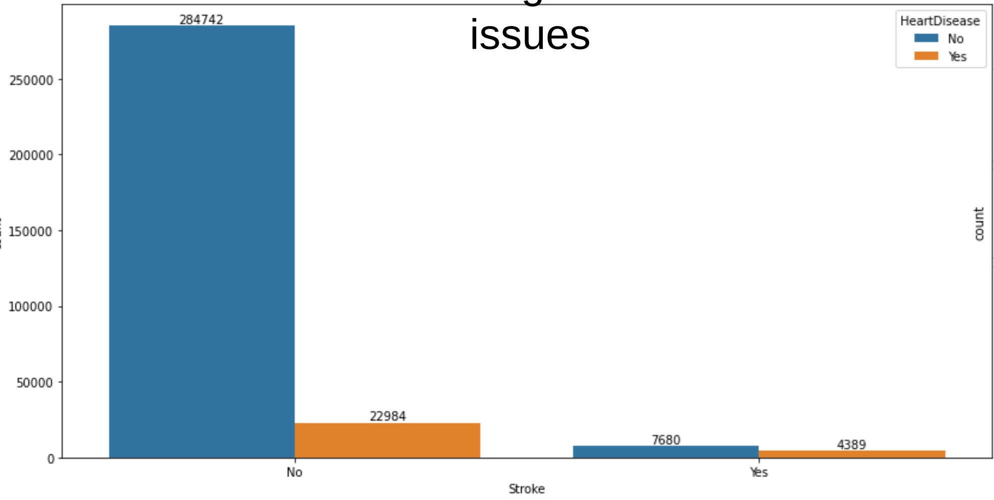


Heart Disease Vs alcohol drink

Drinking alcohol is not big reason of heart disease

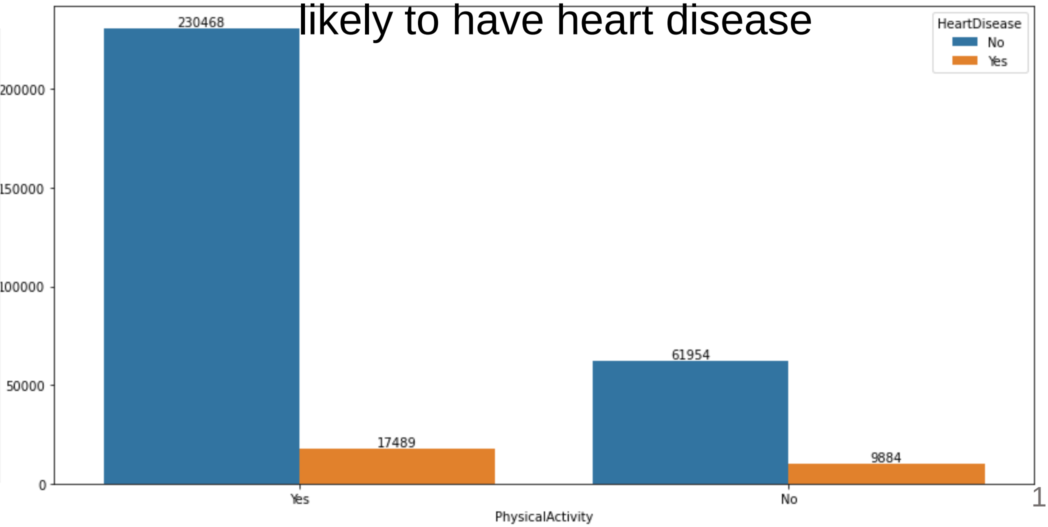
Heart Disease Vs Stroke

Almost 50% of having Stroke means heart issues

