Personal Key Indicators of Heart Disease

MA-541-A Statistical Methods, Spring 2022

Samrudh Nagesh (samrudh10@gmail.com)

Abstract:

Each year, about 659,000 people die due to heart related diseases in the United States of America alone. That's 1 in every 4 deaths. A person having heart disease can be contributed to many factors. These factors can be the persons physical and mental health. This project aims to identify the different factors that can contribute to a person's heart health. This can range from their drinking habits to their sleep and mental health. We can use machine learning algorithms like logistic regression and linear regression to build a model to predict if a person can have a heart disease. By using this dataset, we can also perform other tests such as finding any significant difference of a person's Body Mass Index (BMI) due to their physical activity using ANOVA.

Index terms-Machine learning, linear regression, logistic regression, BMI, ANOVA

Introduction:

Cardiovascular diseases (CVDs) are the leading cause of death globally, taking an estimated 17.9 million lives each year. CVDs are a group of disorders of the heart and blood vessels and include coronary heart disease, cerebrovascular disease, rheumatic heart disease and other conditions. More than four out of five CVD deaths are due to heart attacks and strokes, and one third of these deaths occur prematurely in people under 70 years of age.

The most important behavioral risk factors of heart disease and stroke are unhealthy diet, physical inactivity, tobacco use and harmful use of alcohol. The effects of behavioral risk factors may show up in individuals as raised blood pressure, raised blood glucose, raised blood lipids, and overweight and obesity. These "intermediate risks factors" can be measured in primary care facilities and indicate an increased risk of heart attack, stroke, heart failure and other complications.

According to the research conducted by the Centre for Disease Control and Prevention (CDC), heart disease is the leading cause of death for men, women, and people of most racial and ethnic groups in the United States. One person dies every 36 seconds in the United States from cardiovascular disease. About 659,000 people in the United States die from heart disease each year—that's 1 in every 4 deaths. Heart disease costs the United States about $363 billion each year from 2016 to 2017. This includes the cost of health care services, medicines, and lost productivity due to death.

There are two most common types of heart disease, these are:

Coronary Artery Disease

- Coronary heart disease is the most common type of heart disease, killing 360,900 people in 2019.

- About 18.2 million adults aged 20 and older have CAD (about 6.7%).
- About 2 in 10 deaths from CAD happen in adults less than 65 years old.

Heart Attack

- In the United States, someone has a heart attack every 40 seconds.
- Every year, about 805,000 people in the United States have a heart attack. Of these,
    - 605,000 are a first heart attack
    - 200,000 happen to people who have already had a heart attack
    - About 1 in 5 heart attacks is silent—the damage is done, but the person is not aware of it.

Heart Disease Deaths Vary by Sex, Race, and Ethnicity

Heart disease is the leading cause of death for people of most racial and ethnic groups in the United States, including African American, American Indian, Alaska Native, Hispanic, and white men. For women from the Pacific Islands and Asian American, American Indian, Alaska Native, and Hispanic women, heart disease is second only to cancer.[5]

Below are the percentages of all deaths caused by heart disease in 2015, listed by ethnicity, race, and sex.

Race of Ethnic Group	% of Deaths	Men, %	Women, %
American Indian or Alaska Native	18.3	19.4	17.0
Asian American or Pacific Islander	21.4	22.9	19.9
Black (Non-Hispanic)	23.5	23.9	23.1
White (Non-Hispanic)	23.7	24.9	22.5
Hispanic	20.3	20.6	19.9
All	23.4	24.4	22.3

Factors that can cause risk of heart disease

High blood pressure, high blood cholesterol, and smoking are key risk factors for heart disease. We see these factors are also observed in our dataset which will be analyzed in the next section.

Several other medical conditions and lifestyle choices can also put people at a higher risk for heart disease, including:

- Diabetes
- Overweight and obesity

- Unhealthy diet
- Physical inactivity
- Excessive alcohol use

In this project, we try to analyze the dataset to identify the key indicators of heart disease in men and women, perform regression on the data and perform statistical analysis on the various factors studied in the dataset using Python and R.

The Dataset

The dataset titled 'Personal Key Indicators of Heart Disease' is sourced from the Kaggle website. The dataset contains 18 columns of both numerical and categorical data and there are 319,796 entries which makes it a very large dataset. Most of the columns are categorical data, which is binary, which is the data entries are either 'Yes' or 'No'. The dataset is shown in Fig.2.

