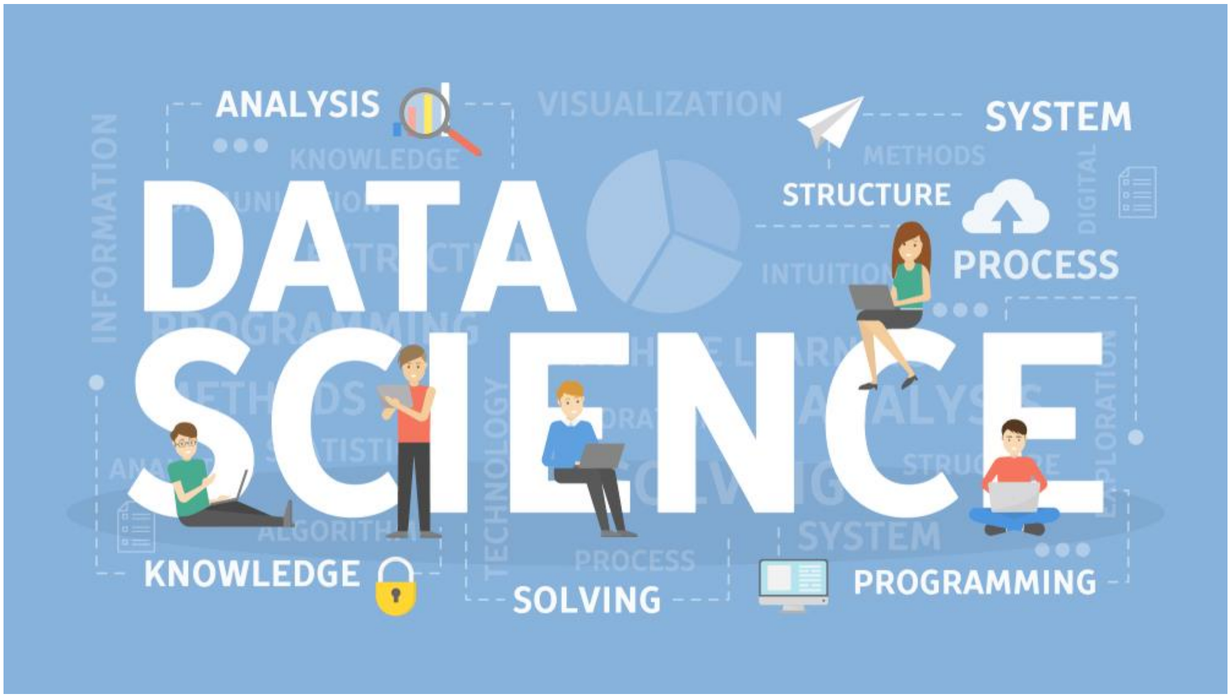
INTRODUCTION TO DATA SCIENCE PROJECT REPORT



Topic : CDC Diabetes Health Indicators

SUBMITTED BY:-

- Harshit Jain (21UCS091)
- Yash Chugh (21UCC234)
- Lokesh Dandwani (21UCS118)
- Smit Patel (21UCS205)

SUBMITTED TO:-

- Dr. Aloke Datta
- Dr. Lal Upendra Pratap Singh
- Dr. Subrat Dash

Table Of Contents

Problem Statement

The objective of this data science endeavor is to examine the CDC Diabetes Health Indicators dataset, encompassing healthcare statistics and lifestyle survey responses from individuals, coupled with their diabetes diagnoses. The dataset's overarching goal is to enhance our comprehension of the correlation between lifestyle choices and diabetes prevalence within the United States. The project entails multiple phases, such as data preprocessing, initial analysis, and the generation of a comprehensive report featuring various visual representations, including plots, charts, and graphs.

The project's core challenge involves delving into the dataset to discern patterns and trends, pinpoint significant factors linked to diabetes, and offer valuable insights into the intricate interplay between lifestyle choices and the occurrence of diabetes.

Dataset Description

The CDC Diabetes Health Indicators dataset is a tabular, multivariate dataset with 253,680 instances and 21 features. It contains healthcare statistics and lifestyle survey information about people in general along with their diagnosis of diabetes.

The dataset is associated with the subject area of health and medicine and is primarily used for classification tasks.

The features in the dataset consist of some demographics, lab test results, and answers to survey questions for each patient. The target variable for classification is whether a patient has diabetes, is pre-diabetic, or healthy. Feature types in the dataset include categorical and integer data.

The dataset was funded by the CDC. Each row represents a person participating in the study. Cross-validation or a fixed train-test split could be used.

The dataset contains sensitive data such as gender, income, and education level.

Dataset Information

Source of Data Set:

The CDC Diabetes Health Indicators dataset was obtained from the official website of the Centres for Disease Control and Prevention (CDC) at https://www.cdc.gov/brfss/annual_data/annual_2014.html.

The dataset is publicly accessible and was linked on 9/25/2023.

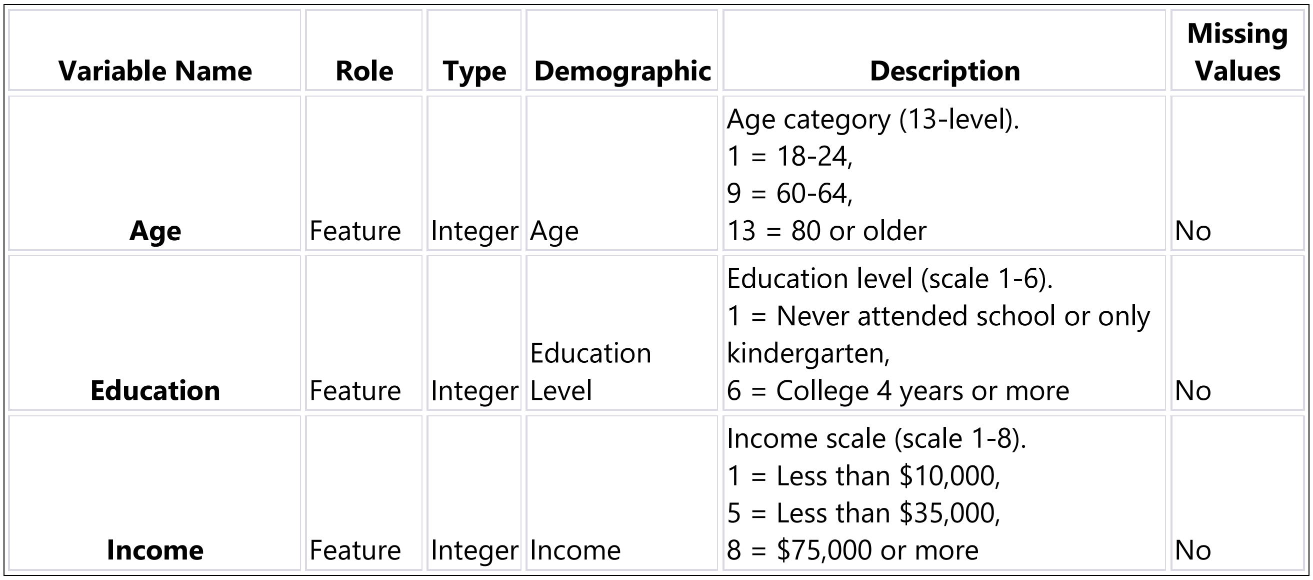
Data set specification:

Data Set Characteristics	Tabular, Multivariate
Feature type	Categorical, Integer
Associated Tasks	Classification
Number of Instances	253680
Number of attributes	21
Area	Health and Medicine

Attribute Information

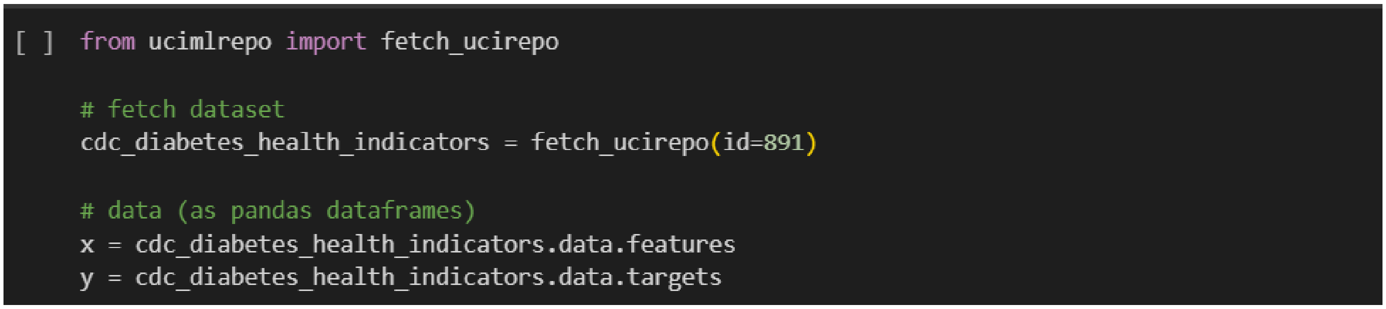
Variable Name	Role	Type	Demographic	Description	Missing Values
ID	Identifier	Integer	Patient ID	Unique identifier assigned to each patient.	No
Diabetes_binary	Target	Binary		Binary indicator for diabetes status. 0 = No diabetes, 1 = Prediabetes or diabetes	No
HighBP	Feature	Binary		Binary indicator for high blood pressure. 0 = No high blood pressure, 1 = High blood pressure	No
HighChol	Feature	Binary		Binary indicator for high cholesterol. 0 = No high cholesterol, 1 = High cholesterol	No
CholCheck	Feature	Binary		Binary indicator for cholesterol checks in the last 5 years. 0 = No cholesterol checks in 5 years, 1 = Cholesterol check in 5 years	No
BMI	Feature	Integer	Body Mass Index	Body Mass Index	No
Smoker	Feature	Binary		Smoking history. 0 = No, 1 = Yes	No
Stroke	Feature	Binary		History of stroke. 0 = No, 1 = Yes	No
HeartDiseaseorAttack	Feature	Binary		History of coronary heart disease (CHD) or myocardial infarction (MI). 0 = No, 1 = Yes	No
PhysActivity	Feature	Binary		Physical activity in the past 30 days - not including job.	No

Variable Name	Role	Type	Demographic	Description	Missing Values
				0 = No, 1 = Yes	
Fruits	Feature	Binary		Consumption of fruits 1 or more times per day. 0 = No, 1 = Yes	No
Veggies	Feature	Binary		Consumption of vegetables 1 or more times per day. 0 = No, 1 = Yes	No
HvyAlcoholConsump	Feature	Binary		Heavy alcohol consumption. 0 = No, 1 = Yes	No
AnyHealthcare	Feature	Binary		Having any kind of health care coverage, including health insurance. 0 = No, 1 = Yes	No
NoDocbcCost	Feature	Binary		Inability to see a doctor in the past 12 months due to cost. 0 = No, 1 = Yes	No
GenHlth	Feature	Integer	General Health	Self-reported general health on a scale of 1-5. 1 = Excellent, 2 = Very Good, 3 = Good, 4 = Fair, 5 = Poor	No
MentHlth	Feature	Integer	Mental Health	Number of days during the past 30 days when mental health was not good (scale 1-30 days).	No
PhysHlth	Feature	Integer	Physical Health	Number of days during the past 30 days when physical health was not good (scale 1-30 days).	No
DiffWalk	Feature	Binary		Serious difficulty walking or climbing stairs. 0 = No, 1 = Yes	No
Sex	Feature	Binary	Sex	Gender. 0 = Female, 1 = Male	No