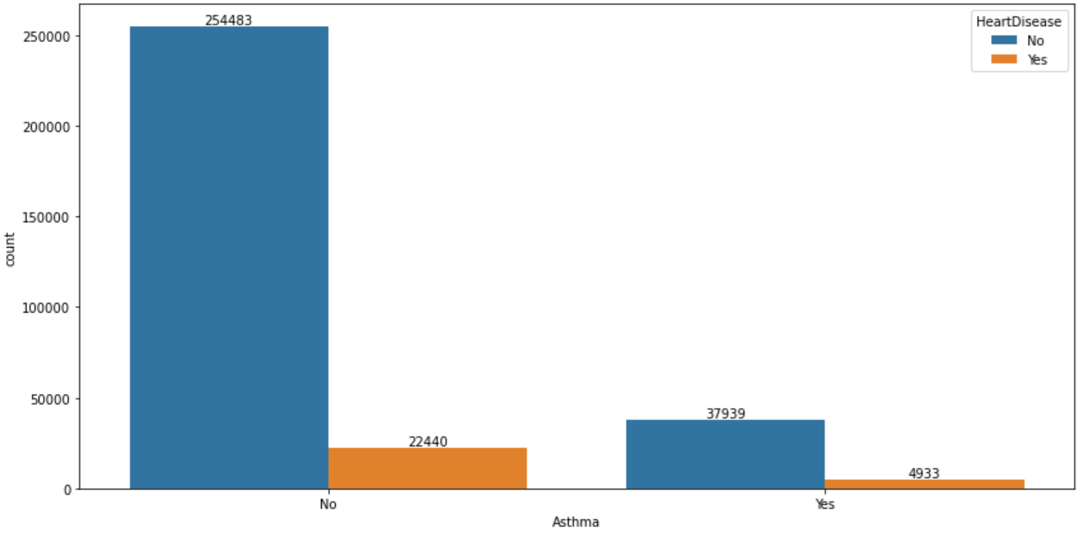


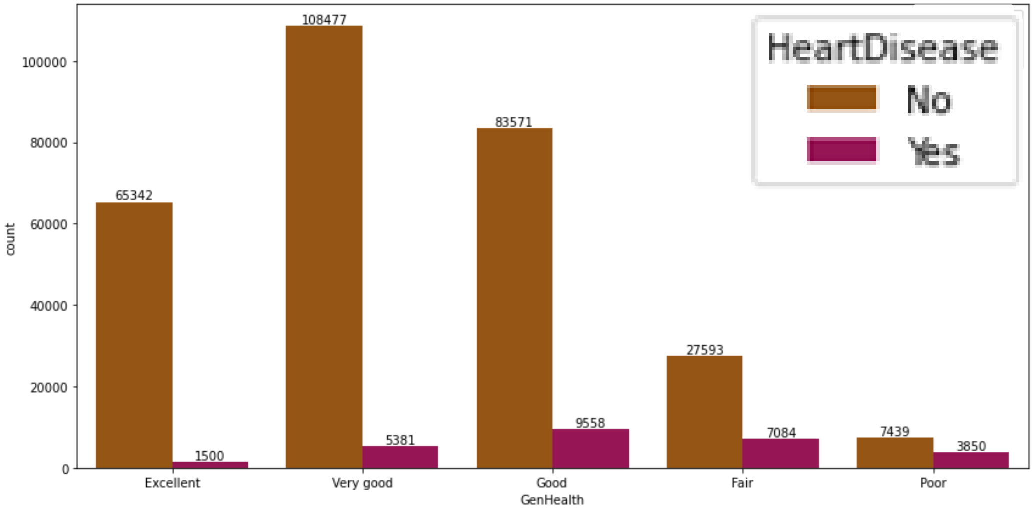
Heart Disease Vs Physical Activity

People who don’t do activities. They are more likely to have heart disease

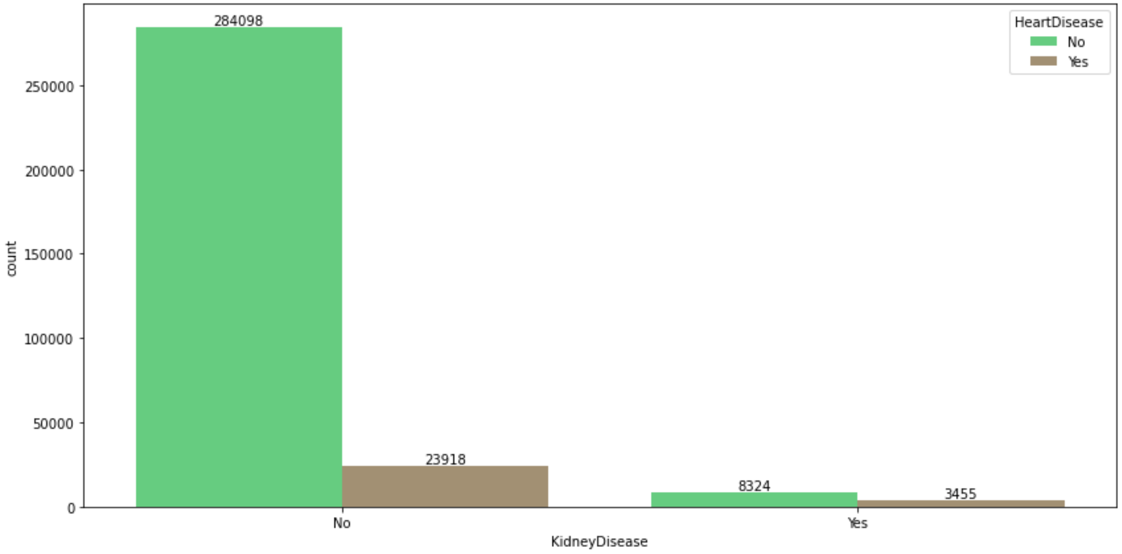