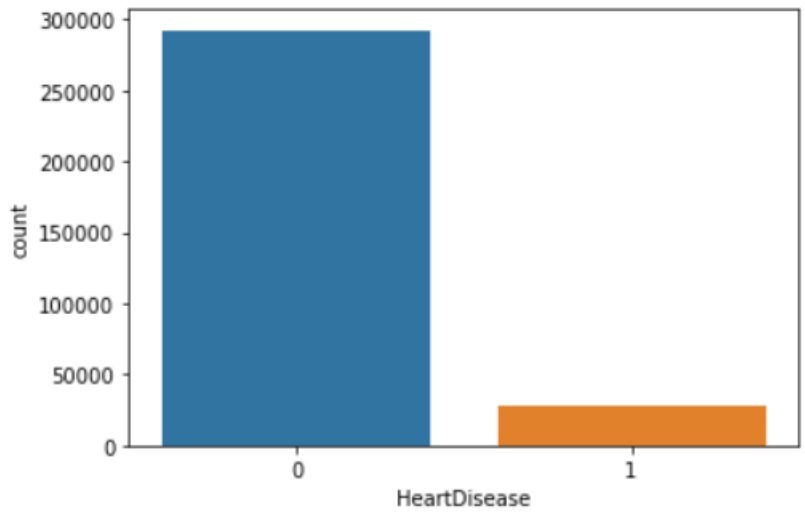
	HeartDisease	BMI	Smoking	AlcoholDrinking	Stroke	PhysicalHealth	MentalHealth	DiffWalking	Sex
0	No	16.60	Yes	No	No	3.0	30.0	No	Female
1	No	20.34	No	No	Yes	0.0	0.0	No	Female
2	No	26.58	Yes	No	No	20.0	30.0	No	Male
3	No	24.21	No	No	No	0.0	0.0	No	Female
4	No	23.71	No	No	No	28.0	0.0	Yes	Female

AgeCategory	Race	Diabetic	PhysicalActivity	GenHealth	SleepTime	Asthma	KidneyDisease	SkinCancer
57.0	White	Yes	Yes	Very good	5.0	Yes	No	Yes
80.0	White	No	Yes	Very good	7.0	No	No	No
67.0	White	Yes	Yes	Fair	8.0	Yes	No	No
77.0	White	No	No	Good	6.0	No	No	Yes
42.0	White	No	Yes	Very good	8.0	No	No	No

Parameters:

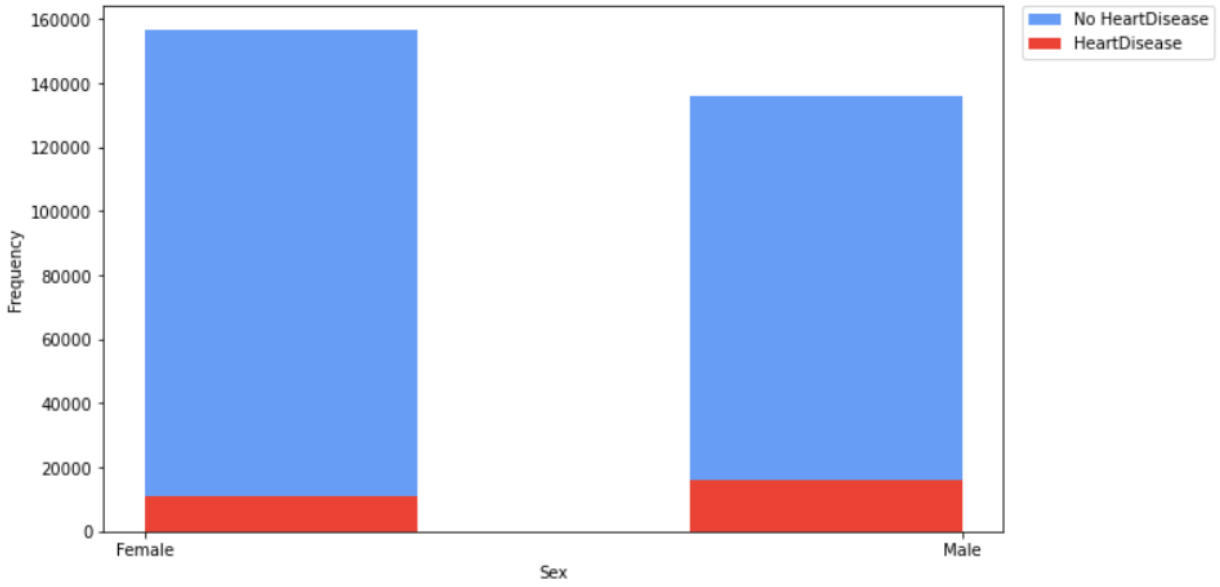
Here are some of the most important features of the dataset:

- **Heart Disease**: This is the parameter that indicates if the person who is suffering from any heart conditions or not. It is a binary categorical data and is the data that we use as the target variable for any regression. Out of the 300,000+ data entries, we observe that about 25,000 individuals are suffering from any type of heart disease. This is reflected in the plot displayed below.



In the Fig, 0 indicates no heart disease while 1 indicates the person has heart disease.

The dataset also features the sex of the person. We observe that the count of females in the dataset is higher than the count of males. The probability of a male with heart disease, is however, higher than the probability of a female with heart disease.

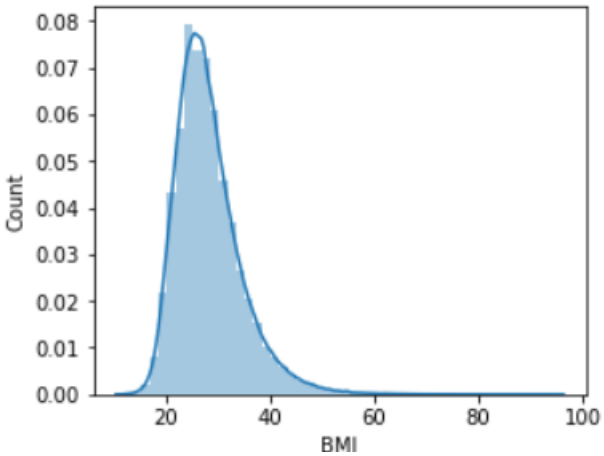


We observe that the probability of Male to have heart disease is 10.61%, while the probability of a Female with heart disease is 6.69%.

