```
!pip install pandas matplotlib seaborn markdown2 weasyprint
```

```
Requirement already satisfied: pandas in /usr/local/lib/python3.11/dist-packages (2.2.2)
Requirement already satisfied: matplotlib in /usr/local/lib/python3.11/dist-packages (3.10.0)
Requirement already satisfied: seaborn in /usr/local/lib/python3.11/dist-packages (0.13.2)
Collecting markdown2
  Downloading markdown2-2.5.3-py3-none-any.whl.metadata (2.1 kB)
Collecting weasyprint
  Downloading weasyprint-64.1-py3-none-any.whl.metadata (3.7 kB)
Requirement already satisfied: numpy>=1.23.2 in /usr/local/lib/python3.11/dist-packages (from pandas) (1.26.4)
Requirement already satisfied: python-dateutil>=2.8.2 in /usr/local/lib/python3.11/dist-packages (from pandas) (2.8.2)
Requirement already satisfied: pytz>=2020.1 in /usr/local/lib/python3.11/dist-packages (from pandas) (2025.1)
Requirement already satisfied: tzdata>=2022.7 in /usr/local/lib/python3.11/dist-packages (from pandas) (2025.1)
Requirement already satisfied: contourpy>=1.0.1 in /usr/local/lib/python3.11/dist-packages (from matplotlib) (1.3.1)
Requirement already satisfied: cycler>=0.10 in /usr/local/lib/python3.11/dist-packages (from matplotlib) (0.12.1)
Requirement already satisfied: fonttools>=4.22.0 in /usr/local/lib/python3.11/dist-packages (from matplotlib) (4.56.0)
Requirement already satisfied: kiwisolver>=1.3.1 in /usr/local/lib/python3.11/dist-packages (from matplotlib) (1.4.8)
Requirement already satisfied: packaging>=20.0 in /usr/local/lib/python3.11/dist-packages (from matplotlib) (24.2)
Requirement already satisfied: pillow>=8 in /usr/local/lib/python3.11/dist-packages (from matplotlib) (11.1.0)
Requirement already satisfied: pyparsing>=2.3.1 in /usr/local/lib/python3.11/dist-packages (from matplotlib) (3.2.1)