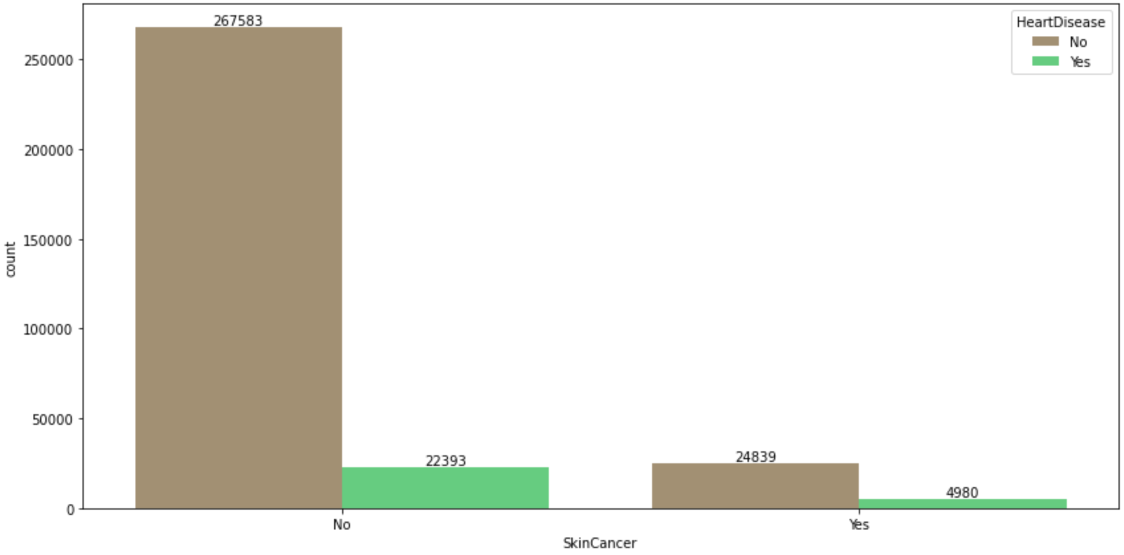
Feature Vs Goal

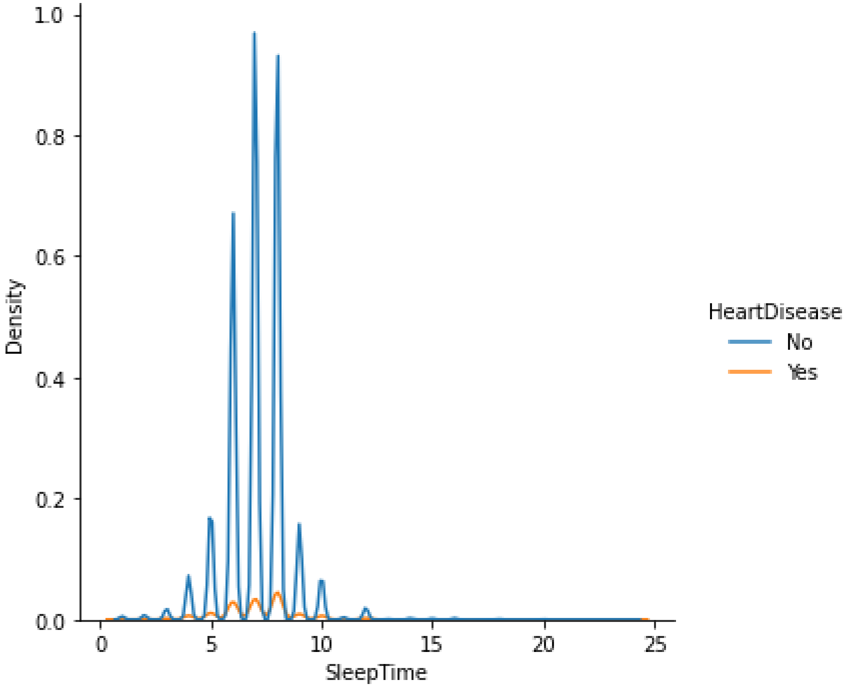




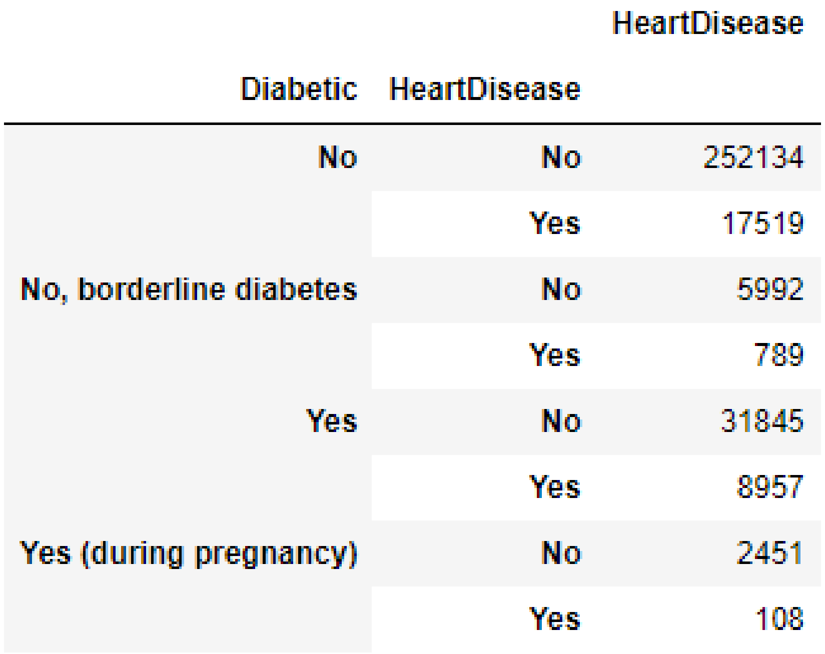
Feature Vs Goal



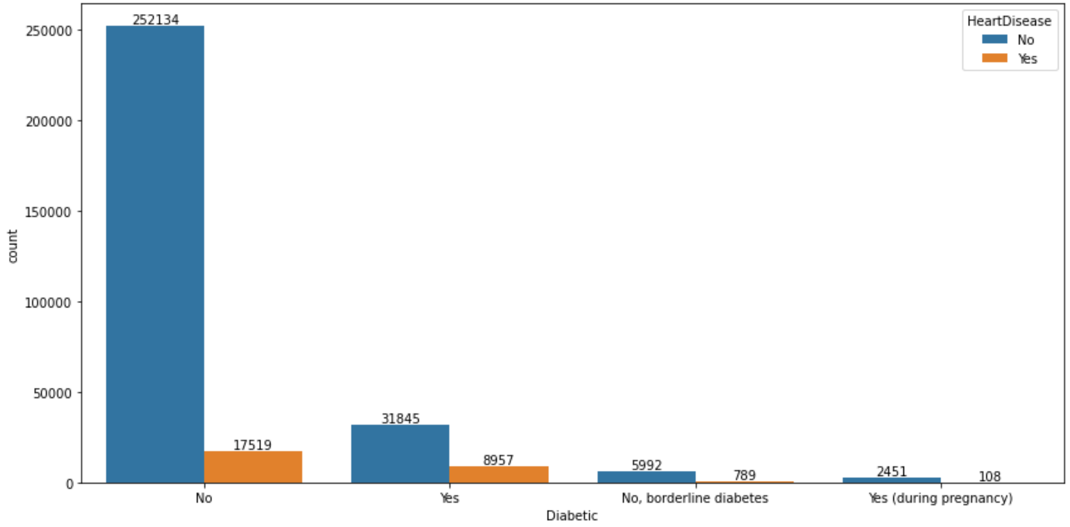




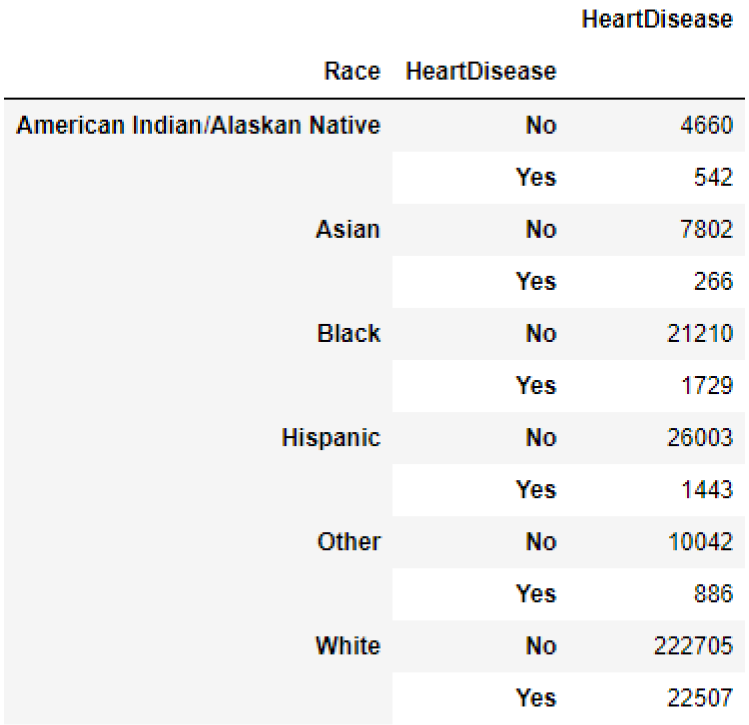