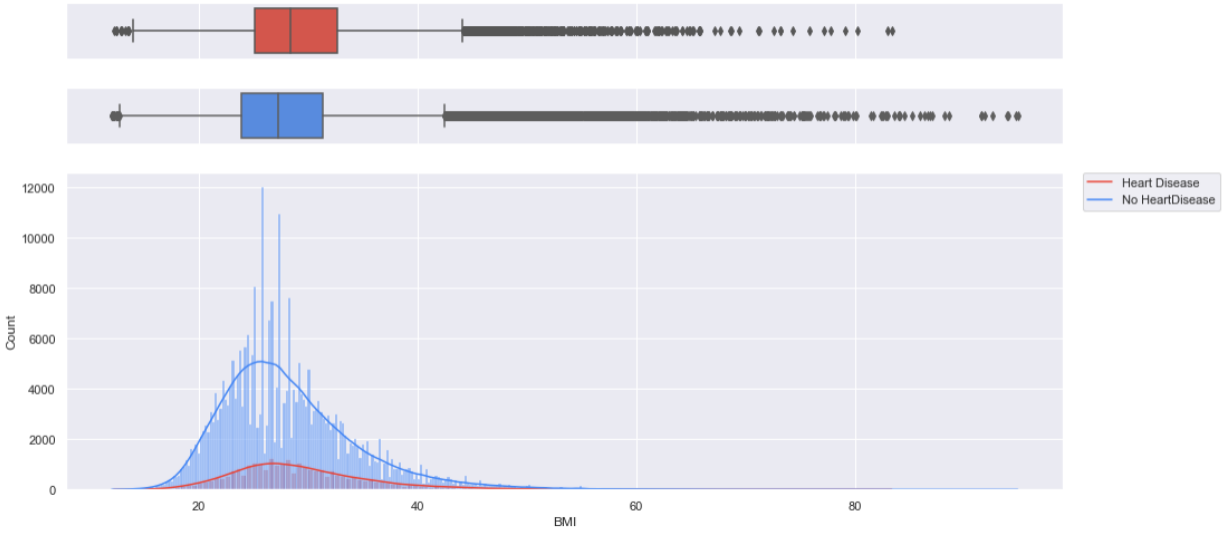
- **Body Mass Index**: Body Mass Index (BMI) is a person's weight in kilograms (or pounds) divided by the square of height in meters (or feet). A high BMI can indicate high body fatness. BMI screens for weight categories that may lead to health problems, but it does not diagnose the body fatness or health of an individual.

	mean	std	min	25%	50%	75%	max
BMI	28.325399	6.356100	12.020000	24.030000	27.340000	31.420000	94.850000

This table shows the mean, min and max of the numerical data BMI in the dataset.

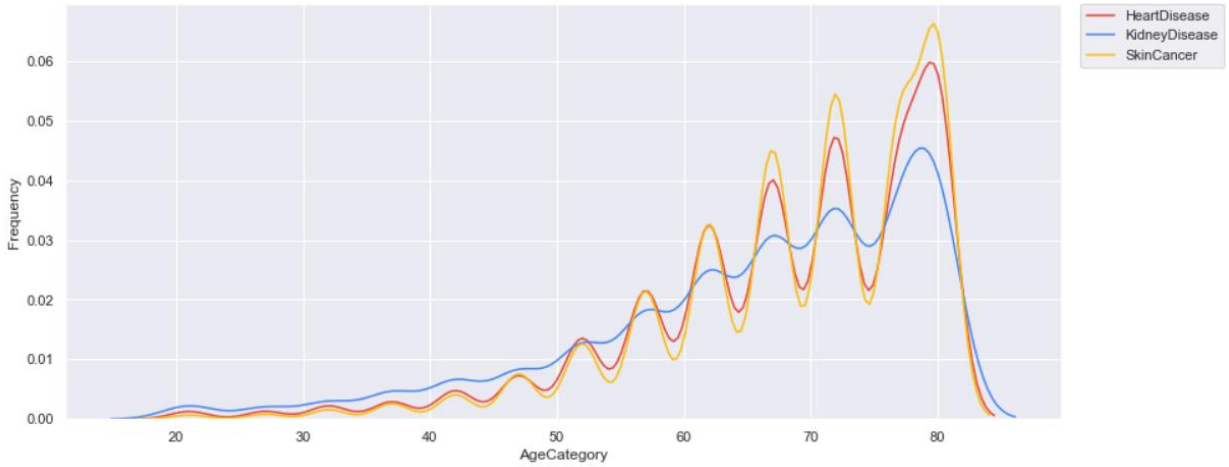


As we can see from the plot in fig that most of the people surveyed have BMI between 20 and 40.