Variable Name	Role	Type	Demographic	Description	Missing Values
Age	Feature	Integer	Age	Age category (13-level). 1 = 18-24, 9 = 60-64, 13 = 80 or older	No
Education	Feature	Integer	Education Level	Education level (scale 1-6). 1 = Never attended school or only kindergarten, 6 = College 4 years or more	No
Income	Feature	Integer	Income	Income scale (scale 1-8). 1 = Less than $10,000, 5 = Less than $35,000, 8 = $75,000 or more	No

Importing Dataset

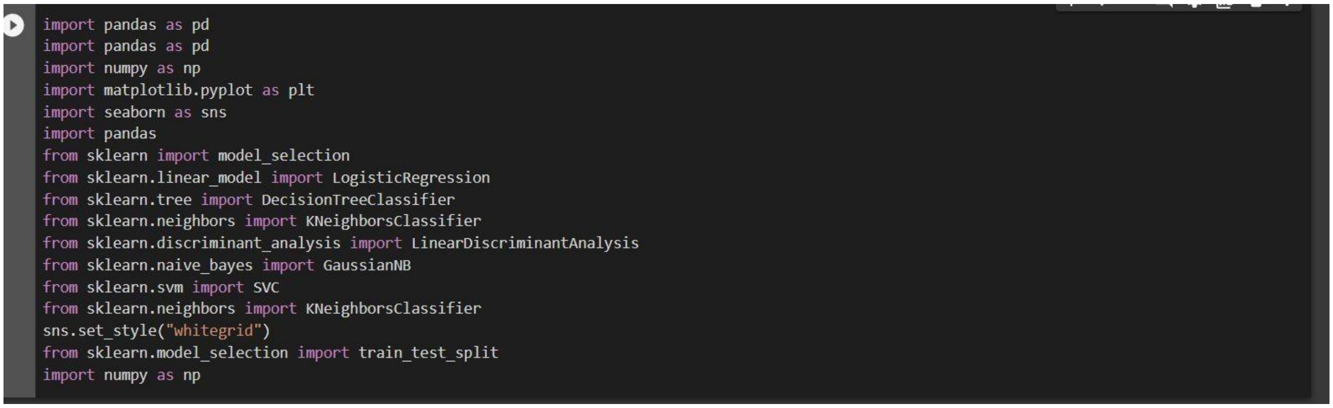


```
[ ]  from ucimlrepo import fetch_ucirepo

     # fetch dataset
     cdc_diabetes_health_indicators = fetch_ucirepo(id=891)

     # data (as pandas dataframes)
     x = cdc_diabetes_health_indicators.data.features
     y = cdc_diabetes_health_indicators.data.targets
```

Figure 1.

Firstly we have imported our dataset from ucimlrepo. Here,
x : feature variable
y : target variable.

Importing Libraries



```
import pandas as pd
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
import pandas
from sklearn import model_selection
from sklearn.linear_model import LogisticRegression
from sklearn.tree import DecisionTreeClassifier
from sklearn.neighbors import KNeighborsClassifier
from sklearn.discriminant_analysis import LinearDiscriminantAnalysis
from sklearn.naive_bayes import GaussianNB
from sklearn.svm import SVC
from sklearn.neighbors import KNeighborsClassifier
sns.set_style("whitegrid")
from sklearn.model_selection import train_test_split
import numpy as np
```

Figure 2.

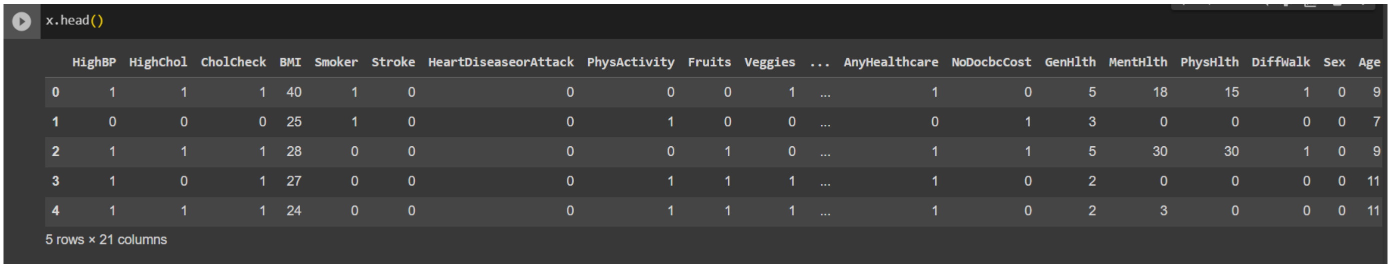
The subsequent action involves incorporating the diverse libraries utilized in our project at a certain juncture.

Data Preprocessing

The objective of data preprocessing is to enhance the data's quality and adapt it for the particular data mining task at hand. In this project, we will conduct data preprocessing on the CDC Diabetes Health Indicators dataset to rectify any discrepancies or gaps in the data.

Through these procedures, the data will undergo transformation into a format that is more efficiently and effectively handled in data mining, machine learning, and other data science endeavors. The methods employed in data preprocessing aim to guarantee precise results and elevate the overall data quality for thorough analysis.

- Visualizing the original data set



```
x.head()
```

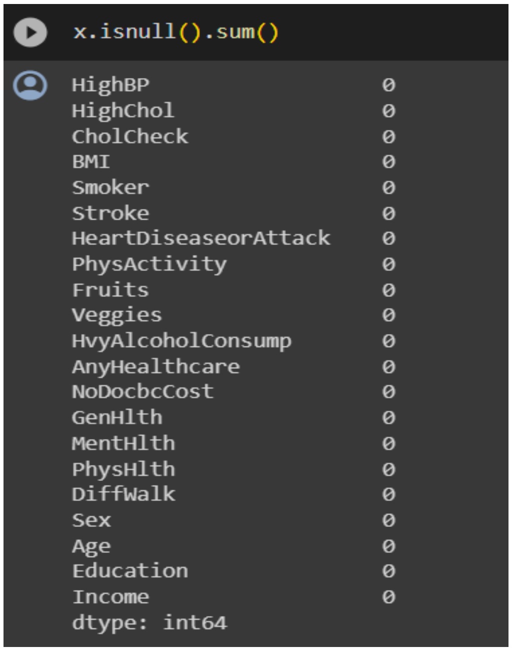
	HighBP	HighChol	CholCheck	BMI	Smoker	Stroke	HeartDiseaseorAttack	PhysActivity	Fruits	Veggies	...	AnyHealthcare	NoDocbcCost	GenHlth	MentHlth	PhysHlth	DiffWalk	Sex	Age
0	1	1	1	40	1	0	0	0	0	1	...	1	0	5	18	15	1	0	9
1	0	0	0	25	1	0	0	1	0	0	...	0	1	3	0	0	0	0	7
2	1	1	1	28	0	0	0	0	1	0	...	1	1	5	30	30	1	0	9
3	1	0	1	27	0	0	0	1	1	1	...	1	0	2	0	0	0	0	11
4	1	1	1	24	0	0	0	1	1	1	...	1	0	2	3	0	0	0	11

5 rows × 21 columns

Figure 3.

- Checking if the data set contains any Null values

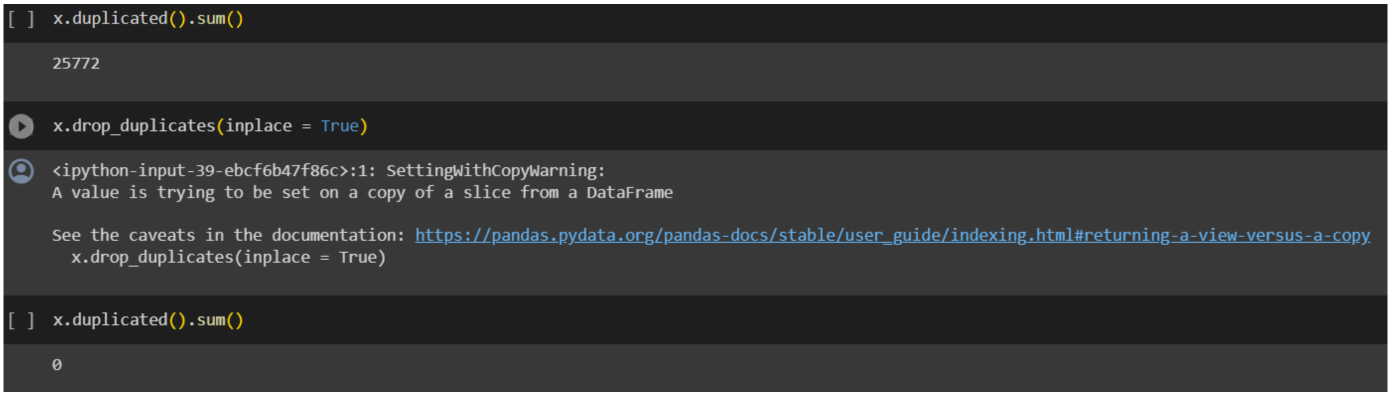
(isnull() is a function which checks for null values in the given data set, and returns Boolean for every attribute used in the data set)



```
x.isnull().sum()

HighBP                  0
HighChol                0
CholCheck               0
BMI                     0
Smoker                  0
Stroke                  0
HeartDiseaseorAttack    0
PhysActivity            0
Fruits                  0
Veggies                 0
HvyAlcoholConsump       0
AnyHealthcare           0
NoDocbcCost             0
GenHlth                 0
MentHlth                0
PhysHlth                0
DiffWalk                0
Sex                     0
Age                     0
Education               0
Income                  0
dtype: int64
```

Figure 4.

- Checking whether the data set contains any duplicate values.



```
[ ] x.duplicated().sum()

    25772

    x.drop_duplicates(inplace = True)

    <ipython-input-39-ebcf6b47f86c>:1: SettingWithCopyWarning:
    A value is trying to be set on a copy of a slice from a DataFrame

    See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy
      x.drop_duplicates(inplace = True)

[ ] x.duplicated().sum()

    0
```

Figure 5.

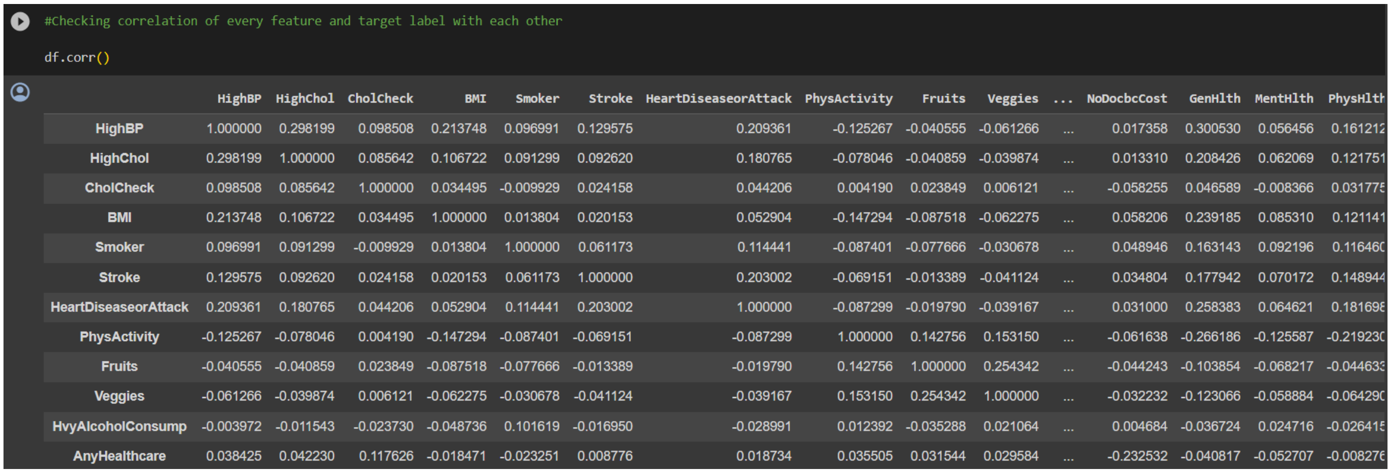
Here we have first checked all the duplicates and then dropped all the duplicates.