Collecting pydyf>=0.11.0 (from weasyprint)
  Downloading pydyf-0.11.0-py3-none-any.whl.metadata (2.5 kB)
Requirement already satisfied: cffi>=0.6 in /usr/local/lib/python3.11/dist-packages (from weasyprint) (1.17.1)
Collecting tinyhtml5>=2.0.0b1 (from weasyprint)
  Downloading tinyhtml5-2.0.0-py3-none-any.whl.metadata (2.9 kB)
Requirement already satisfied: tinycss2>=1.4.0 in /usr/local/lib/python3.11/dist-packages (from weasyprint) (1.4.0)
Collecting cssselect2>=0.1 (from weasyprint)
  Downloading cssselect2-0.7.0-py3-none-any.whl.metadata (2.9 kB)
Collecting Pyphen>=0.9.1 (from weasyprint)
  Downloading pyphen-0.17.2-py3-none-any.whl.metadata (3.2 kB)
Requirement already satisfied: pycparser in /usr/local/lib/python3.11/dist-packages (from cffi>=0.6->weasyprint) (2.22)
Requirement already satisfied: webencodings in /usr/local/lib/python3.11/dist-packages (from cssselect2>=0.1->weasyprint) (0.5.1)
Collecting brotli>=1.0.1 (from fonttools[woff]>=4.0.0->weasyprint)
  Downloading Brotli-1.1.0-cp311-cp311-manylinux_2_17_x86_64.manylinux2014_x86_64.whl.metadata (5.5 kB)
Collecting zopfli>=0.1.4 (from fonttools[woff]>=4.0.0->weasyprint)
  Downloading zopfli-0.2.3.post1-cp311-cp311-manylinux_2_17_x86_64.manylinux2014_x86_64.whl.metadata (2.9 kB)
Requirement already satisfied: six>=1.5 in /usr/local/lib/python3.11/dist-packages (from python-dateutil>=2.8.2->pandas) (1.17.0)
Downloading markdown2-2.5.3-py3-none-any.whl (48 kB)
   ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 48.5/48.5 kB 1.2 MB/s eta 0:00:00
Downloading weasyprint-64.1-py3-none-any.whl (302 kB)
   ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 302.0/302.0 kB 4.2 MB/s eta 0:00:00
Downloading cssselect2-0.7.0-py3-none-any.whl (15 kB)
