In [ ]:
```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
```

In [ ]:
```
# reading in our dataset

data_2 = pd.read_csv('brffs_diabetes.csv')
```

In [ ]:
```
# Understanding our dataset

data_2.head(), data_2.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 253680 entries, 0 to 253679
Data columns (total 22 columns):
 #   Column                Non-Null Count   Dtype  
---  ------                --------------   -----  
 0   Diabetes_012          253680 non-null  float64
 1   HighBP                253680 non-null  float64
 2   HighChol              253680 non-null  float64
 3   CholCheck             253680 non-null  float64
 4   BMI                   253680 non-null  float64
 5   Smoker                253680 non-null  float64
 6   Stroke                253680 non-null  float64
 7   HeartDiseaseorAttack  253680 non-null  float64
 8   PhysActivity          253680 non-null  float64
 9   Fruits                253680 non-null  float64
 10  Veggies               253680 non-null  float64
 11  HvyAlcoholConsump     253680 non-null  float64
 12  AnyHealthcare         253680 non-null  float64
 13  NoDocbcCost           253680 non-null  float64
 14  GenHlth               253680 non-null  float64
 15  MentHlth              253680 non-null  float64
 16  PhysHlth              253680 non-null  float64
 17  DiffWalk              253680 non-null  float64
 18  Sex                   253680 non-null  float64
 19  Age                   253680 non-null  float64
 20  Education             253680 non-null  float64
 21  Income                253680 non-null  float64
```

```
dtypes: float64(22)
memory usage: 42.6 MB
```

Out[ ]:
```
(   Diabetes_012  HighBP  HighChol  CholCheck   BMI  Smoker  Stroke  \
 0           0.0     1.0       1.0        1.0  40.0     1.0     0.0
 1           0.0     0.0       0.0        0.0  25.0     1.0     0.0
 2           0.0     1.0       1.0        1.0  28.0     0.0     0.0
 3           0.0     1.0       0.0        1.0  27.0     0.0     0.0
 4           0.0     1.0       1.0        1.0  24.0     0.0     0.0

    HeartDiseaseorAttack  PhysActivity  Fruits  ...  AnyHealthcare  \
 0                   0.0           0.0     0.0  ...            1.0
 1                   0.0           1.0     0.0  ...            0.0
 2                   0.0           0.0     1.0  ...            1.0
 3                   0.0           1.0     1.0  ...            1.0
 4                   0.0           1.0     1.0  ...            1.0

    NoDocbcCost  GenHlth  MentHlth  PhysHlth  DiffWalk  Sex   Age  Education  \
 0          0.0      5.0      18.0      15.0       1.0  0.0   9.0        4.0
 1          1.0      3.0       0.0       0.0       0.0  0.0   7.0        6.0
 2          1.0      5.0      30.0      30.0       1.0  0.0   9.0        4.0
 3          0.0      2.0       0.0       0.0       0.0  0.0  11.0        3.0
 4          0.0      2.0       3.0       0.0       0.0  0.0  11.0        5.0

    Income
 0     3.0
 1     1.0
 2     8.0
 3     6.0
 4     4.0

 [5 rows x 22 columns],
 None)
```

In [ ]:
```
data_2.describe()
```

Out[ ]:

	Diabetes_012	HighBP	HighChol	CholCheck	BMI	Smoker	Stroke	HeartDiseaseorAttack	PhysActivity
count	253680.000000	253680.000000	253680.000000	253680.000000	253680.000000	253680.000000	253680.000000	253680.000000	253680.000000
mean	0.296921	0.429001	0.424121	0.962670	28.382364	0.443169	0.040571	0.094186	0.756544
std	0.698160	0.494934	0.494210	0.189571	6.608694	0.496761	0.197294	0.292087	0.429169
min	0.000000	0.000000	0.000000	0.000000	12.000000	0.000000	0.000000	0.000000	0.000000