Correlation Calculation

- Positive Correlations: HighBP, HighChol, BMI, Stroke, HeartDisease, GenHlth, MentHlth, PhysHlth, DiffWalk, Age have positive correlations with Diabetes.
- Negative Correlations: PhysActivity, Fruits, Veggies, HvyAlcoholConsump, Education, Income show negative correlations with Diabetes.
- Weak Correlations: CholCheck, Smoker, AnyHealthcare, NoDocbcCost, Sex have weaker correlations with Diabetes.

- **Interpretation**: As positive correlation attributes increase, likelihood of Diabetes increases; as negative correlation attributes increase, likelihood decreases.

Cautions: Correlation doesn't imply causation. Consider additional factors and nonlinear relationships for a comprehensive analysis.



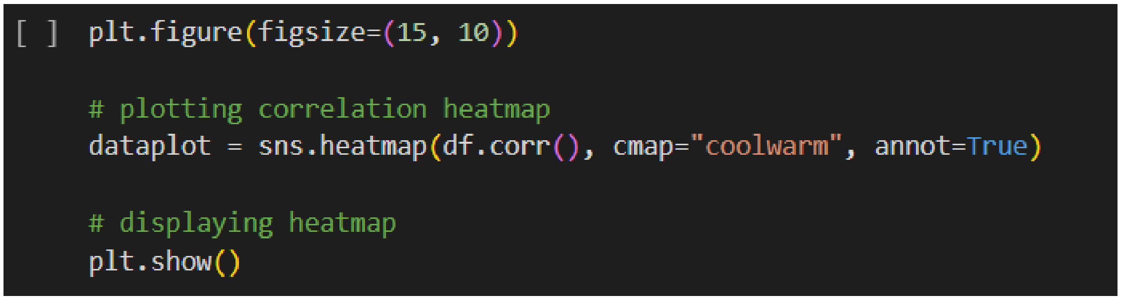
```
#Checking correlation of every feature and target label with each other

df.corr()
```

	HighBP	HighChol	CholCheck	BMI	Smoker	Stroke	HeartDiseaseorAttack	PhysActivity	Fruits	Veggies	...	NoDocbcCost	GenHlth	MentHlth	PhysHlth
HighBP	1.000000	0.298199	0.098508	0.213748	0.096991	0.129575	0.209361	-0.125267	-0.040555	-0.061266	...	0.017358	0.300530	0.056456	0.161212
HighChol	0.298199	1.000000	0.085642	0.106722	0.091299	0.092620	0.180765	-0.078046	-0.040859	-0.039874	...	0.013310	0.208426	0.062069	0.121751
CholCheck	0.098508	0.085642	1.000000	0.034495	-0.009929	0.024158	0.044206	0.004190	0.023849	0.006121	...	-0.058255	0.046589	-0.008366	0.031775
BMI	0.213748	0.106722	0.034495	1.000000	0.013804	0.020153	0.052904	-0.147294	-0.087518	-0.062275	...	0.058206	0.239185	0.085310	0.121141
Smoker	0.096991	0.091299	-0.009929	0.013804	1.000000	0.061173	0.114441	-0.087401	-0.077666	-0.030678	...	0.048946	0.163143	0.092196	0.116460
Stroke	0.129575	0.092620	0.024158	0.020153	0.061173	1.000000	0.203002	-0.069151	-0.013389	-0.041124	...	0.034804	0.177942	0.070172	0.148944
HeartDiseaseorAttack	0.209361	0.180765	0.044206	0.052904	0.114441	0.203002	1.000000	-0.087299	-0.019790	-0.039167	...	0.031000	0.258383	0.064621	0.181698
PhysActivity	-0.125267	-0.078046	0.004190	-0.147294	-0.087401	-0.069151	-0.087299	1.000000	0.142756	0.153150	...	-0.061638	-0.266186	-0.125587	-0.219230
Fruits	-0.040555	-0.040859	0.023849	-0.087518	-0.077666	-0.013389	-0.019790	0.142756	1.000000	0.254342	...	-0.044243	-0.103854	-0.068217	-0.044633
Veggies	-0.061266	-0.039874	0.006121	-0.062275	-0.030678	-0.041124	-0.039167	0.153150	0.254342	1.000000	...	-0.032232	-0.123066	-0.058884	-0.064290
HvyAlcoholConsump	-0.003972	-0.011543	-0.023730	-0.048736	0.101619	-0.016950	-0.028991	0.012392	-0.035288	0.021064	...	0.004684	-0.036724	0.024716	-0.026415
AnyHealthcare	0.038425	0.042230	0.117626	-0.018471	-0.023251	0.008776	0.018734	0.035505	0.031544	0.029584	...	-0.232532	-0.040817	-0.052707	-0.008276

Figure 6.

Heatmap Visualization

To provide a comprehensive view of the correlation matrix, we employed the heatmap. The heatmap visually represents the strength of correlations between features. Features with higher positive correlations are represented in warmer colors, while features with higher negative correlations are represented in cooler colors.



```
[ ]  plt.figure(figsize=(15, 10))

     # plotting correlation heatmap
     dataplot = sns.heatmap(df.corr(), cmap="coolwarm", annot=True)

     # displaying heatmap
     plt.show()
```

Figure 7.

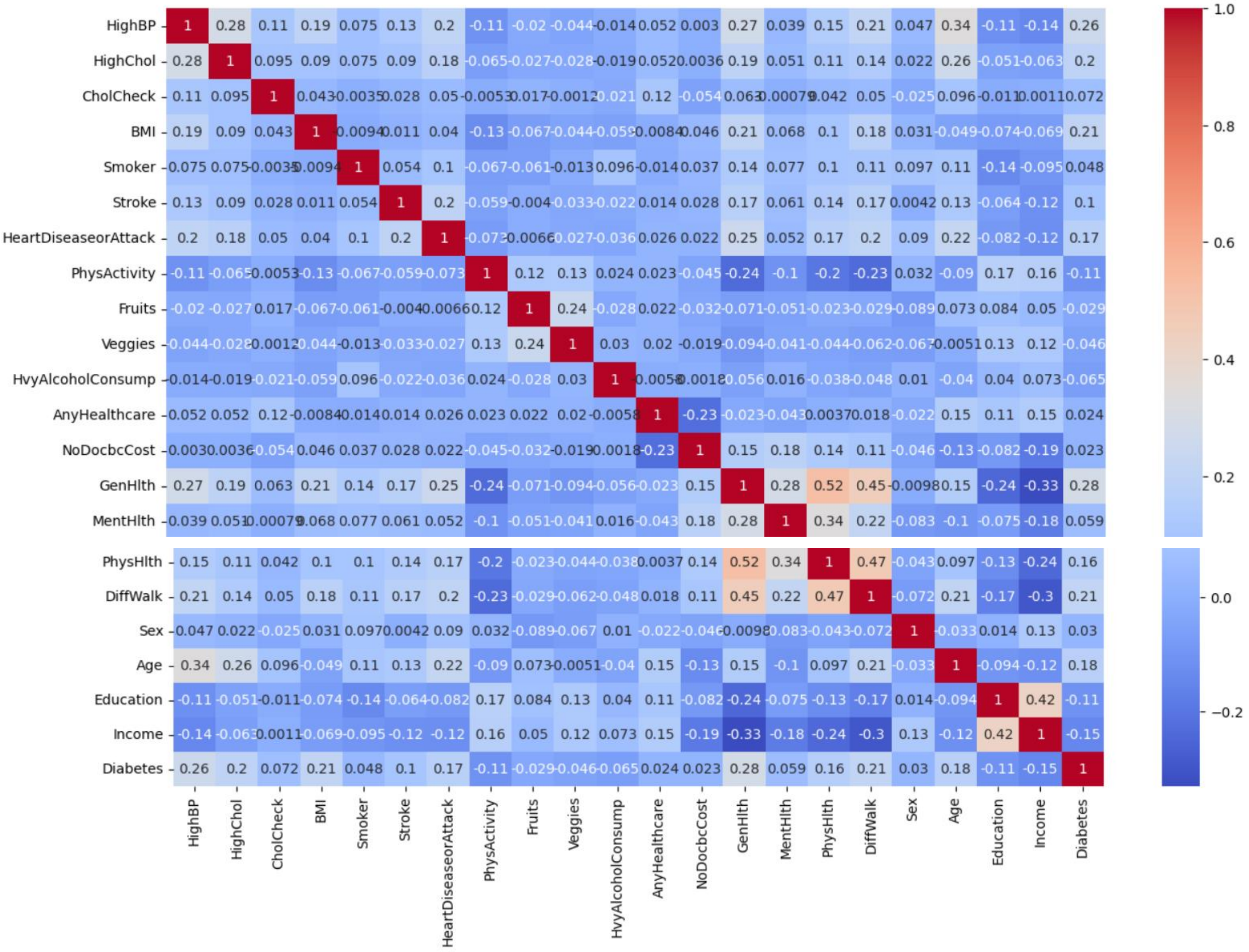


Figure 8.

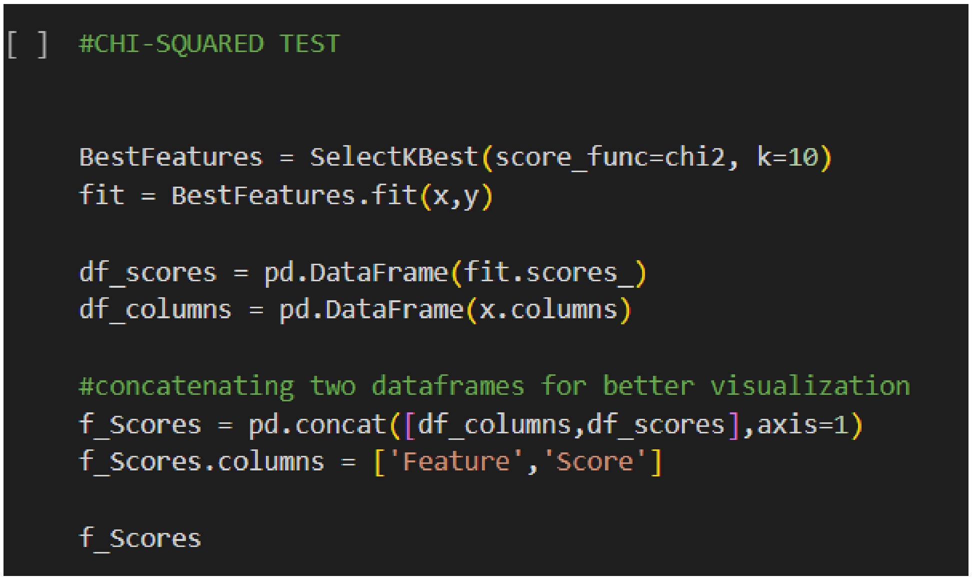
Chi-Squared Test for Categorical Variables

The chi-squared test is a statistical method used to determine if there is a significant association between two categorical variables. In our analysis, we applied the chi-squared test to assess the independence of certain categorical features within the CDC Diabetes Health Indicators dataset.