Feature Vs Goal



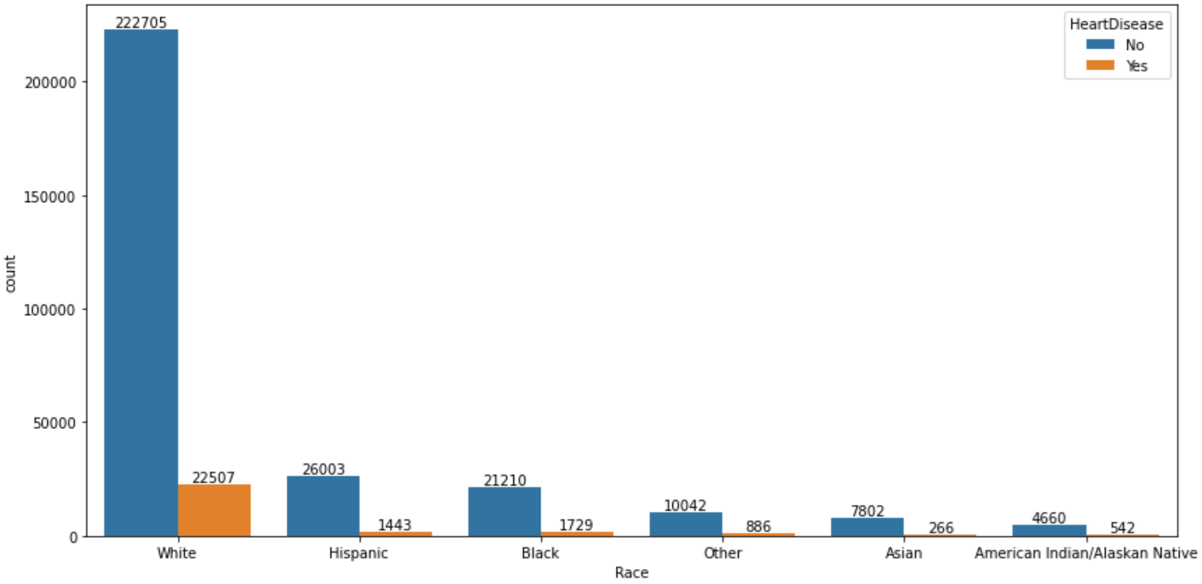
		HeartDisease
Diabetic	HeartDisease	
No	No	252134
	Yes	17519
No, borderline diabetes	No	5992
	Yes	789
Yes	No	31845
	Yes	8957
Yes (during pregnancy)	No	2451
	Yes	108



Feature Vs Goal