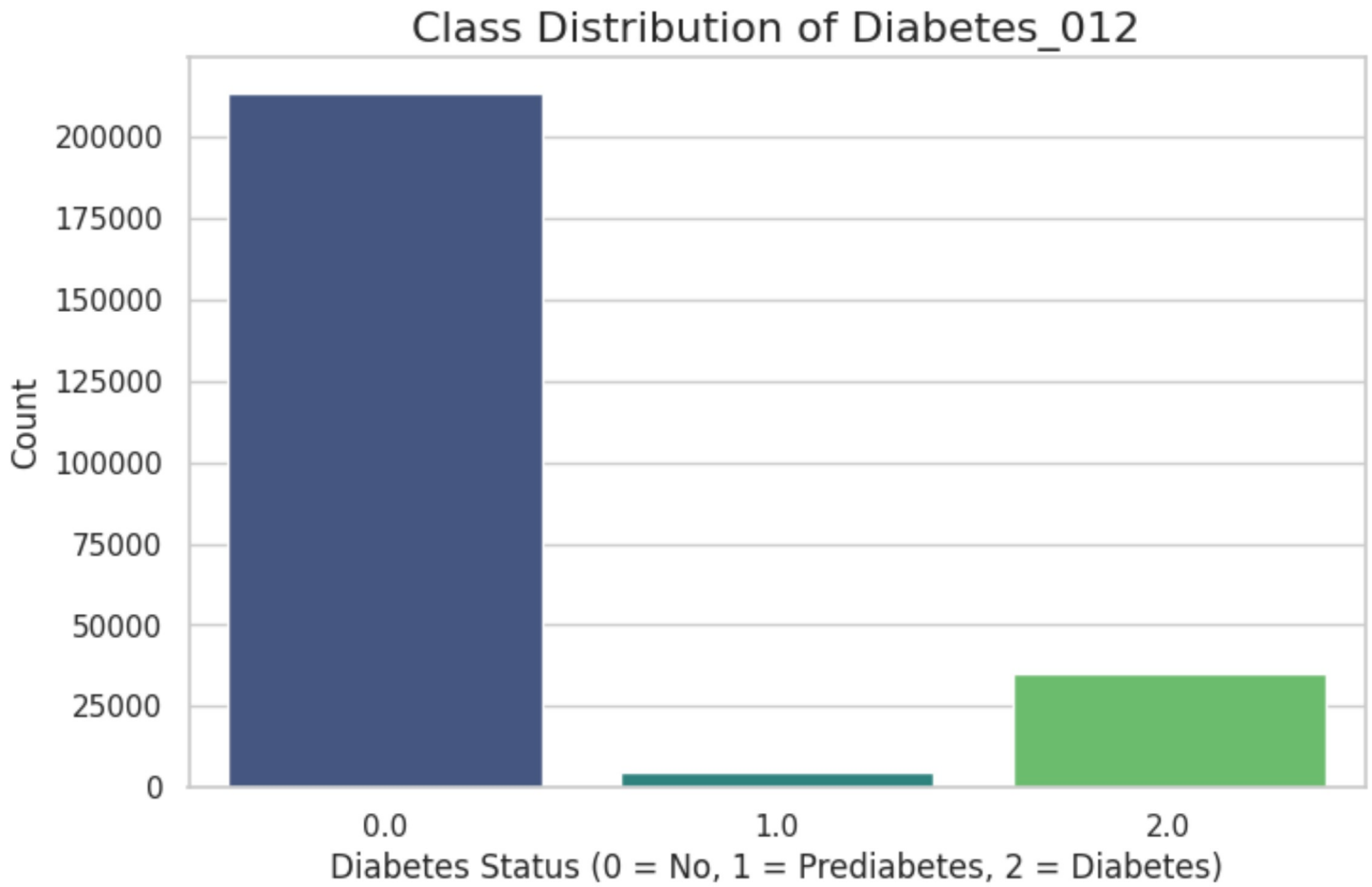
	Diabetes_012	HighBP	HighChol	CholCheck	BMI	Smoker	Stroke	HeartDiseaseorAttack	PhysActivity
25%	0.000000	0.000000	0.000000	1.000000	24.000000	0.000000	0.000000	0.000000	1.000000
50%	0.000000	0.000000	0.000000	1.000000	27.000000	0.000000	0.000000	0.000000	1.000000
75%	0.000000	1.000000	1.000000	1.000000	31.000000	1.000000	0.000000	0.000000	1.000000
max	2.000000	1.000000	1.000000	1.000000	98.000000	1.000000	1.000000	1.000000	1.000000

8 rows × 22 columns

In [ ]:

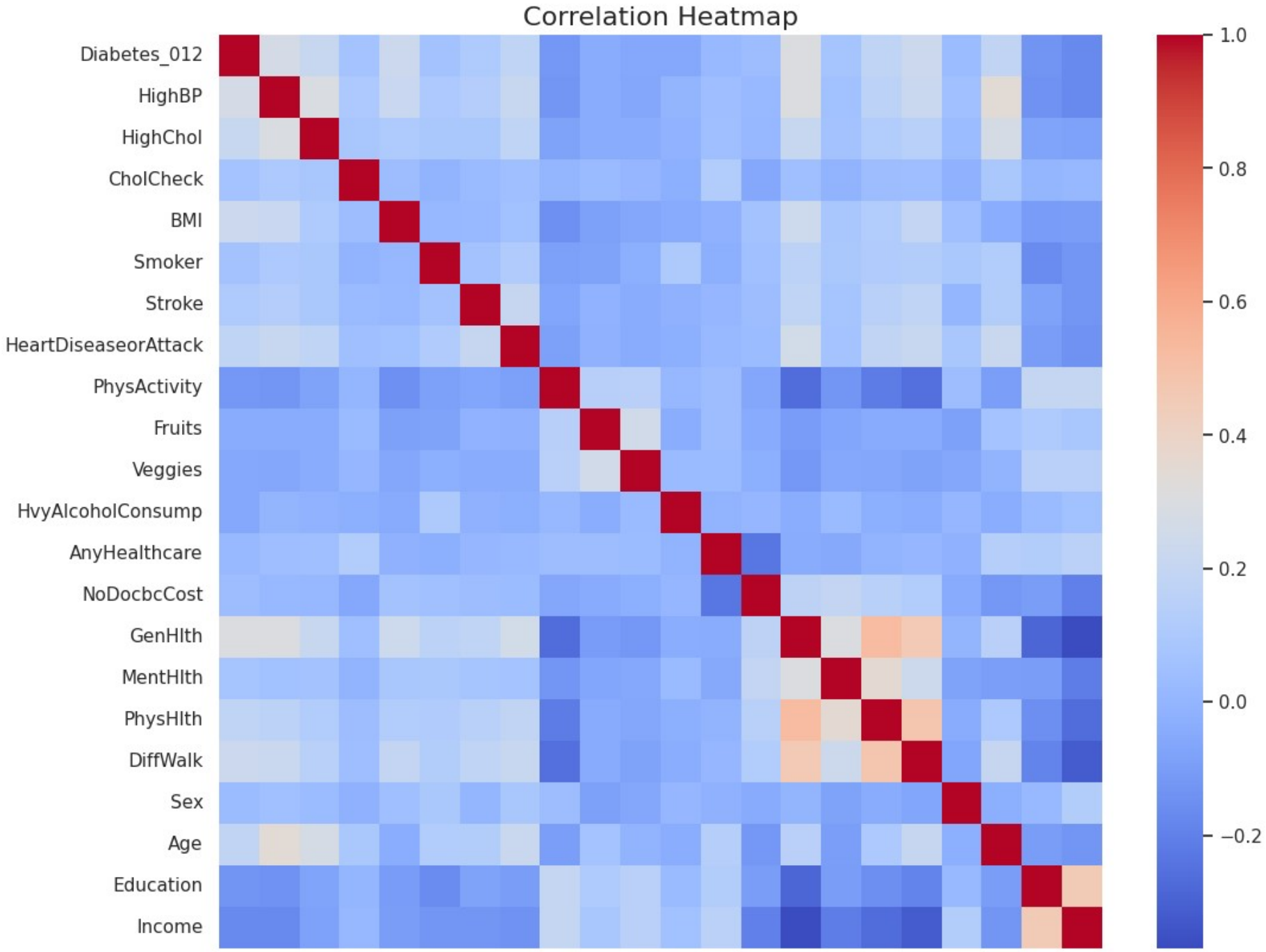
```
# Analyzing our target column: Diabetes_012 - what is the class distribution? is it balanced?

sns.set_theme(style="whitegrid")

class_counts = data_2['Diabetes_012'].value_counts()
print(class_counts)
plt.figure(figsize=(8, 5))
sns.barplot(x=class_counts.index, y=class_counts.values, palette="viridis")
plt.title("Class Distribution of Diabetes_012", fontsize=16)
plt.xlabel("Diabetes Status (0 = No, 1 = Prediabetes, 2 = Diabetes)", fontsize=12)
plt.ylabel("Count", fontsize=12)
plt.show()
```