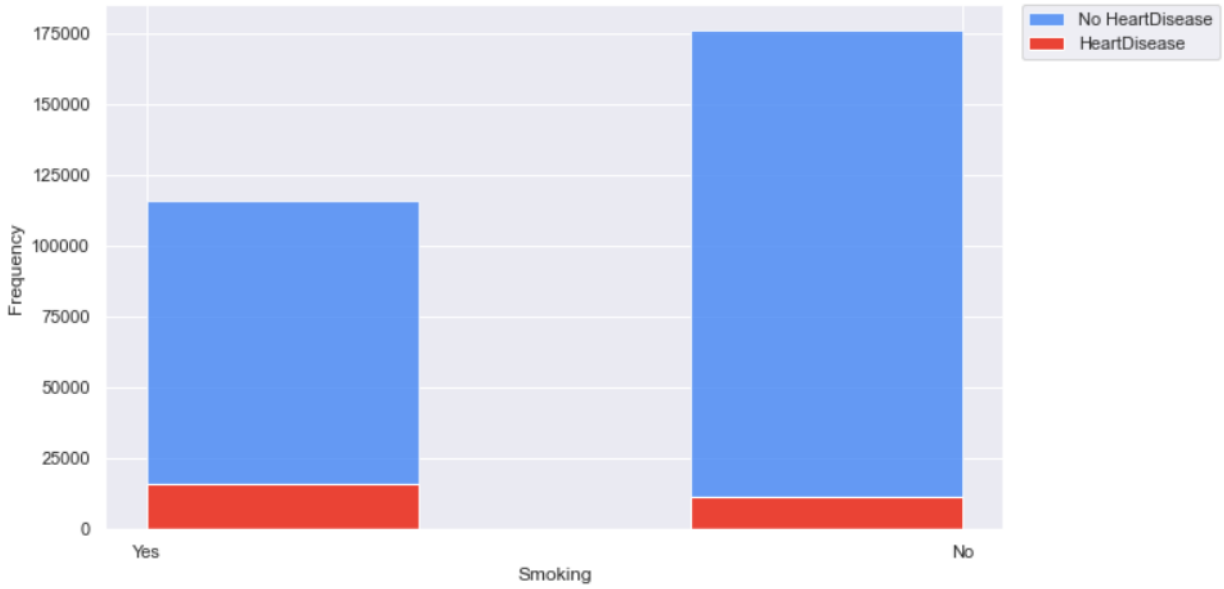
While comparing the BMI of people with heart disease and without, we see that people with heart disease have slightly higher BMI than people without heart disease.

- **Age Category**: In the dataset, the age category varies from 18 years to 80 and older. Heart disease can occur in people of all ages. We analyse how age affects the likelihood of not only heart disease, but also kidney disease and skin cancer.



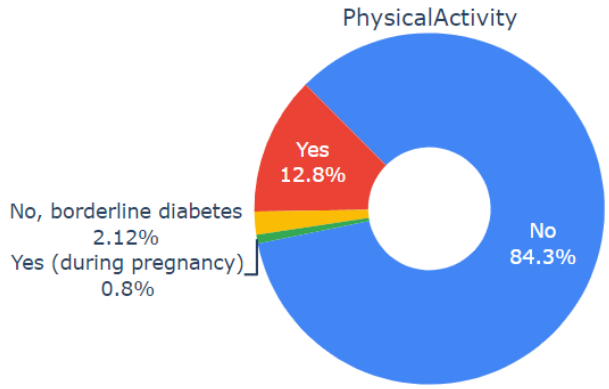
From the plot we observe that people aged around 80 years have a high probability of having heart disease, kidney disease and skin cancer.

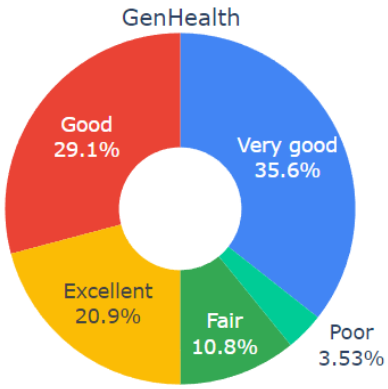
- **Smoking**: Smoking increases the formation of plaque in blood vessels. Coronary Heart Disease occurs when arteries that carry blood to the heart muscle are narrowed by plaque or blocked by clots. Chemicals in cigarette smoke cause the blood to thicken and form clots inside veins and arteries.



From the above plot, we observe that the probability of heart disease if you smoke is 12.15% and the probability of heart disease if you don't smoke is 6.03%. So, you have twice the chance of having heart disease if you are a smoker than a non-smoker.

- **General Health and Physical Activity**: A person's general health and physical activity are very important statistic when it comes to heart disease.

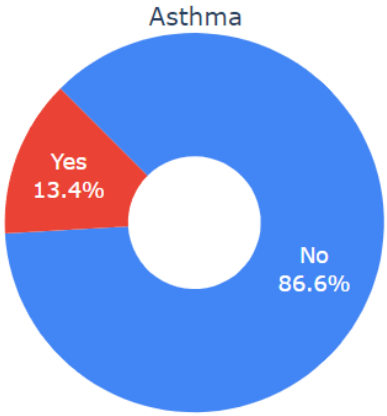


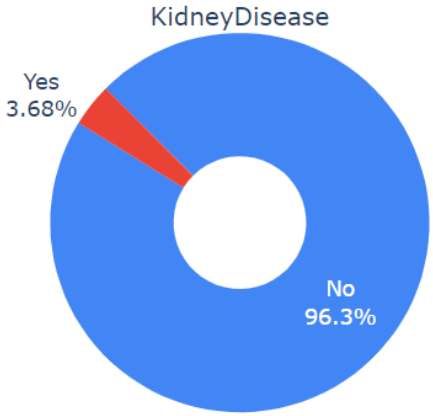


We observe that the people surveyed have good general health.