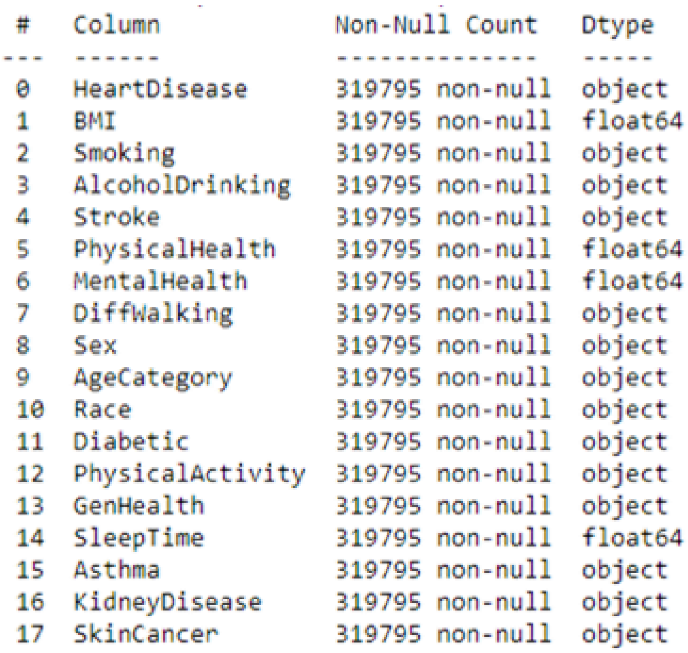


		HeartDisease
Race	HeartDisease	
American Indian/Alaskan Native	No	4660
	Yes	542
Asian	No	7802
	Yes	266
Black	No	21210
	Yes	1729
Hispanic	No	26003
	Yes	1443
Other	No	10042
	Yes	886
White	No	222705
	Yes	22507