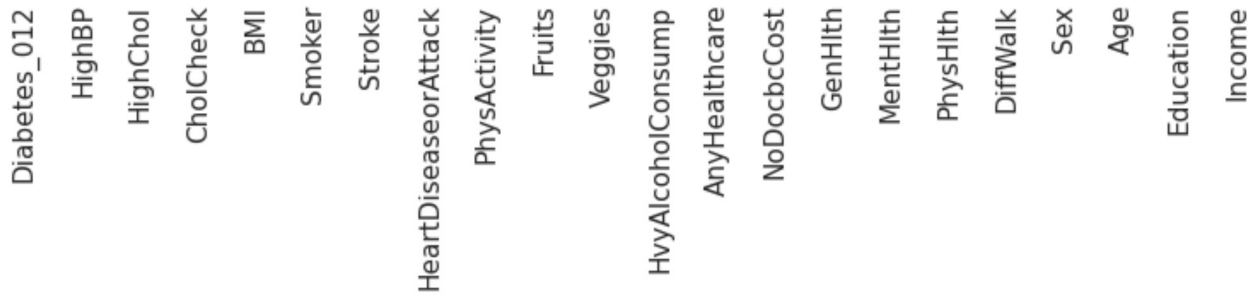
```
Diabetes_012
0.0    213703
2.0     35346
1.0      4631
Name: count, dtype: int64
<ipython-input-10-ff549c419562>:8: FutureWarning:

Passing `palette` without assigning `hue` is deprecated and will be removed in v0.14.0. Assign the `x` variable to `hue` and set
`legend=False` for the same effect.

  sns.barplot(x=class_counts.index, y=class_counts.values, palette="viridis")
```



In [ ]:

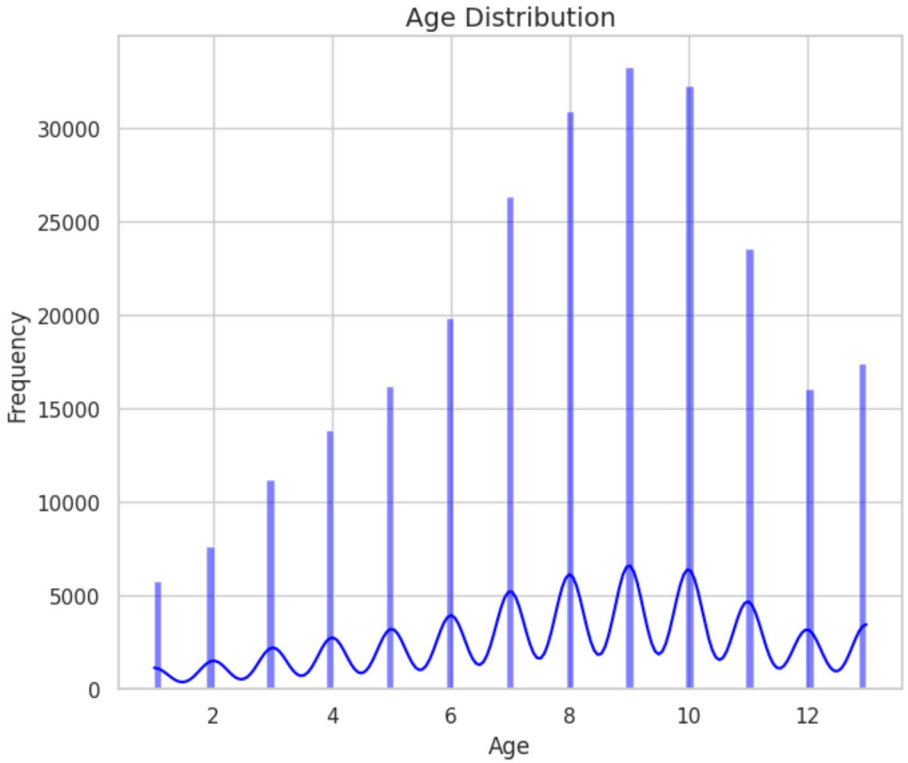
```
# Correlation heatmap to visualize correlations between continuous variables like BMI, blood pressure, cholesterol, etc

plt.figure(figsize=(12, 10))
correlation_matrix = data_2.corr()
sns.heatmap(correlation_matrix, cmap="coolwarm", annot=False, fmt=".2f")
plt.title("Correlation Heatmap", fontsize=16)
plt.show()
```