Downloading pydyf-0.11.0-py3-none-any.whl (8.1 kB)
Downloading pyphen-0.17.2-py3-none-any.whl (2.1 MB)
   ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 2.1/2.1 MB 19.7 MB/s eta 0:00:00
Downloading tinyhtml5-2.0.0-py3-none-any.whl (39 kB)
Downloading Brotli-1.1.0-cp311-cp311-manylinux_2_17_x86_64.manylinux2014_x86_64.whl (2.9 MB)
   ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 2.9/2.9 MB 25.6 MB/s eta 0:00:00
Downloading zopfli-0.2.3.post1-cp311-cp311-manylinux_2_17_x86_64.manylinux2014_x86_64.whl (850 kB)
   ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 850.6/850.6 kB 15.0 MB/s eta 0:00:00
Installing collected packages: brotli, zopfli, tinyhtml5, Pyphen, pydyf, markdown2, cssselect2, weasyprint
Successfully installed Pyphen-0.17.2 brotli-1.1.0 cssselect2-0.7.0 markdown2-2.5.3 pydyf-0.11.0 tinyhtml5-2.0.0 weasyprint-64.1 zopfli-0
```

```
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
import markdown2
import weasyprint
```

```
from google.colab import files
uploaded = files.upload()
```

Choose Files No file chosen Upload widget is only available when the cell has been executed in the current browser session. Please rerun this cell to enable.

```
df = pd.read_csv("diabetes_012_health_indicators_BRFSS2015.csv")
```

```
# Display basic info
print("Dataset Information:")
print(df.info())
```

```
Dataset Information:
<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
None
```

```
# Show first few rows
print("\nFirst 5 Rows:")
print(df.head())
```

```
First 5 Rows:
   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]
```

```
# Show column names
print("\nColumn Names:")
print(df.columns)
```

```
Column Names:
Index(['Diabetes_012', 'HighBP', 'HighChol', 'CholCheck', 'BMI', 'Smoker',