Other features in the dataset include diseases such as

- Asthma
- Stroke
- Diabetes
- Kidney Disease
- Skin Cancer





Analysis on numeric data:

The numerical data include BMI, Physical Health, Mental Health, Age Category and Sleep Time. We find the min, max, and mean of these parameters.

	min	mean	max
BMI	12.020000	28.325399	94.850000
PhysicalHealth	0.000000	3.371710	30.000000
MentalHealth	0.000000	3.898366	30.000000
AgeCategory	21.000000	54.355759	80.000000
SleepTime	1.000000	7.097075	24.000000

By observing the max of these numerical data, we can clearly observe some false data in the dataset such as sleep time of 24 hours and BMI of 94.85.

Statistical Analysis

In this section we discuss the results of different types of statistical analysis on the dataset such as one-way ANOVA and two-way ANOVA using Python.

Data Preprocessing:

We first perform data preprocessing and data manipulation on the dataset's numerical and categorical variables.

First, we analyze the dataset to find any null values present in the dataset. If we find any null values, for numerical data we replace them with the mean, for categorical data we delete the data entry entirely. But fortunately, this dataset does not have any null values.

```
In [9]:  1 df.isnull().sum()
```

```
Out[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
```

We then perform preprocessing on categorical data such as Age Category and Race as these features along with BMI are used for our statistical analysis.

```
In [10]: 1 df.AgeCategory.unique()

Out[10]: array(['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'],
               dtype=object)
```

```
In [11]: 1 df.groupby('AgeCategory').BMI.count()

Out[11]: AgeCategory
         18-24          21064
         25-29          16955
         30-34          18753
         35-39          20550
         40-44          21006
         45-49          21791
         50-54          25382
         55-59          29757
         60-64          33686
         65-69          34151
         70-74          31065
         75-79          21482
         80 or older    24153
         Name: BMI, dtype: int64
```

We find the different categories in our Age Category data. We then reduce these categories to 4 categories being Young (age between 18 and 29), Adult (age between 30 and 49), Old (age between 50 and 69) and Very Old (age 70 and older). We also use dummy variables for the Race dataset, giving them numeric values 0-5.

The dataset after data preprocessing is as follows.

	index	HeartDisease	BMI	Smoking	AlcoholDrinking	Stroke	PhysicalHealth	MentalHealth	DiffWalking
0	44797	No	25.06	No	No	No	0.0	2.0	No
1	194160	No	32.28	No	No	No	0.0	0.0	No
2	173073	No	30.81	No	No	No	1.0	3.0	No
3	171625	No	32.28	No	No	No	0.0	0.0	No
4	28145	No	25.79	No	No	No	0.0	7.0	No

Sex	AgeCategory	Race	Diabetic	PhysicalActivity	GenHealth	SleepTime	Asthma	KidneyDisease	SkinCancer
Male	Young	0	No	Yes	Excellent	8.0	No	No	No
Male	Young	5	No	Yes	Excellent	6.0	No	No	No
Male	Young	0	No	Yes	Good	8.0	No	No	No
Male	Young	0	Yes	Yes	Very good	8.0	No	No	No
Female	Young	0	No	Yes	Excellent	7.0	No	No	No

1) The Shapiro-Wilk test:

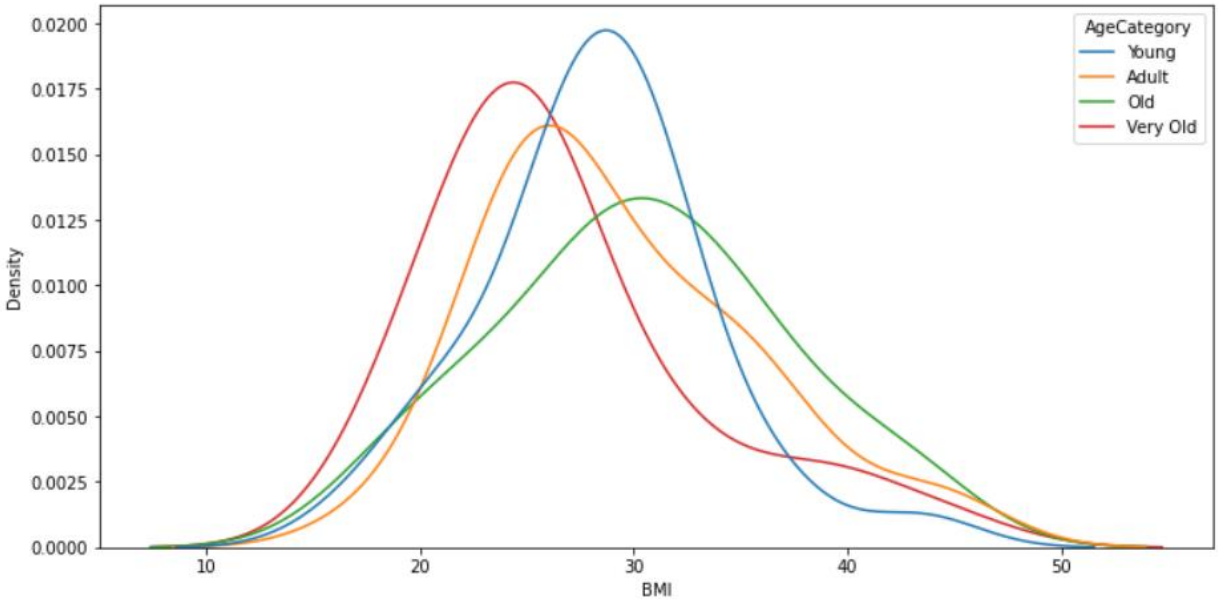
First, we test for normality. The Shapiro–Wilk test is a test of normality in frequentist statistics. The Shapiro–Wilk test tests the null hypothesis that a sample x_1, ..., x_n came from a normally distributed population. The test statistic is