Some Inferences:

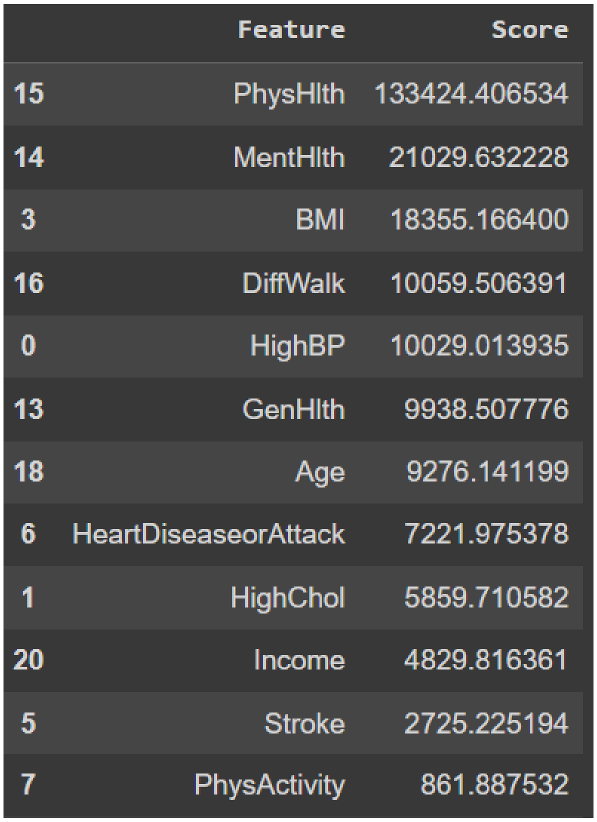
PhysHlth (Physical Health):
Highest chi-squared score i.e, strongly associated with 'Diabetes.
Indicates that physical health is quite important predicting diabetes.

PhysActivity (Physical Activity):
Moderate chi-squared score and negative correlation with 'Diabetes.
Highlights the protective effect of physical activity against diabetes.



```
[ ]  #CHI-SQUARED TEST


     BestFeatures = SelectKBest(score_func=chi2, k=10)
     fit = BestFeatures.fit(x,y)

     df_scores = pd.DataFrame(fit.scores_)
     df_columns = pd.DataFrame(x.columns)

     #concatenating two dataframes for better visualization
     f_Scores = pd.concat([df_columns,df_scores],axis=1)
     f_Scores.columns = ['Feature','Score']

     f_Scores
```

Figure 9.

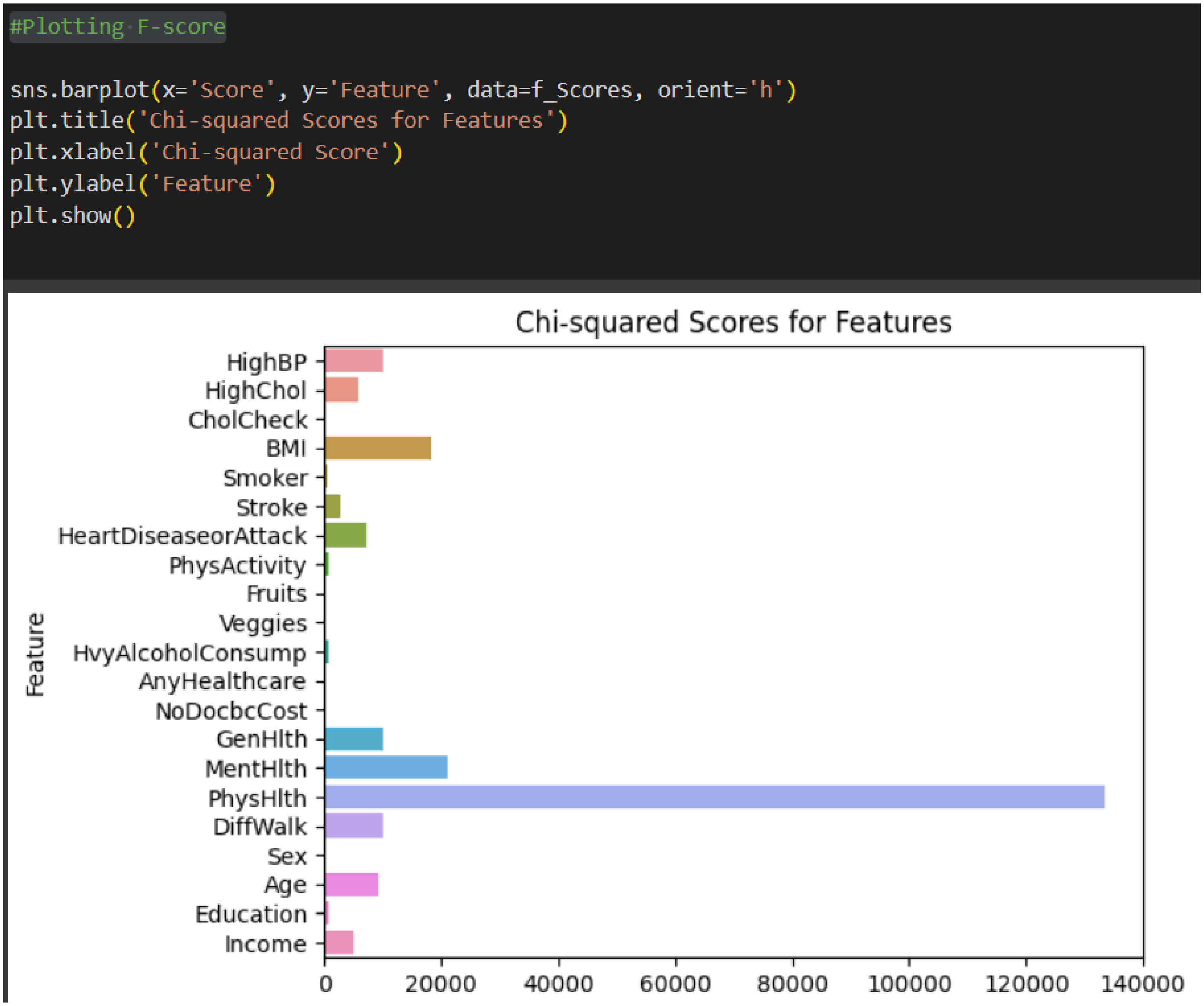


	Feature	Score
15	PhysHlth	133424.406534
14	MentHlth	21029.632228
3	BMI	18355.166400
16	DiffWalk	10059.506391
0	HighBP	10029.013935
13	GenHlth	9938.507776
18	Age	9276.141199
6	HeartDiseaseorAttack	7221.975378
1	HighChol	5859.710582
20	Income	4829.816361
5	Stroke	2725.225194
7	PhysActivity	861.887532

Figure 10.

Plotting F-score (Bar-Plot)

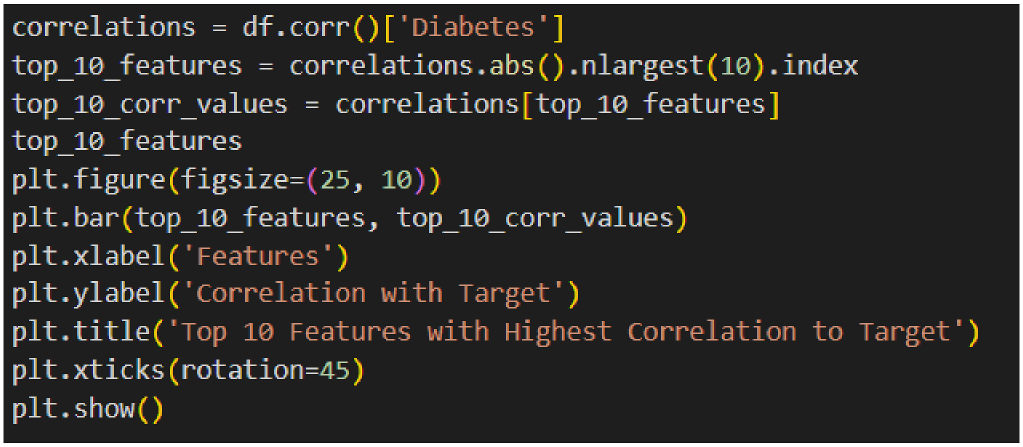
The term "F score" commonly denotes the chi-squared statistic computed for individual features during the process of feature selection employing the chi-squared test. The chi-squared statistic assesses the extent of association between categorical variables.. It aids in evaluating the relationship between each distinct feature and the target variable in the context of feature selection.



```
#Plotting F-score

sns.barplot(x='Score', y='Feature', data=f_Scores, orient='h')
plt.title('Chi-squared Scores for Features')
plt.xlabel('Chi-squared Score')
plt.ylabel('Feature')
plt.show()
```

Figure 11.

Plotting the top 10 features which have the highest correlation (Bar-Plot)



```
correlations = df.corr()['Diabetes']
top_10_features = correlations.abs().nlargest(10).index
top_10_corr_values = correlations[top_10_features]
top_10_features
plt.figure(figsize=(25, 10))
plt.bar(top_10_features, top_10_corr_values)
plt.xlabel('Features')
plt.ylabel('Correlation with Target')
plt.title('Top 10 Features with Highest Correlation to Target')
plt.xticks(rotation=45)
plt.show()
```

Figure 12.

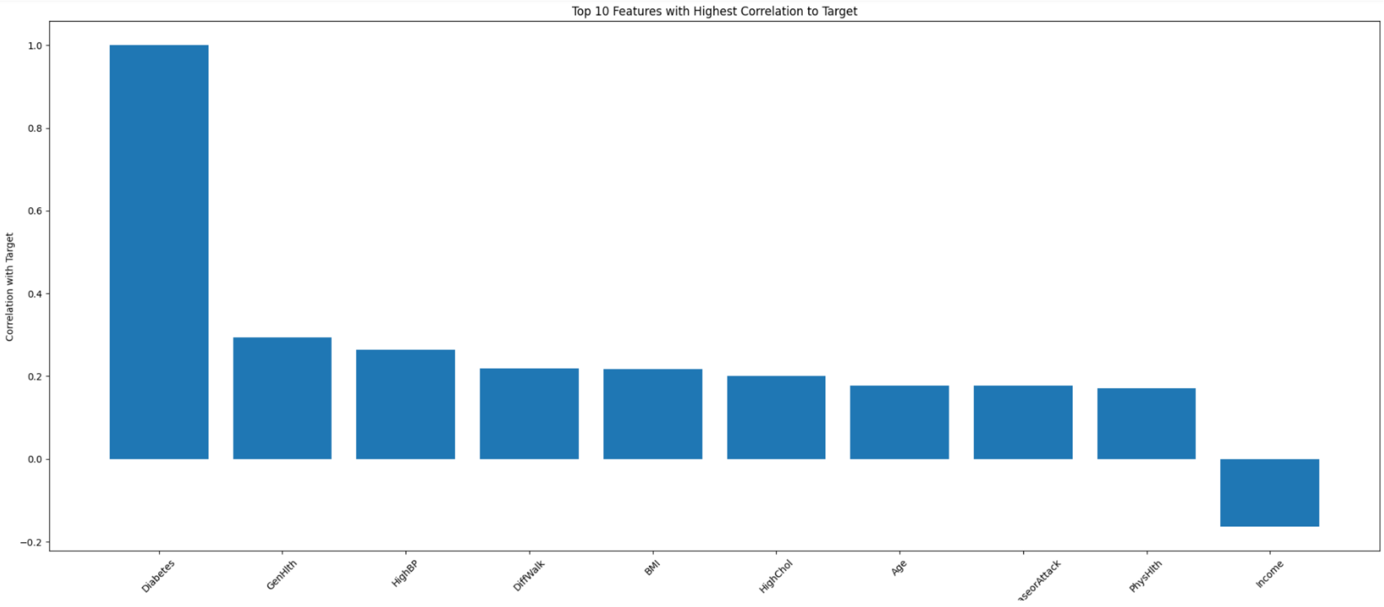


Figure 13.