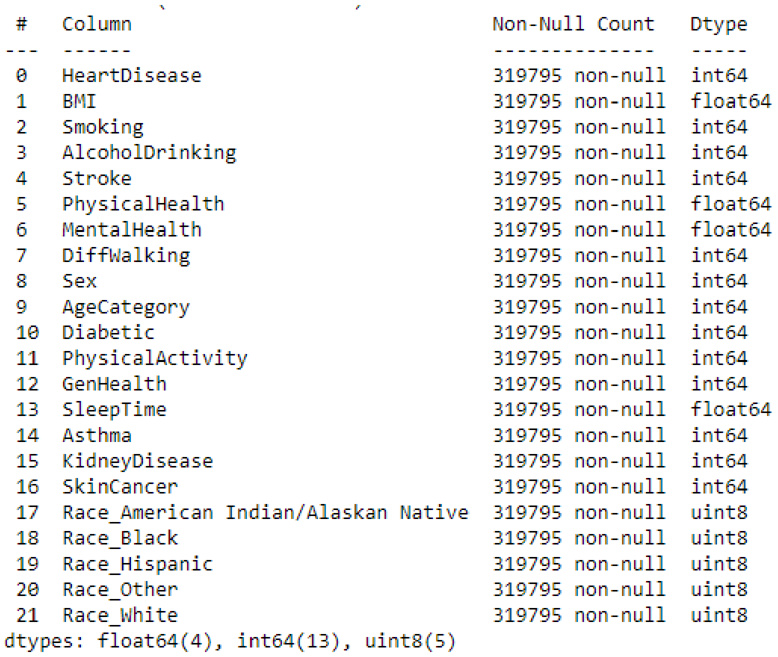
Data preprocessing

Data encoding



```
#   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
```

No missing Values
in the data set



```
 #   Column                             Non-Null Count   Dtype
---  ------                             --------------   -----
 0   HeartDisease                       319795 non-null  int64
 1   BMI                                319795 non-null  float64
 2   Smoking                            319795 non-null  int64
 3   AlcoholDrinking                    319795 non-null  int64
 4   Stroke                             319795 non-null  int64
 5   PhysicalHealth                     319795 non-null  float64
 6   MentalHealth                       319795 non-null  float64
 7   DiffWalking                        319795 non-null  int64
 8   Sex                                319795 non-null  int64
 9   AgeCategory                        319795 non-null  int64
 10  Diabetic                           319795 non-null  int64
 11  PhysicalActivity                   319795 non-null  int64
 12  GenHealth                          319795 non-null  int64
 13  SleepTime                          319795 non-null  float64
 14  Asthma                             319795 non-null  int64
 15  KidneyDisease                      319795 non-null  int64
 16  SkinCancer                         319795 non-null  int64
 17  Race_American Indian/Alaskan Native 319795 non-null  uint8
 18  Race_Black                         319795 non-null  uint8
 19  Race_Hispanic                      319795 non-null  uint8
 20  Race_Other                         319795 non-null  uint8
 21  Race_White                         319795 non-null  uint8
dtypes: float64(4), int64(13), uint8(5)
```