$$W = \frac{\left(\sum_{i=1}^{n} a_i x_{(i)}\right)^2}{\sum_{i=1}^{n}(x_i - \overline{x})^2},$$

- x is the *i*th order statistic, i.e., the *i*th-smallest number in the sample

The null-hypothesis of this test is that the population is normally distributed. Thus, if the p value is less than the chosen alpha level, then the null hypothesis is rejected and there is evidence that the data tested are not normally distributed. On the other hand, if the p value is greater than the chosen alpha level, then the null hypothesis (that the data came from a normally distributed population) cannot be rejected (e.g., for an alpha level of .05, a data set with a p value of less than .05 rejects the null hypothesis that the data are from a normally distributed population).

The BMI vs Age Category plot after data preprocessing is shown in the plot shown below.



We group the means of BMI with their respective age categories and perform the Shapiros test.

```
In [17]: 1 df_sampled.groupby('AgeCategory').BMI.mean()

Out[17]: AgeCategory
         Adult       29.328333
         Old         30.377000
         Very Old    26.701000
         Young       28.274000
         Name: BMI, dtype: float64
```

Code:

st.shapiro(df_sample_young.BMI),st.shapiro(df_sample_adult.BMI),st.shapiro(df_sample_old.BMI),st.shapiro(df_sample_very_old.BMI)

Output:

```
(ShapiroResult(statistic=0.965241014957428, pvalue=0.41835784912109375),
 ShapiroResult(statistic=0.9382164478302002, pvalue=0.08144113421440125),
 ShapiroResult(statistic=0.9818514585494995, pvalue=0.8723570108413696),
 ShapiroResult(statistic=0.8910850882530212, pvalue=0.005118906497955322))
```

We observe that the p-value of our test is greater than our alpha value, therefore we can go ahead with our normality assumption.

2) Bartlett's test

In statistics, Bartlett's test, named after Maurice Stevenson Bartlett, is used to test homoscedasticity, that is, if multiple samples are from populations with equal variances. Some statistical tests, such as the analysis of variance, assume that variances are equal across groups or samples, which can be verified with Bartlett's test.

```
In [19]: 1 statistics,pvalue=st.bartlett(df_sample_young.BMI,df_sample_adult.BMI,
         2                              df_sample_old.BMI,df_sample_very_old.BMI)
         3 print(" statistics = {}\n pvalue ={}".format(statistics,pvalue))

         statistics = 1.7737612075649847
         pvalue =0.6206613681678832
```

We observe that the p-value is greater than the alpha value, therefore we can assess the homogeneity of variance in our sample of BMI and age.

3) One-Way ANOVA

In statistics, one-way analysis of variance (abbreviated one-way ANOVA) is a technique that can be used to compare whether two sample's means are significantly different or not (using the F

distribution). This technique can be used only for numerical response data, the "Y", usually one variable, and numerical or (usually) categorical input data, the "X", always one variable, hence "one-way".

In [20]:
```
import statsmodels.api as sm
from statsmodels.formula.api import ols

model = ols('BMI ~ C(AgeCategory)', data=df_sampled).fit()
sm.stats.anova_lm(model, typ=2)
```

Out[20]:

	sum_sq	df	F	PR(>F)
C(AgeCategory)	221.430863	3.0	1.875101	0.137645
Residual	4566.150037	116.0	NaN	NaN

We see that Out[20] displays the one way ANOVA table for BMI and Age Category. We also observe from the ANOVA table that our test statistic F is less than our alpha value (2%), that is we have at least one BMI mean that statistically differs from the other. Therefore, we run the post-hoc test to find the pair that differs. We use the Tukey method to adjust our p-value rather than t-test so that we can have a global type I error of alpha.

4) Tukey's test

We use the Tukey's test to find the pairs that statistically differ in our ANOVA table. We adjust our p-value so that we can have a global type I error of alpha.

In [21]:
```
from statsmodels.stats.multicomp import pairwise_tukeyhsd
tukey=pairwise_tukeyhsd(endog=df_sampled.BMI,groups=df_sampled.AgeCategory,
                        alpha=0.05)
print(tukey)
```

```
  Multiple Comparison of Means - Tukey HSD, FWER=0.05
====================================================
 group1   group2  meandiff p-adj   lower  upper  reject
----------------------------------------------------
   Adult      Old   1.0487 0.9163  -3.174 5.2713  False
   Adult Very Old  -2.6273 0.3704   -6.85 1.5953  False
   Adult    Young  -1.0543 0.9151  -5.277 3.1683  False
     Old Very Old   -3.676 0.1113 -7.8987 0.5467  False
     Old    Young   -2.103 0.5659 -6.3257 2.1197  False
Very Old    Young    1.573 0.7662 -2.6497 5.7957  False
----------------------------------------------------
```

Mean difference is mean_group2 - mean_group1, so we have a statistical difference in BMI means between Old and Young category. Young's BMI is -3 which is 4 lower than Old's BMI.