Now based on the correlation we have calculated the top 9 factors that have the highest impact on determining whether a person is likely to have diabetes or not.

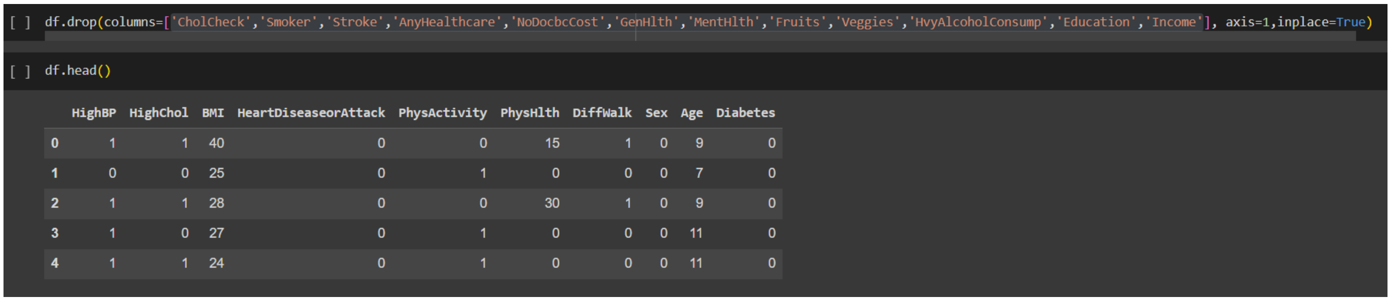
These 9 attributes are:-

1.)General Health 2).High BP 3.)Difficulty in Walking 4.)BMI

5.)High cholesterol 6.)Age 7.)Heart Disease Attack

8.)Physical Health 9.)Income

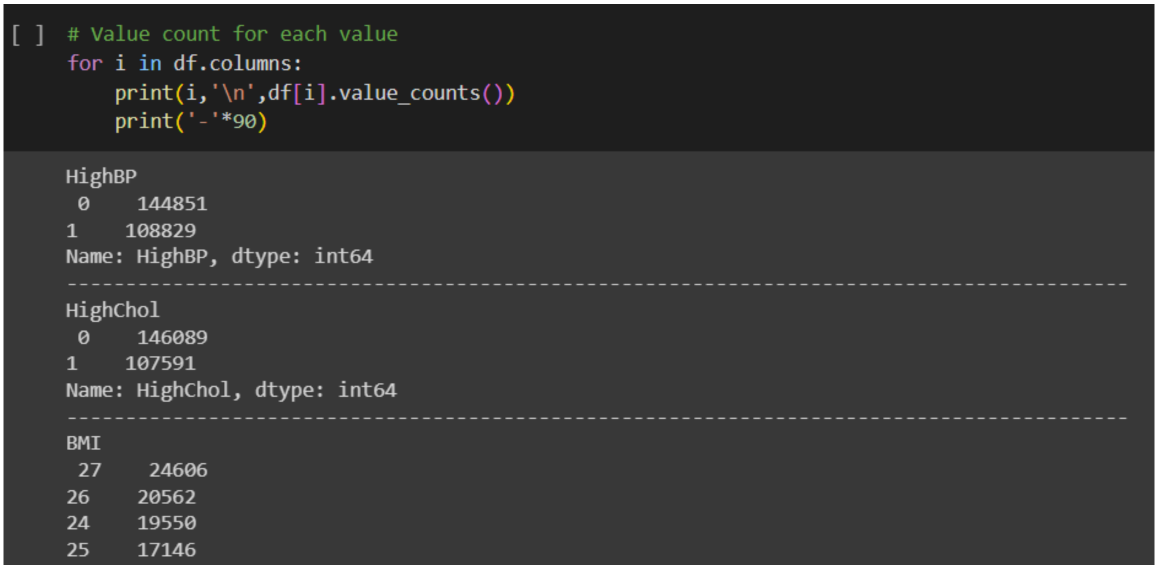
- Dropping the features that have very low correlation with the target.



```
[ ] df.drop(columns=['CholCheck','Smoker','Stroke','AnyHealthcare','NoDocbcCost','GenHlth','MentHlth','Fruits','Veggies','HvyAlcoholConsump','Education','Income'], axis=1,inplace=True)

[ ] df.head()
```

	HighBP	HighChol	BMI	HeartDiseaseorAttack	PhysActivity	PhysHlth	DiffWalk	Sex	Age	Diabetes
0	1	1	40	0	0	15	1	0	9	0
1	0	0	25	0	1	0	0	0	7	0
2	1	1	28	0	0	30	1	0	9	0
3	1	0	27	0	1	0	0	0	11	0
4	1	1	24	0	1	0	0	0	11	0

Figure 14.



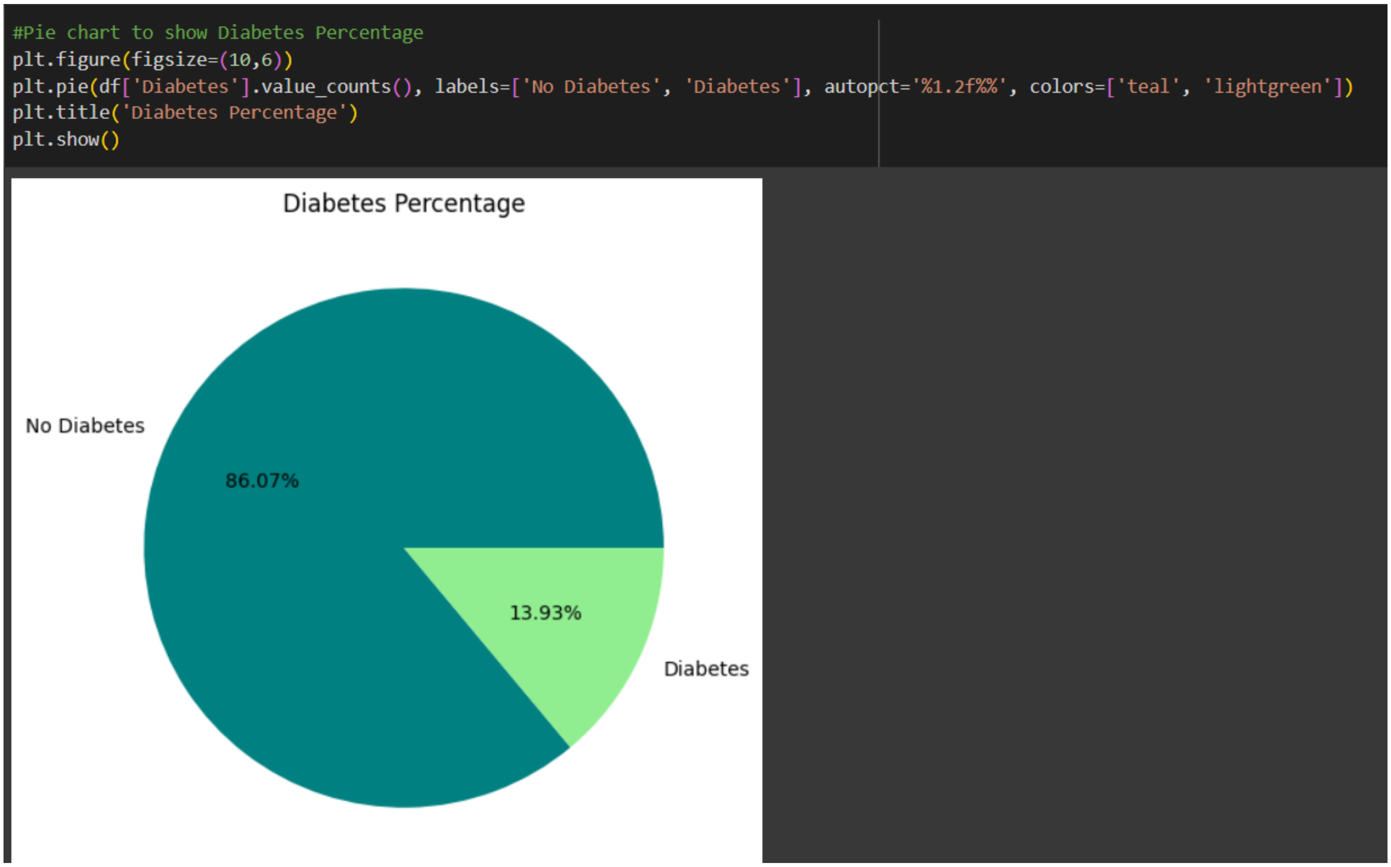
```
[ ] # Value count for each value
    for i in df.columns:
        print(i,'\n',df[i].value_counts())
        print('-'*90)

    HighBP
     0    144851
    1    108829
    Name: HighBP, dtype: int64
    ------------------------------------------------------------------------------------------
    HighChol
     0    146089
    1    107591
    Name: HighChol, dtype: int64
    ------------------------------------------------------------------------------------------
    BMI
     27    24606
    26    20562
    24    19550
    25    17146
```

Figure 15.

EDA(EXPLORATORY DATA ANALYSIS)

Exploratory Data Analysis stands as a pivotal stage in comprehending the attributes and trends inherent in a dataset. In this section, we delve into key aspects of the CDC Diabetes Health Indicators dataset to gain insights and guide further analysis.