       'Stroke', 'HeartDiseaseorAttack', 'PhysActivity', 'Fruits', 'Veggies',
       'HvyAlcoholConsump', 'AnyHealthcare', 'NoDocbcCost', 'GenHlth',
       'MentHlth', 'PhysHlth', 'DiffWalk', 'Sex', 'Age', 'Education',
       'Income'],
      dtype='object')
```

```
# Check for missing values
print("\nMissing Values:")
print(df.isnull().sum())
```

```

Missing Values:
Diabetes_012            0
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
```

```
# Summary statistics
print("\nSummary Statistics:")
print(df.describe())
```

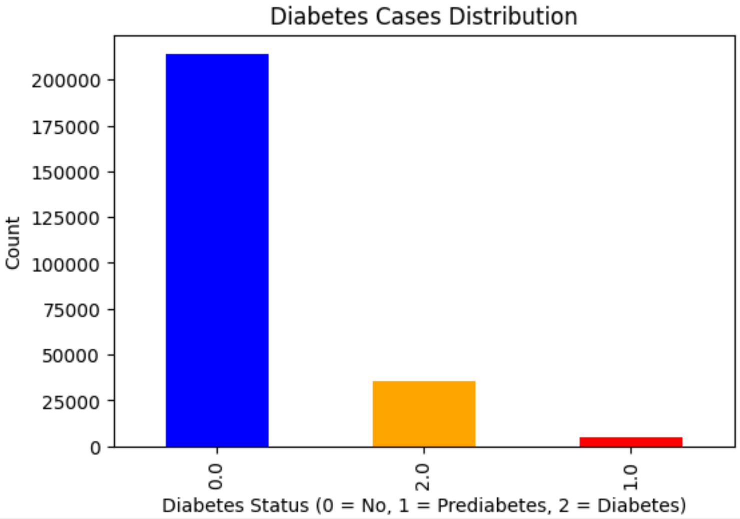
```

Summary Statistics:
        Diabetes_012         HighBP       HighChol      CholCheck  \
count  253680.000000  253680.000000  253680.000000  253680.000000   
mean        0.296921       0.429001       0.424121       0.962670   
std         0.698160       0.494934       0.494210       0.189571   
min         0.000000       0.000000       0.000000       0.000000   
25%         0.000000       0.000000       0.000000       1.000000   
50%         0.000000       0.000000       0.000000       1.000000   
75%         0.000000       1.000000       1.000000       1.000000   
max         2.000000       1.000000       1.000000       1.000000   

                 BMI         Smoker         Stroke  HeartDiseaseorAttack  \
count  253680.000000  253680.000000  253680.000000         253680.000000   
mean       28.382364       0.443169       0.040571              0.094186   
std         6.608694       0.496761       0.197294              0.292087   
min        12.000000       0.000000       0.000000              0.000000   
25%        24.000000       0.000000       0.000000              0.000000   
50%        27.000000       0.000000       0.000000              0.000000   