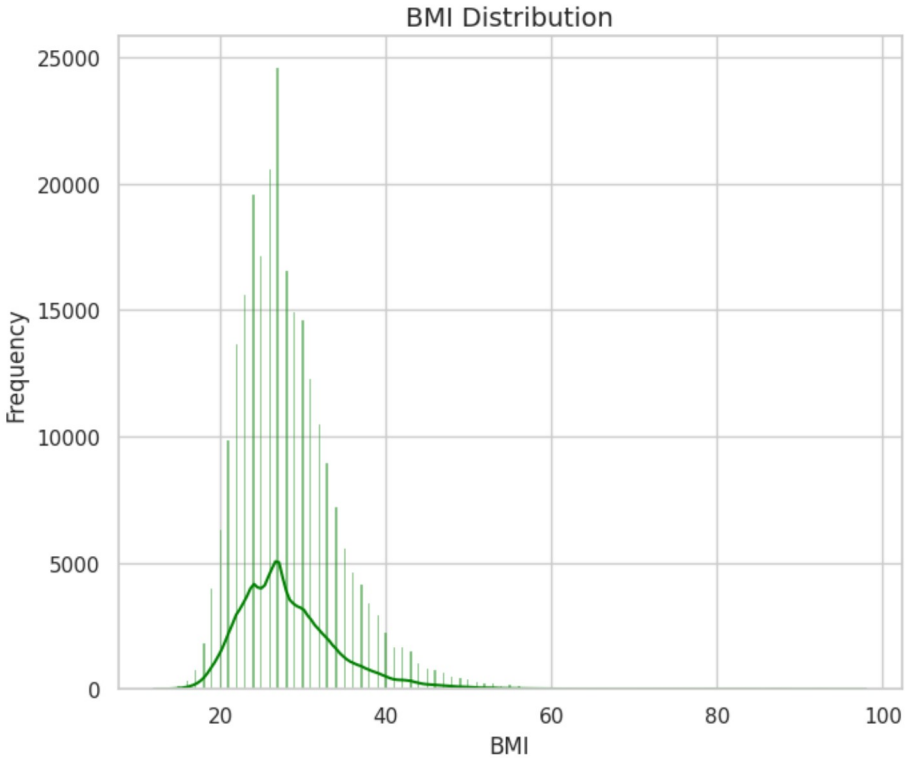

Correlation Heatmap
Diabetes_012
HighBP
HighChol
CholCheck
BMI
Smoker
Stroke
HeartDiseaseorAttack
PhysActivity
Fruits
Veggies
HvyAlcoholConsump
AnyHealthcare
NoDocbcCost
GenHlth
MentHlth
PhysHlth
DiffWalk
Sex
Age
Education
Income
1.0
0.8
0.6
0.4
0.2
0.0
−0.2



In [ ]:

```
# Histograms for Age and BMI: two common variables to understand trends across the patient population

fig, axes = plt.subplots(1, 2, figsize=(14, 6))
sns.histplot(data_2['Age'], kde=True, ax=axes[0], color="blue")
axes[0].set_title("Age Distribution", fontsize=14)
axes[0].set_xlabel("Age")
axes[0].set_ylabel("Frequency")

sns.histplot(data_2['BMI'], kde=True, ax=axes[1], color="green")
axes[1].set_title("BMI Distribution", fontsize=14)
axes[1].set_xlabel("BMI")
axes[1].set_ylabel("Frequency")

plt.tight_layout()
plt.show()
```