Pie chart (Diabetes and No-Diabetes)



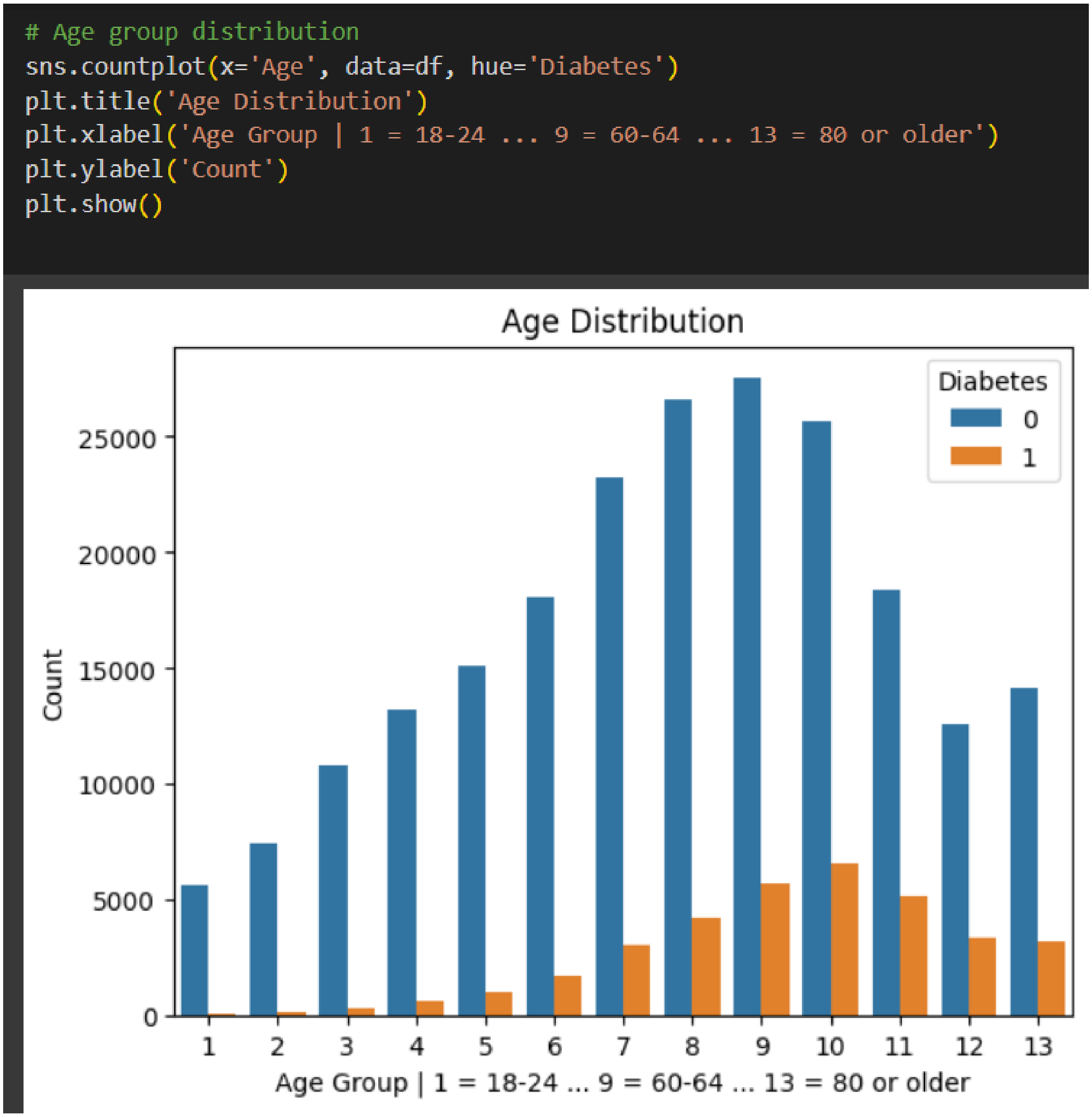
```
#Pie chart to show Diabetes Percentage
plt.figure(figsize=(10,6))
plt.pie(df['Diabetes'].value_counts(), labels=['No Diabetes', 'Diabetes'], autopct='%1.2f%%', colors=['teal', 'lightgreen'])
plt.title('Diabetes Percentage')
plt.show()
```

Figure 16.

Inferences drawn from the pie chart:

- 13.93% of individuals within the dataset are diagnosed with diabetes.
- 86.07% of the individuals in the dataset do not have diabetes.

Age group distribution(Bar-Plot)



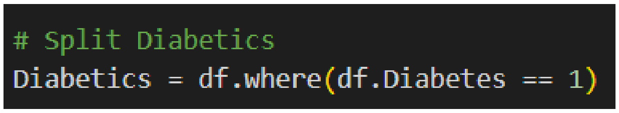
```
# Age group distribution
sns.countplot(x='Age', data=df, hue='Diabetes')
plt.title('Age Distribution')
plt.xlabel('Age Group | 1 = 18-24 ... 9 = 60-64 ... 13 = 80 or older')
plt.ylabel('Count')
plt.show()
```

Figure 17.

Observations derived from the age distribution graph of individuals with diabetes:

- Diabetes exhibits a higher occurrence among older adults.
- The prevalence of diabetes escalates with advancing age.
- The 80-or-older age group demonstrates the highest diabetes prevalence.
- Significant diabetes prevalence is also notable in the 60-80 age range.
- Although lower, diabetes prevalence remains noteworthy in younger adults within the analysed dataset.

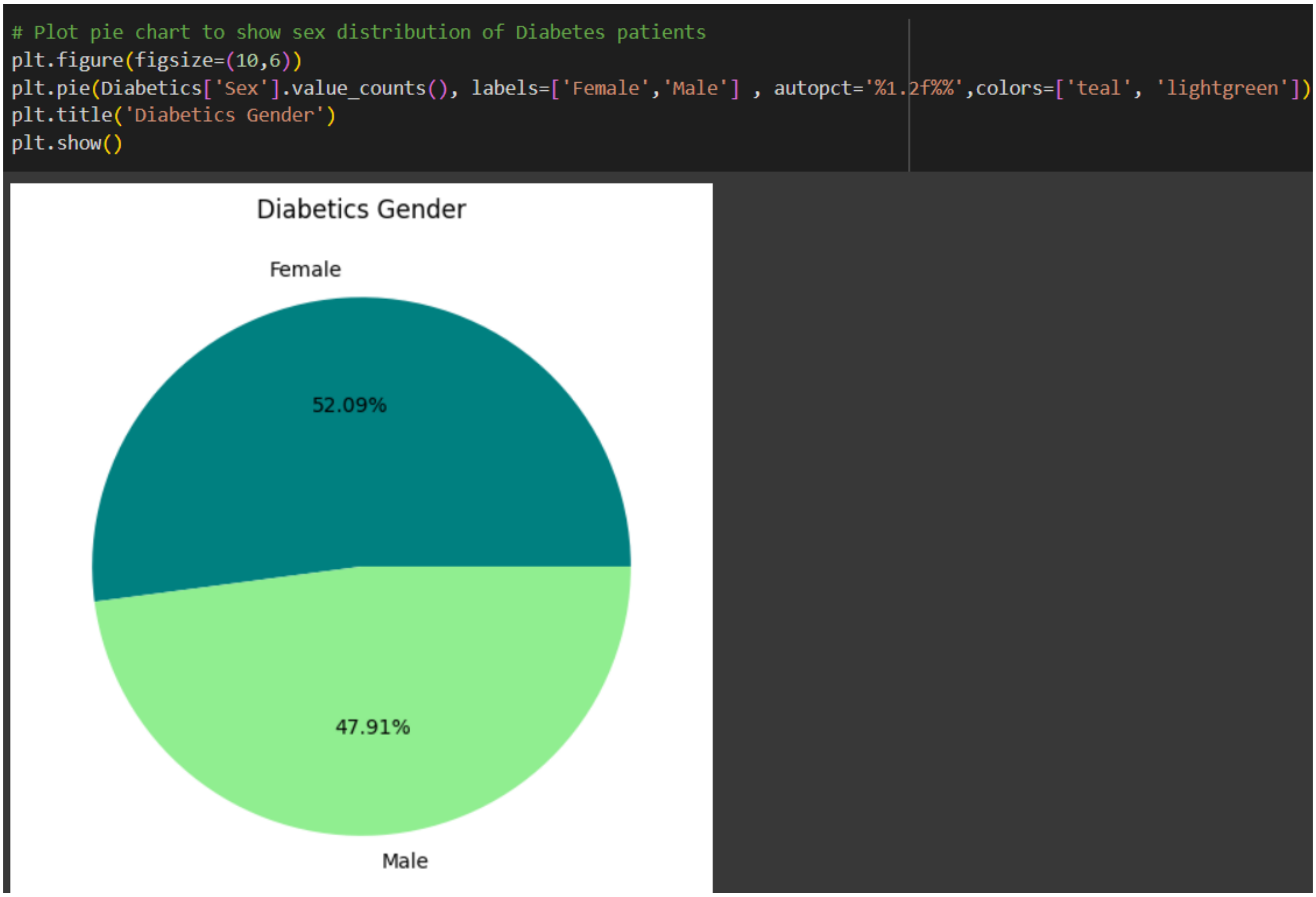
Now we are doing analysis for only those people who have diabetes.



```
# Split Diabetics
Diabetics = df.where(df.Diabetes == 1)
```

Figure 18.

Pie chart to show the sex distribution of Diabetes patients.



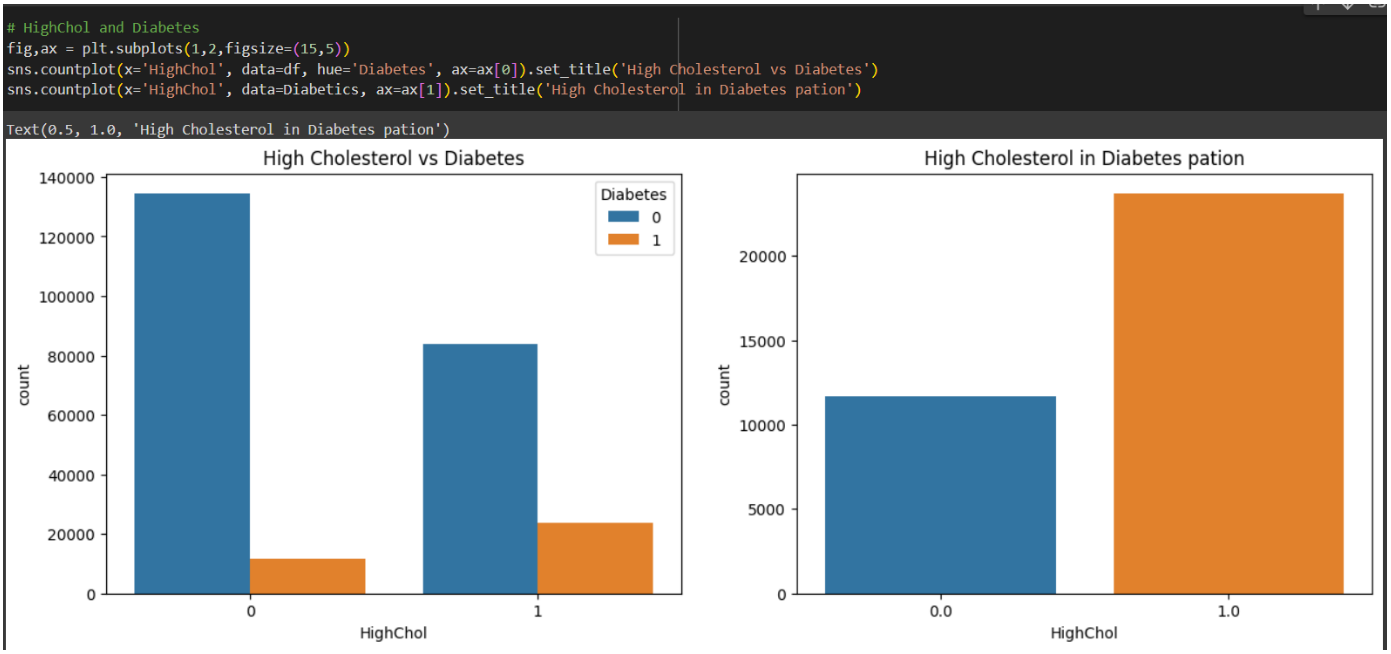
```
# Plot pie chart to show sex distribution of Diabetes patients
plt.figure(figsize=(10,6))
plt.pie(Diabetics['Sex'].value_counts(), labels=['Female','Male'] , autopct='%1.2f%%',colors=['teal', 'lightgreen'])
plt.title('Diabetics Gender')
plt.show()
```

Figure 19.