75%        31.000000       1.000000       0.000000              0.000000   
max        98.000000       1.000000       1.000000              1.000000   

        PhysActivity         Fruits  ...  AnyHealthcare    NoDocbcCost  \
count  253680.000000  253680.000000  ...  253680.000000  253680.000000   
mean        0.756544       0.634256  ...       0.951053       0.084177   
std         0.429169       0.481639  ...       0.215759       0.277654   
min         0.000000       0.000000  ...       0.000000       0.000000   
25%         1.000000       0.000000  ...       1.000000       0.000000   
50%         1.000000       1.000000  ...       1.000000       0.000000   
75%         1.000000       1.000000  ...       1.000000       0.000000   
max         1.000000       1.000000  ...       1.000000       1.000000   

             GenHlth       MentHlth       PhysHlth       DiffWalk  \
count  253680.000000  253680.000000  253680.000000  253680.000000   
mean        2.511392       3.184772       4.242081       0.168224   
std         1.068477       7.412847       8.717951       0.374066   
min         1.000000       0.000000       0.000000       0.000000   
25%         2.000000       0.000000       0.000000       0.000000   
50%         2.000000       0.000000       0.000000       0.000000   
75%         3.000000       2.000000       3.000000       0.000000   
max         5.000000      30.000000      30.000000       1.000000   

                 Sex            Age      Education         Income  
count  253680.000000  253680.000000  253680.000000  253680.000000  
mean        0.440342       8.032119       5.050434       6.053875  
std         0.496429       3.054220       0.985774       2.071148  
min         0.000000       1.000000       1.000000       1.000000  