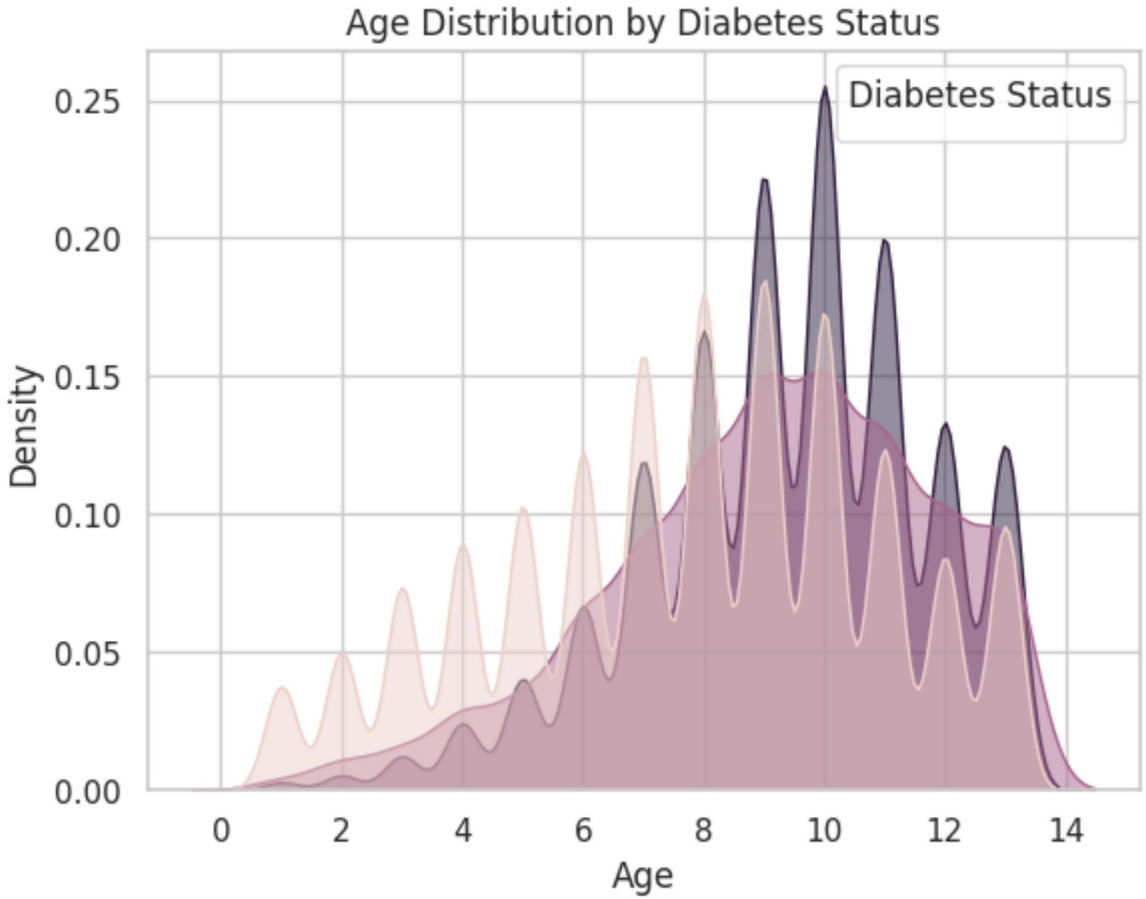




In [ ]:

```
# Kernel Density Estimate (KDE) plots for seeing age across Diabetes Status

#for age:

sns.kdeplot(data=data_2, x="Age", hue="Diabetes_012", fill=True, common_norm=False, alpha=0.5)
plt.title("Age Distribution by Diabetes Status")
plt.xlabel("Age")
plt.ylabel("Density")
plt.legend(title="Diabetes Status")
plt.show()
```

```
WARNING:matplotlib.legend:No artists with labels found to put in legend.  Note that artists whose label start with an undersc
are ignored when legend() is called with no argument.
```