The breakdown in the pie chart suggests that a majority of those identified with diabetes are female, comprising 52.09%, while males make up 47.91%. This deduction is drawn from the data showcased in the pie chart and is in line with findings observed in various studies indicating a higher likelihood of diabetes diagnosis in women compared to men.

There are various conceivable explanations for this gender contrast in diabetes prevalence. One plausible factor is the increased likelihood of women being overweight or obese, a significant risk factor for diabetes. Another potential factor is the heightened probability of women encountering gestational diabetes, a distinct form of diabetes that can emerge during pregnancy.

Plot for High Cholesterol and Diabetes



```
# HighChol and Diabetes
fig,ax = plt.subplots(1,2,figsize=(15,5))
sns.countplot(x='HighChol', data=df, hue='Diabetes', ax=ax[0]).set_title('High Cholesterol vs Diabetes')
sns.countplot(x='HighChol', data=Diabetics, ax=ax[1]).set_title('High Cholesterol in Diabetes pation')
```

```
Text(0.5, 1.0, 'High Cholesterol in Diabetes pation')
```

Figure 20.

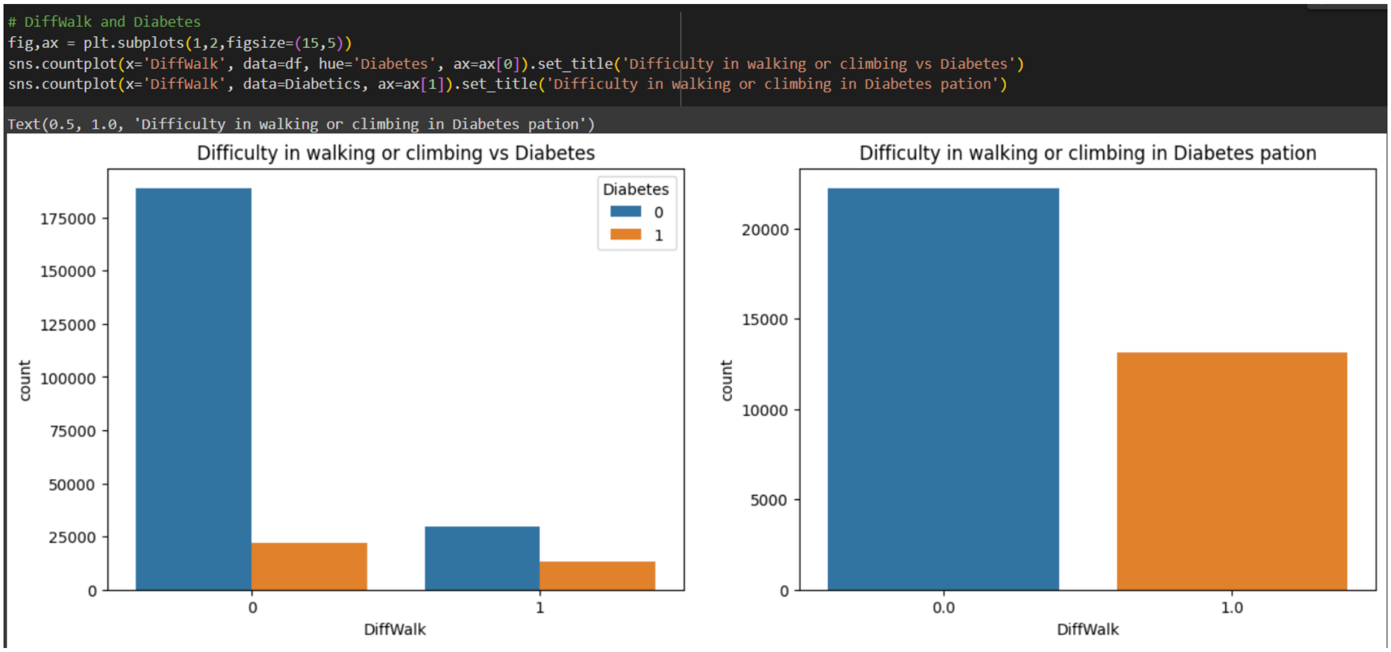
Insights drawn from the graph depicting the count of individuals with diabetes based on their cholesterol levels:

- The count of individuals with diabetes exhibiting high cholesterol surpasses those with low cholesterol.

- A predominant proportion of individuals with diabetes manifests high cholesterol.

- High cholesterol emerges as a significant risk factor for diabetes.
- Individuals diagnosed with diabetes are advised to proactively address their cholesterol levels through measures like adopting a nutritious diet, engaging in regular exercise, and considering medication if necessary.

Plot for Difficulty in Walking and Diabetes



```
# DiffWalk and Diabetes
fig,ax = plt.subplots(1,2,figsize=(15,5))
sns.countplot(x='DiffWalk', data=df, hue='Diabetes', ax=ax[0]).set_title('Difficulty in walking or climbing vs Diabetes')
sns.countplot(x='DiffWalk', data=Diabetics, ax=ax[1]).set_title('Difficulty in walking or climbing in Diabetes pation')
```

```
Text(0.5, 1.0, 'Difficulty in walking or climbing in Diabetes pation')
```

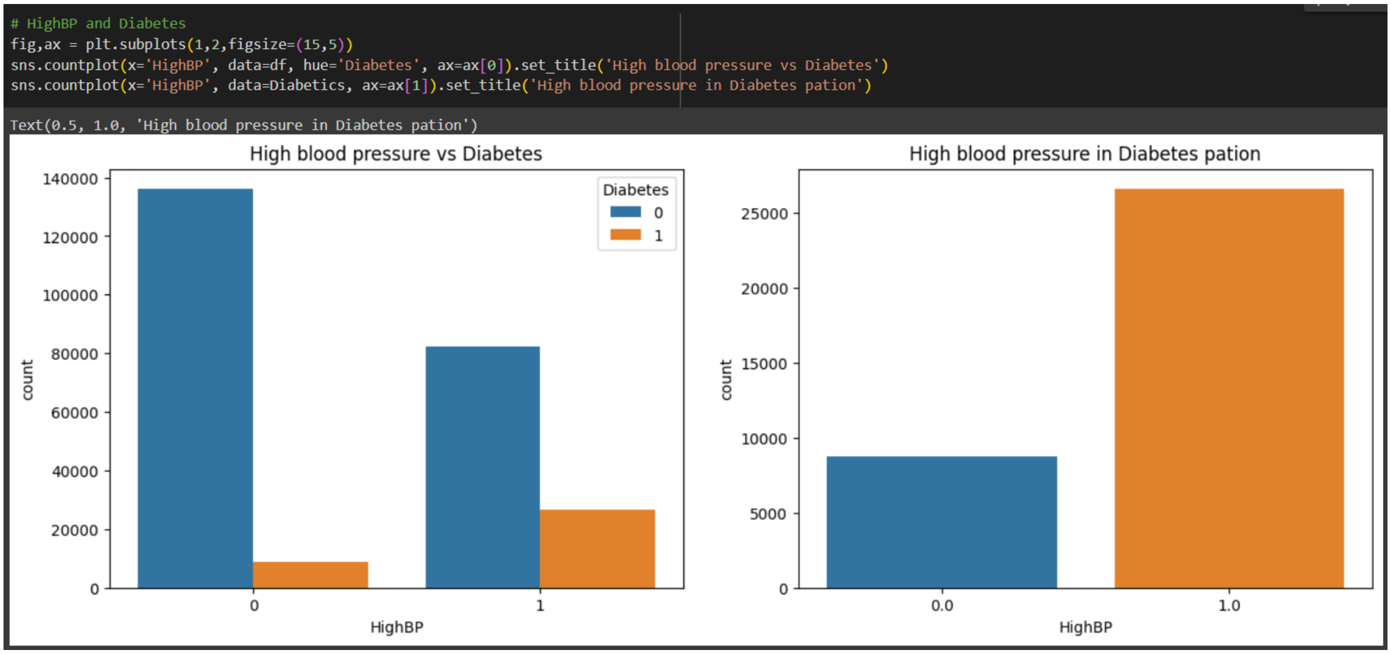
Figure 21.

Inferred from the graphical representation, the subsequent conclusions can be drawn:

- Individuals diagnosed with diabetes are more susceptible to facing obstacles in walking or climbing in comparison to those without diabetes.

- The utmost prevalence of challenges in walking or climbing is noted among individuals with diabetes who demonstrate severe neuropathy.

Plot for High BP and Diabetes



```
# HighBP and Diabetes
fig,ax = plt.subplots(1,2,figsize=(15,5))
sns.countplot(x='HighBP', data=df, hue='Diabetes', ax=ax[0]).set_title('High blood pressure vs Diabetes')
sns.countplot(x='HighBP', data=Diabetics, ax=ax[1]).set_title('High blood pressure in Diabetes pation')
```

```
Text(0.5, 1.0, 'High blood pressure in Diabetes pation')
```

Figure 22.

Conclusions drawn from the visual depiction are outlined below:

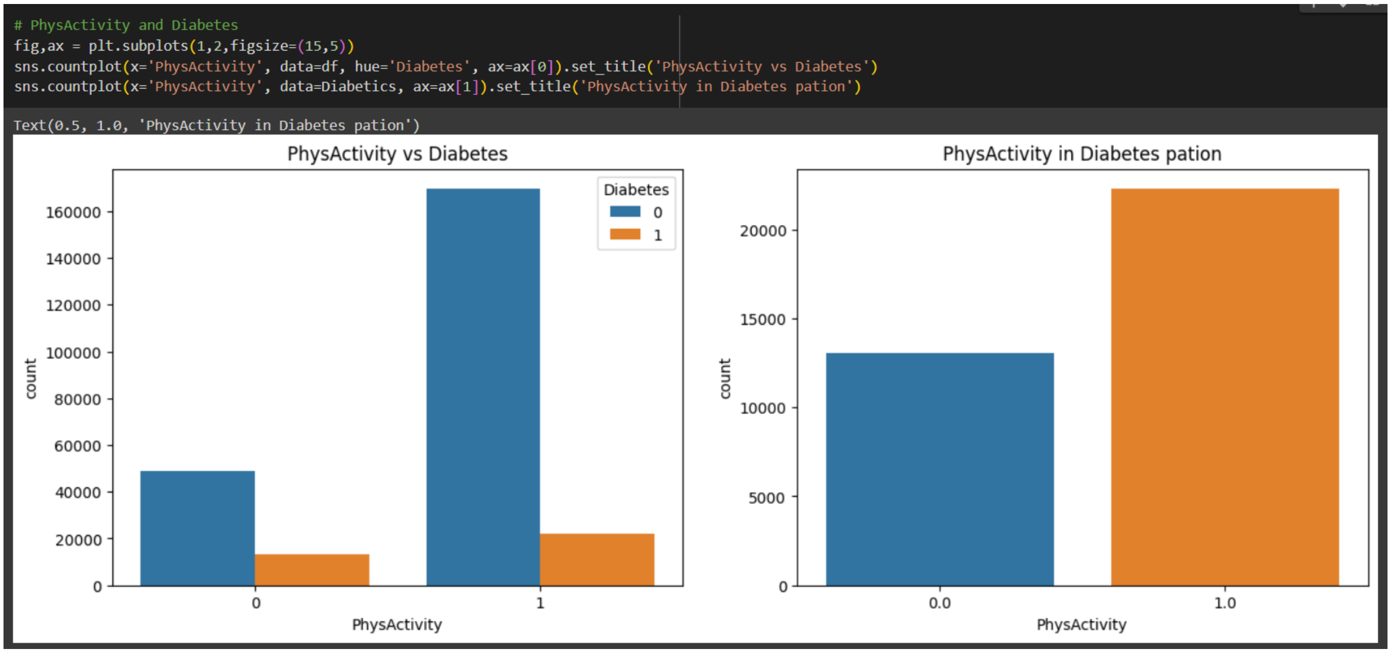
- Diabetes is identified as a risk factor for heightened blood pressure.