5) Two-Way ANOVA

We use two-way ANOVA to find the relation between Age Category and Physical Activity with BMI.

In [22]:

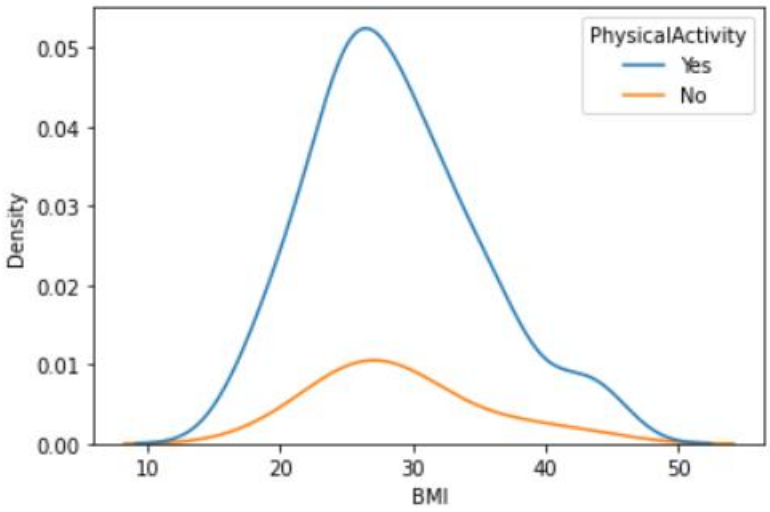
```
import statsmodels.api as sm
from statsmodels.formula.api import ols

model = ols('BMI ~ C(AgeCategory) + C(PhysicalActivity) ',
            data=df_sampled).fit()
sm.stats.anova_lm(model, typ=2)
```

Out[22]:

	sum_sq	df	F	PR(>F)
C(AgeCategory)	226.583956	3.0	1.904809	0.132718
C(PhysicalActivity)	6.261634	1.0	0.157918	0.691817
Residual	4559.888403	115.0	NaN	NaN

We observe that there is no significant difference in BMI mean from people with physical activity and others.



From the above plot, we see that there isn't any significant difference of BMI mean due to physical activity.

Regression Analysis

1) PCA

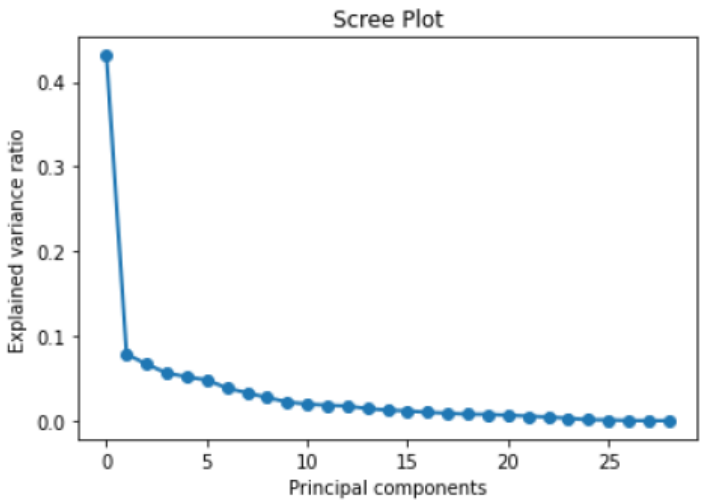
PCA is used for dimensionality reduction. Principal component analysis (PCA) is the process of computing the principal components and using them to perform a change of basis on the data, sometimes using only the first few principal components and ignoring the rest.

Here, we apply pca on our dataset to first find how many components are sufficient for 90% of covariance matrix.

In [60]:
```
start_pca=time.time()
from sklearn.decomposition import PCA
pca = PCA()
df_pca = pd.DataFrame(pca.fit_transform(scaled_df.T))
```

Scree-plot

In [61]:
```
plt.plot(pca.explained_variance_ratio_, 'o-', linewidth=2)
plt.title("Scree Plot")
plt.xlabel('Principal components')
plt.ylabel('Explained variance ratio')
plt.show()
```