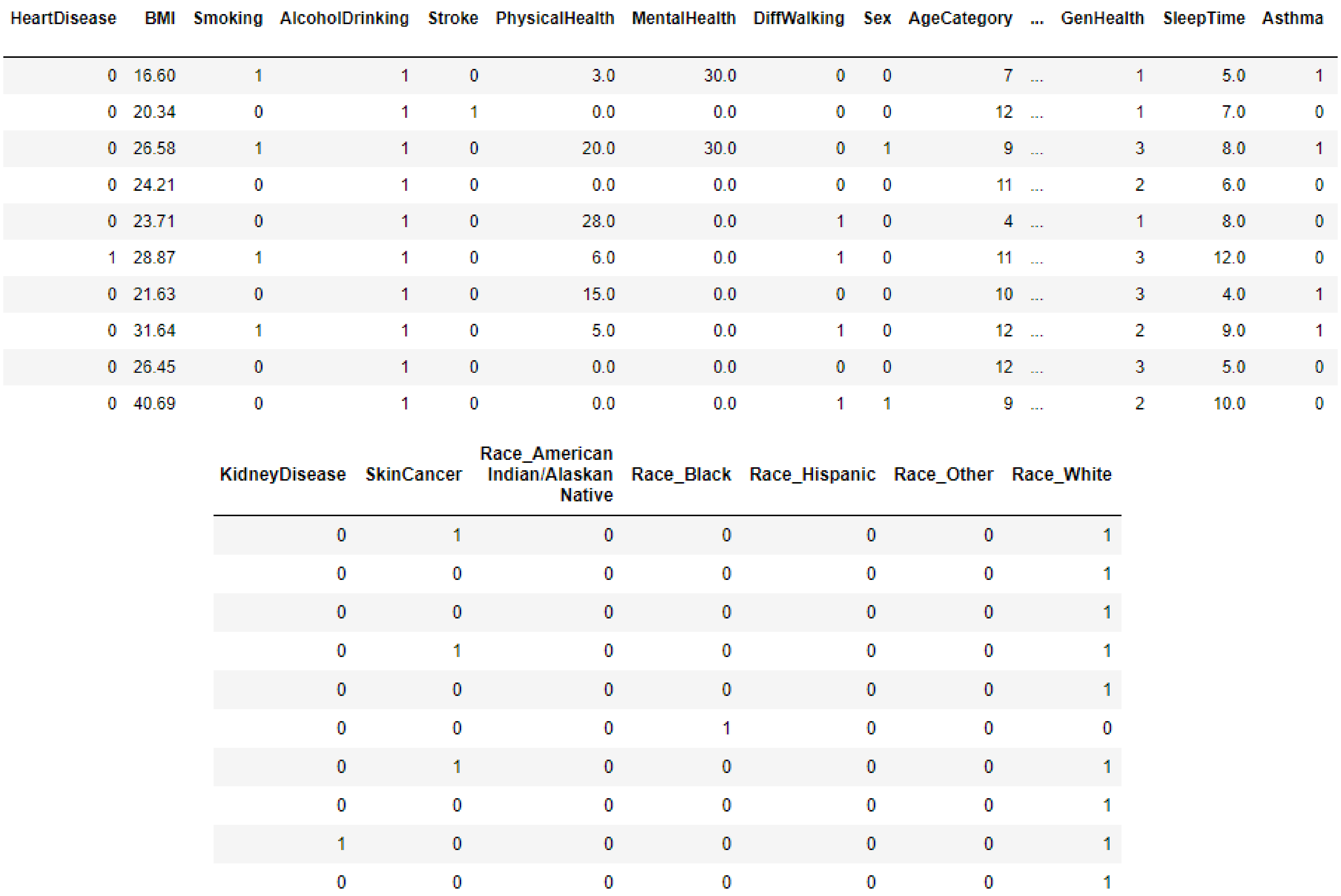
Data balancing

RandomOverSampler <> RandomUnderSampler

Data Outliers

There is no outliers in the data set

Data preprocessing



HeartDisease	BMI	Smoking	AlcoholDrinking	Stroke	PhysicalHealth	MentalHealth	DiffWalking	Sex	AgeCategory	...	GenHealth	SleepTime	Asthma
0	16.60	1	1	0	3.0	30.0	0	0	7	...	1	5.0	1
0	20.34	0	1	1	0.0	0.0	0	0	12	...	1	7.0	0
0	26.58	1	1	0	20.0	30.0	0	1	9	...	3	8.0	1
0	24.21	0	1	0	0.0	0.0	0	0	11	...	2	6.0	0
0	23.71	0	1	0	28.0	0.0	1	0	4	...	1	8.0	0
1	28.87	1	1	0	6.0	0.0	1	0	11	...	3	12.0	0
0	21.63	0	1	0	15.0	0.0	0	0	10	...	3	4.0	1
0	31.64	1	1	0	5.0	0.0	1	0	12	...	2	9.0	1
0	26.45	0	1	0	0.0	0.0	0	0	12	...	3	5.0	0
0	40.69	0	1	0	0.0	0.0	1	1	9	...	2	10.0	0

KidneyDisease	SkinCancer	Race_American Indian/Alaskan Native	Race_Black	Race_Hispanic	Race_Other	Race_White
0	1	0	0	0	0	1
0	0	0	0	0	0	1
0	0	0	0	0	0	1
0	1	0	0	0	0	1
0	0	0	0	0	0	1
0	0	0	1	0	0	0
0	1	0	0	0	0	1
0	0	0	0	0	0	1
1	0	0	0	0	0	1
0	0	0	0	0	0	1