- Those with diabetes are more pre-disposed to encountering high blood pressure when contrasted with individuals without diabetes.

- The probability of experiencing elevated blood pressure increases with the severity of diabetes.

- Individuals with a diabetes diagnosis are advised to implement strategies for controlling their blood pressure, encompassing adherence to a nutritious diet, regular engagement in exercise, and potential consideration of medication.

Plot for Physical Activity and Diabetes



```
# PhysActivity and Diabetes
fig,ax = plt.subplots(1,2,figsize=(15,5))
sns.countplot(x='PhysActivity', data=df, hue='Diabetes', ax=ax[0]).set_title('PhysActivity vs Diabetes')
sns.countplot(x='PhysActivity', data=Diabetics, ax=ax[1]).set_title('PhysActivity in Diabetes pation')
```

```
Text(0.5, 1.0, 'PhysActivity in Diabetes pation')
```

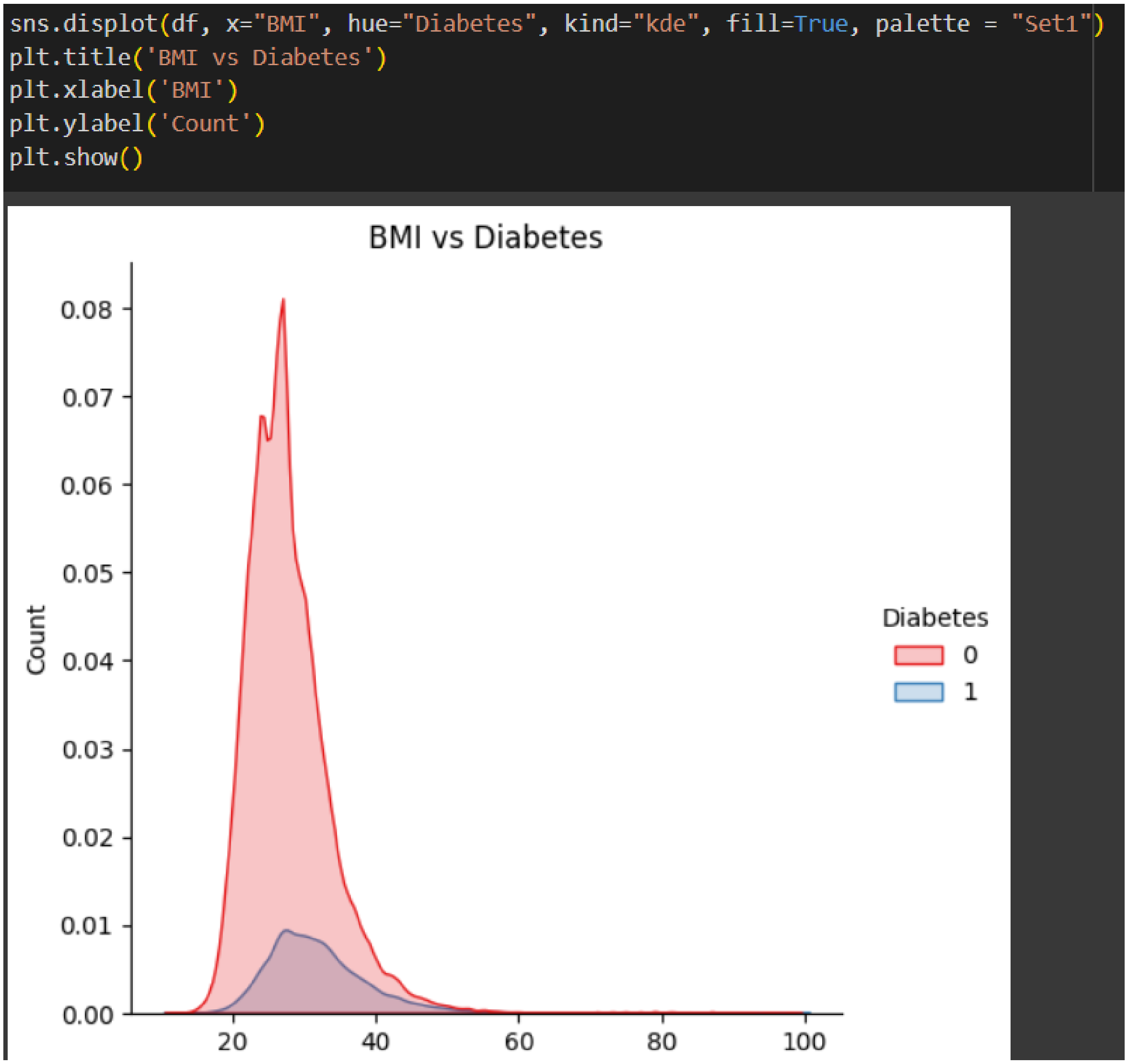
Figure 23.

Inferences derived from the visual representation are summarized as follows:

- A distinct positive correlation is observed between diabetes and levels of physical activity.

- Those with higher activity levels exhibit a diminished susceptibility to diabetes.
- Engaging in physical activity is recognized as both a preventive measure against diabetes and a strategic approach for managing the condition in individuals already diagnosed with it.

Plot for BMI and Diabetes



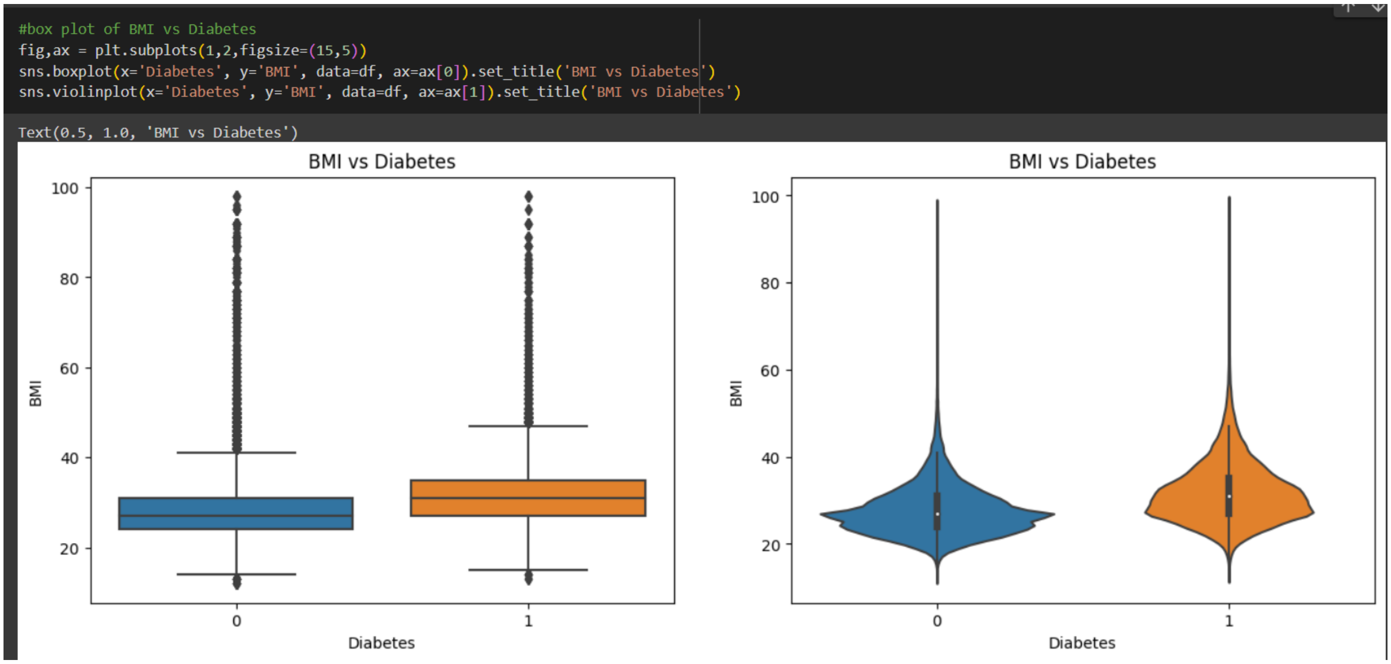
```
sns.displot(df, x="BMI", hue="Diabetes", kind="kde", fill=True, palette = "Set1")
plt.title('BMI vs Diabetes')
plt.xlabel('BMI')
plt.ylabel('Count')
plt.show()
```

Figure 24.

Inferences derived from the depiction of BMI vs. Diabetes include:

- A positive correlation is evident between BMI and diabetes.
- Individuals with elevated BMIs are more prone to diabetes.
- Obesity emerges as a significant contributor to the risk of diabetes. These deductions align with findings from various studies indicating BMI as a robust predictor of diabetes risk. The association between obesity and diabetes risk is attributed to its potential to induce insulin resistance.

Box plot of BMI vs Diabetes



```
#box plot of BMI vs Diabetes
fig,ax = plt.subplots(1,2,figsize=(15,5))
sns.boxplot(x='Diabetes', y='BMI', data=df, ax=ax[0]).set_title('BMI vs Diabetes')
sns.violinplot(x='Diabetes', y='BMI', data=df, ax=ax[1]).set_title('BMI vs Diabetes')
```

```
Text(0.5, 1.0, 'BMI vs Diabetes')
```

Figure 25.

Inferences drawn from the box plot of BMI vs Diabetes:

- People with diabetes have a higher median BMI than people without diabetes.
- There is a wider range of BMI values in people with diabetes than in people without diabetes.
- Some people with diabetes have a normal BMI, but the majority have a BMI that is overweight or obese.
- Obesity is a major risk factor for diabetes, but not everyone with obesity will develop diabetes.

The box plot shows the distribution of BMI values for people with diabetes and people without diabetes. The median BMI for people with diabetes is 29.5, which is classified as overweight. The median BMI for people without diabetes is 25.1, which is classified as normal weight.

Conclusion

- Male and female are equally vulnerable for Diabetes.
- People older than 45 are more vulnerable for diabetes then the younger ones when the age increase the number of diabetic people also increase.
- More than half of the diabetics are obese , almost half of the pre diabetics are obese.
- Percentages of diabetics and pre diabetics who suffers from obesity and Overweight are much higher than percentage of non-diabetic who suffers from obesity and overweight.
- Genth has a major effect on diabetes. When General Health is not good then the risk of diabetes increases rapidly.

- Mental Health is a major factor which causes Diabetes. When Mental Health is not stable for long time then the risk of diabetes increases.
- Physical activity reduces the risk of diabetes.

REFERENCES

- https://scikit-learn.org/stable

- https://seaborn.pydata.org/introduction.html

- https://matplotlib.org/3.1.1/tutorials/index.html

- https://medium.com/@kaanerdenn/exploratory-data-analysis-eda-bf98e8586e6f