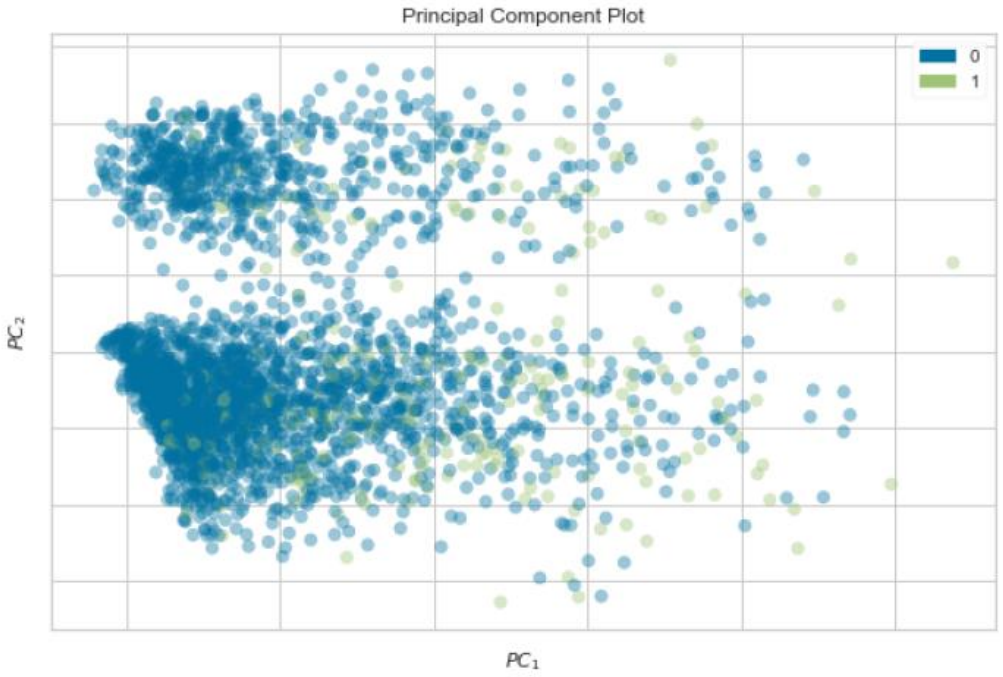
From the above scree plot, we see that the variance is high for the first two Principal Components and significantly reduces for the rest.

```
cumulative_sum = np.cumsum(pca.explained_variance_ratio_)
print("Cumulative Sum --> ", cumulative_sum)
```

```
Cumulative Sum -->  [0.43119535 0.50987826 0.5763854  0.63268324 0.68428818 0.73238754
 0.77075746 0.80361182 0.83088601 0.85286413 0.87241854 0.89050997
 0.90785456 0.92198788 0.93440184 0.9458333  0.9557441  0.96440894
 0.97197337 0.97933716 0.98573253 0.99098247 0.99517778 0.99787973
 0.99901296 0.99983203 1.         1.         1.         ]
```

We observe that 90% of the covariance matrix is explained in 13 PC's.

```
pca_visualisation_2d(df, "HeartDisease")
```



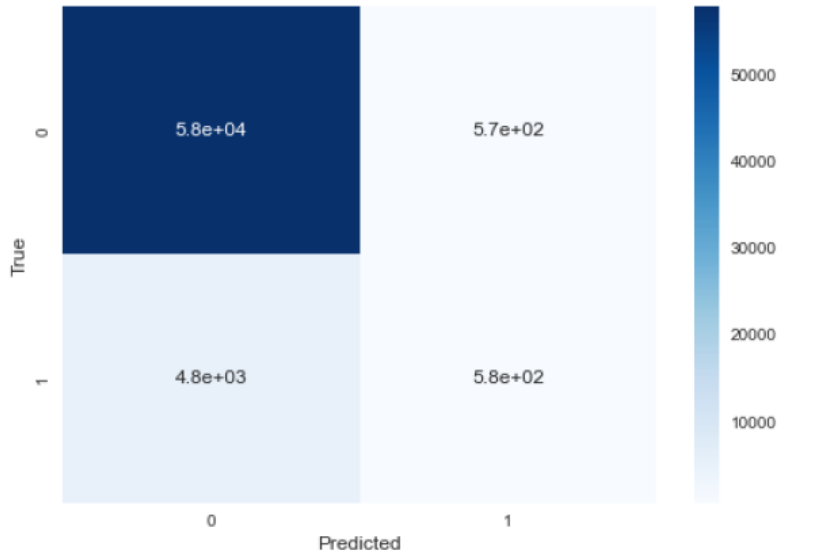
This is the scatter plot for the first two PC's of our dataset.

2) Logistic Regression using PCA

After dimensionality reduction using PCA, we used the reduced dataset to perform logistic regression and observe the train accuracy, test accuracy, precision, recall and F1 score.

```
Train accuracy: 0.91
Test accuracy: 0.92

Test confusion_matrix
Accuracy_score: 0.915273847308432
Precission_score: 0.5043478260869565
Recall_score: 0.10683367102597163
F1-score: 0.17631858945128437
Time taken for Logistic regression with PCA= 1.2476081848144531
```



We find the results of our metric test on logistic regression after PCA. We find no significant difference in accuracy before and after PCA for this dataset.

Conclusion

By using the heart disease dataset, we have performed statistical and regression analysis on some of the categorical data. We have applied concepts of Shapiro-Wilk test, Bartlett's test, One-way ANOVA, Tukey's test, and Two-Way ANOVA obtained various results. We have also performed regression using Logistic Regression and reduced the dimensions of our dataset using PCA. There are many other statistical analysis that can be performed using this detailed and informative dataset which will give us a better understanding on how various factors can contribute to heart disease.

References

[1] https://www.cdc.gov/heartdisease/facts

[2] https://www.who.int/health-topics/cardiovascular-diseases#tab=tab_1

[3] https://en.wikipedia.org