Conclusion

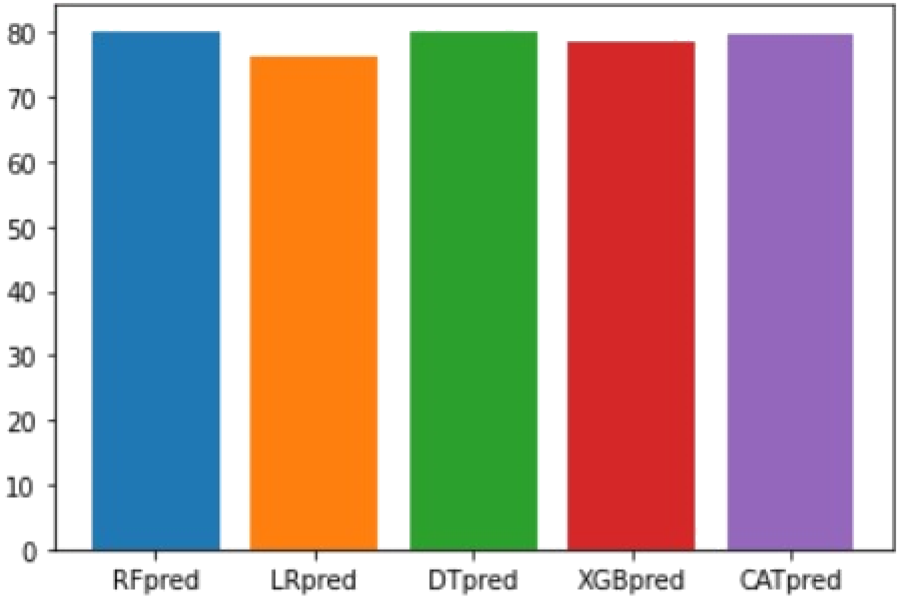
After Exploring the relation between each feature with the heart disease we found that these are a lot for factors that most probably makes heart disease happened.
1. Smoking
2. Difficulty of Walking
3. History of health (fair and poor)
4. Diabetic
5. over 65 years old

Recommendation (Business solution)

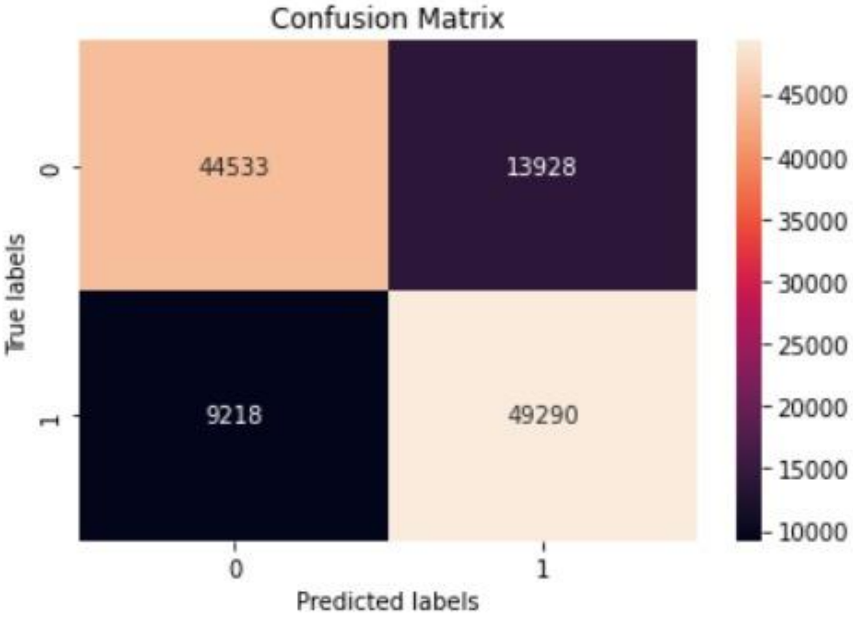
Game board that reduce bad habits like smoking and Alcohol and increasing doing sort of exercises

Modeling & Evaluation

```
RFpred:80.21185100325727
LRpred:76.30996246868828
DTpred:80.18620318203969
XGBpred:78.641349417367
CATpred:79.5364583778608
```



Random Forest (Best Score)



```
              precision    recall  f1-score   support

           0       0.83      0.76      0.79     58461
           1       0.78      0.84      0.81     58508

    accuracy                           0.80    116969
   macro avg       0.80      0.80      0.80    116969
weighted avg       0.80      0.80      0.80    116969
```

SAMSUNG

Together for Tomorrow!
Enabling People
Education for Future Generations