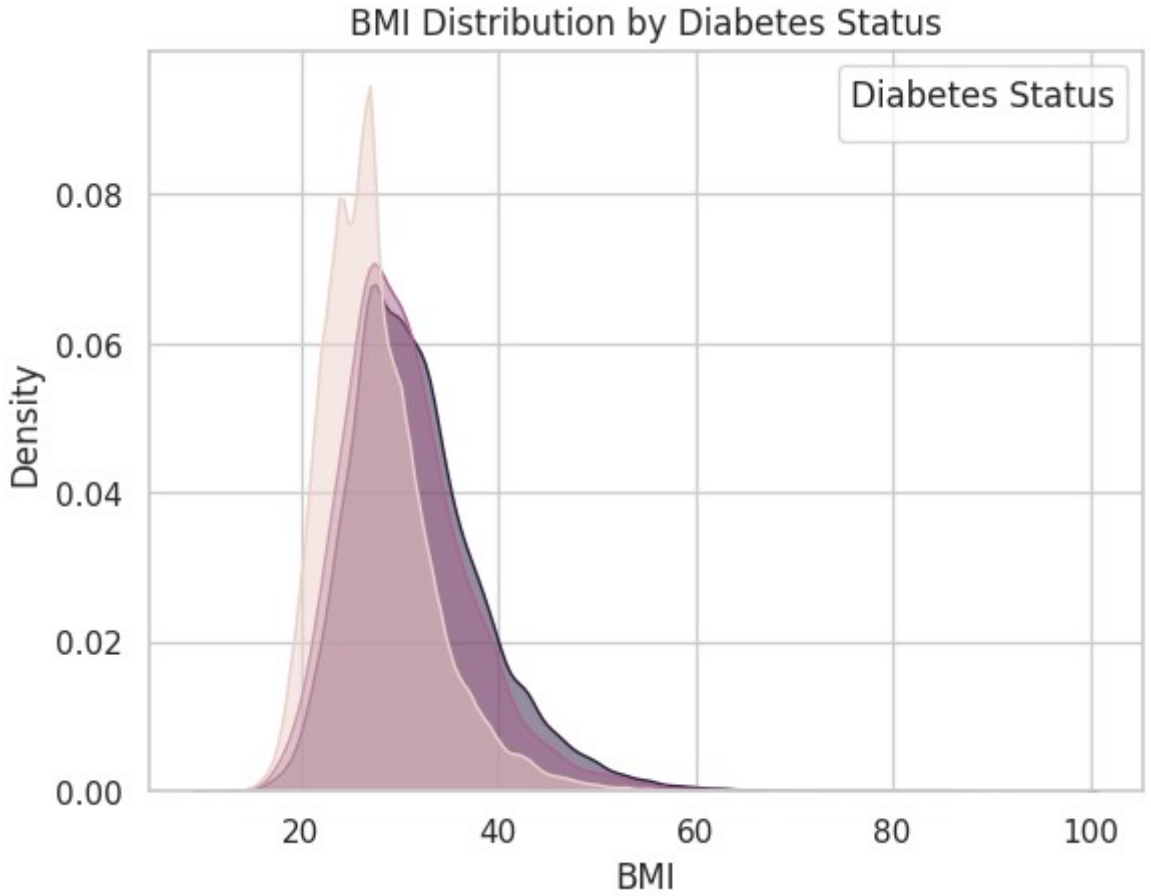


In [ ]:

```
# Kernel Density Estimate (KDE) plots for seeing BMI across Diabetes Status

#for BMI:

sns.kdeplot(data=data_2, x="BMI", hue="Diabetes_012", fill=True, common_norm=False, alpha=0.5)
plt.title("BMI Distribution by Diabetes Status")
plt.xlabel("BMI")
plt.ylabel("Density")
plt.legend(title="Diabetes Status")
plt.show()
```

```
WARNING:matplotlib.legend:No artists with labels found to put in legend.  Note that artists whose label start with an undersc
```

```
are ignored when legend() is called with no argument.
```



Correlation Heatmap

- The heatmap shows the correlation coefficients between various variables in the dataset, ranging from -1 (strong negative correlation) to +1 (strong positive correlation).
- Variables such as Diabetes_012, HighBP, BMI, Age, and Income are included.

Key Observations:

A: Strong Correlations:

- Diabetes_012 has moderate positive correlations with variables like HighBP (high blood pressure) and BMI, suggesting that these factors are associated with higher diabetes prevalence.
- Age correlates strongly with DiffWalk (difficulty walking), indicating mobility issues are more common in older individuals.

B: Negative Correlations:

- Income shows a weak negative correlation with variables like HighBP and BMI, implying higher-income individuals might be less prone to these conditions.
- Physical activity (PhysActivity) has negative correlations with BMI and HighBP, suggesting more active individuals have healthier profiles.

C: Clusters of Relationships:

- Health-related behaviors (like Fruits and Veggies consumption, PhysActivity) cluster together, indicating similar patterns among these variables.
- Mental health (e.g., MentHlth) has weak correlations with physical conditions but aligns with variables like GenHlth.

Implications:

- Understanding these relationships can help target interventions. For example, promoting PhysActivity may address issues related to both BMI and HighBP, which are associated with diabetes prevalence.

Age Distribution

Description:

- The distribution is visualized in bins (age groups) represented by numerical categories.
- The periodic spikes suggest that age data is grouped into distinct categories (possibly representing decades or cohorts).

Key Observations:

Spike Pattern:

- The graph indicates a higher representation in specific age groups, possibly reflecting larger population cohorts or survey design. For example, spikes may occur at ages like 30, 40, 50, etc. Younger and older categories have fewer individuals, likely due to fewer respondents or smaller

age groups.

Evenly Distributed Base:

- Between the spikes, the frequency does not drop to zero, indicating representation across all age ranges, though uneven. Implications: This distribution should be considered when analyzing age-dependent variables. If the dataset is biased toward certain age groups, the results could be skewed.