25%         0.000000       6.000000       4.000000       5.000000  
```

```
50%         0.000000       8.000000       5.000000       7.000000
75%         1.000000      10.000000       6.000000       8.000000
max         1.000000      13.000000       6.000000       8.000000

[8 rows x 22 columns]
```

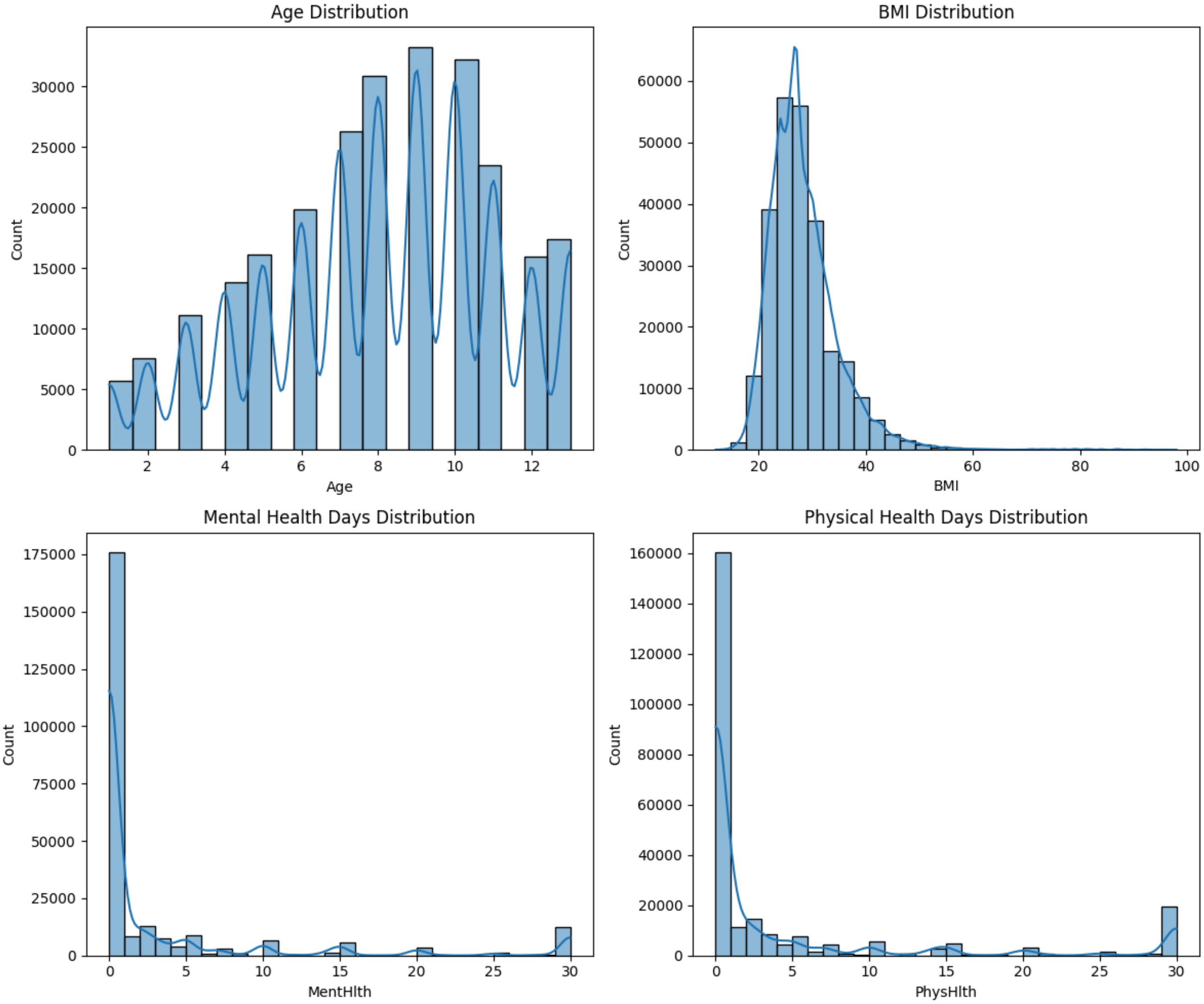
```
df.columns = df.columns.str.strip()  # Remove leading/trailing spaces
```

```
df.fillna(df.median(), inplace=True)  # Replace NaN with median values
```

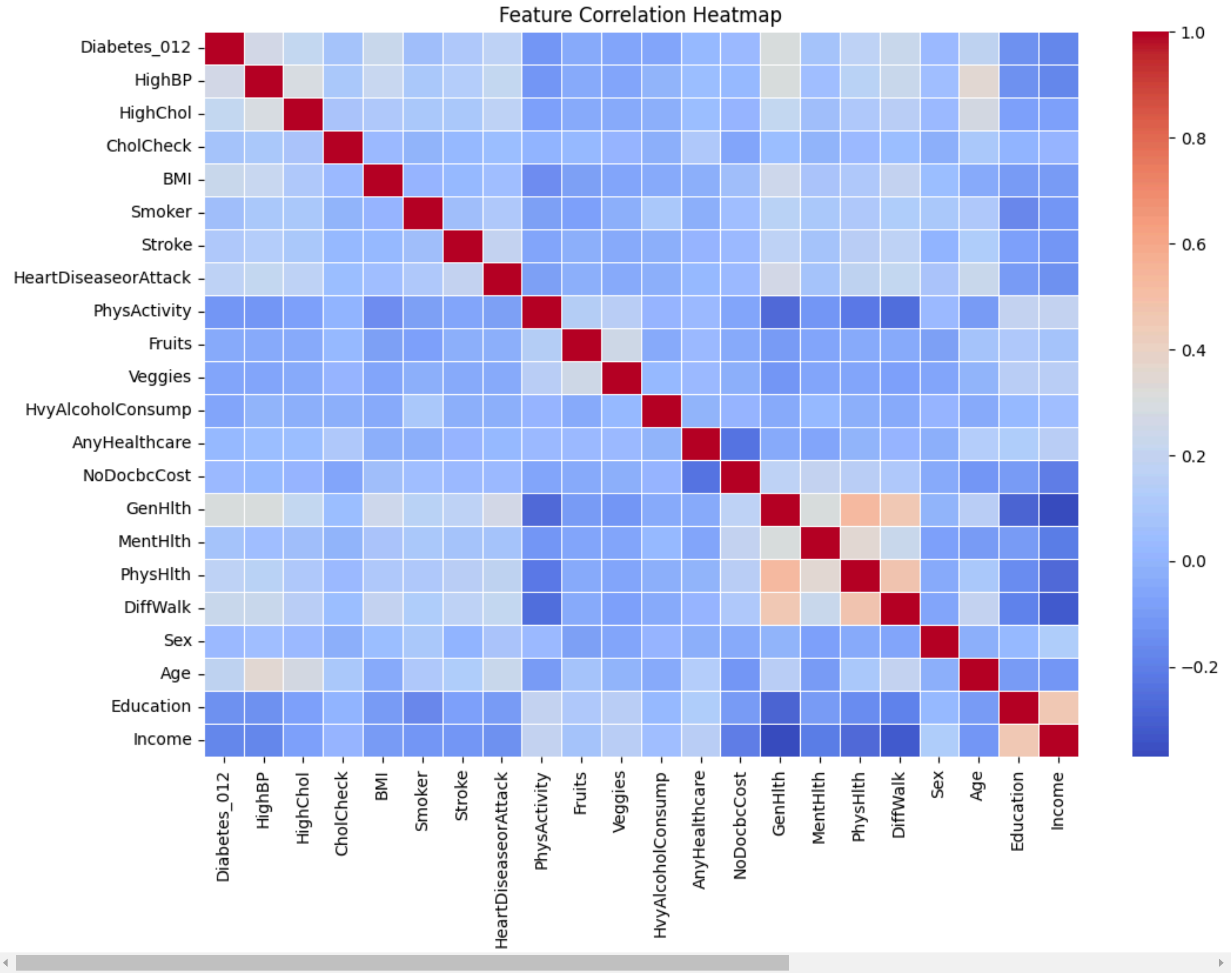
```
import matplotlib.pyplot as plt

plt.figure(figsize=(6,4))
df['Diabetes_012'].value_counts().plot(kind='bar', color=['blue', 'orange', 'red'])
plt.title("Diabetes Cases Distribution")
plt.xlabel("Diabetes Status (0 = No, 1 = Prediabetes, 2 = Diabetes)")
plt.ylabel("Count")
plt.show()
```



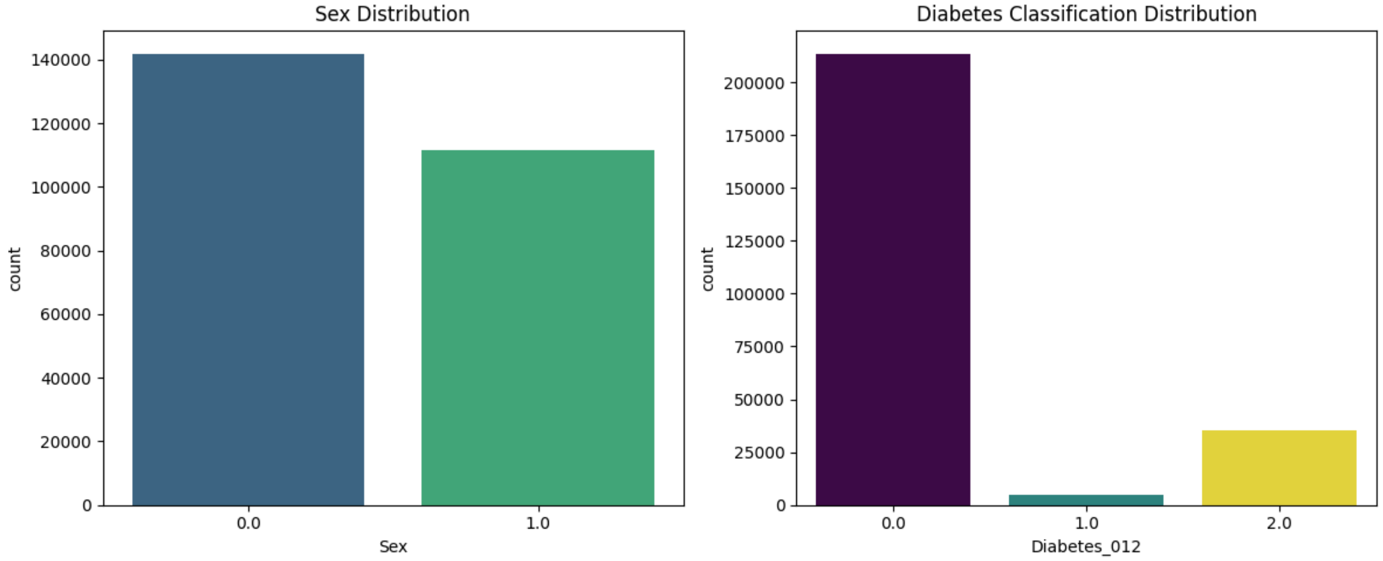
```
# Distribution Plots
fig, axes = plt.subplots(2, 2, figsize=(12, 10))

sns.histplot(df['Age'], bins=20, kde=True, ax=axes[0, 0])
axes[0, 0].set_title('Age Distribution')

sns.histplot(df['BMI'], bins=30, kde=True, ax=axes[0, 1])
axes[0, 1].set_title('BMI Distribution')

sns.histplot(df['MentHlth'], bins=30, kde=True, ax=axes[1, 0])
axes[1, 0].set_title('Mental Health Days Distribution')

sns.histplot(df['PhysHlth'], bins=30, kde=True, ax=axes[1, 1])
axes[1, 1].set_title('Physical Health Days Distribution')

plt.tight_layout()
plt.savefig("feature_distributions.png")
plt.show()
```



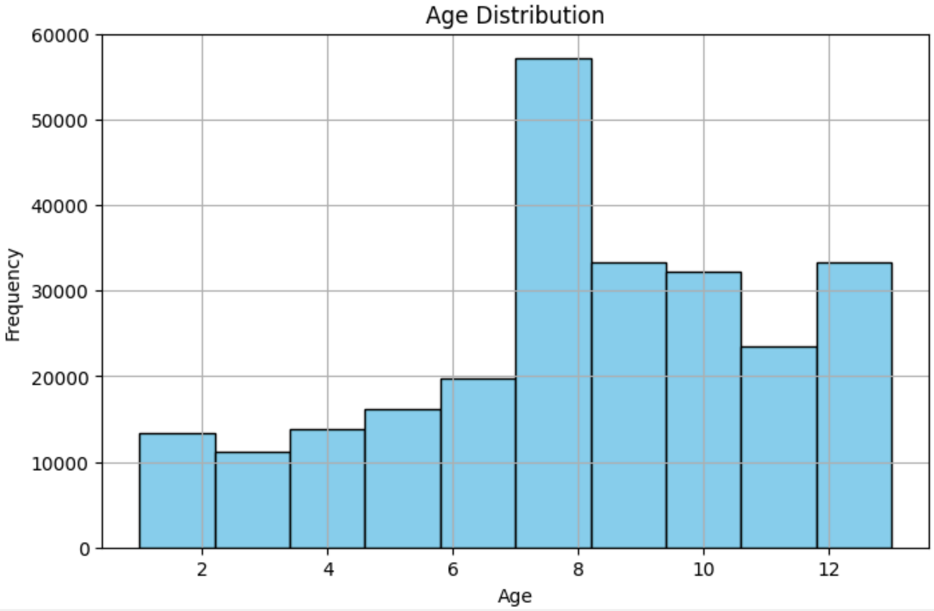
```
# Correlation Heatmap
plt.figure(figsize=(12, 8))
sns.heatmap(df.corr(numeric_only=True), annot=False, cmap='coolwarm', linewidths=0.5)
plt.title("Feature Correlation Heatmap")
plt.savefig("correlation_heatmap.png")
plt.show()
```



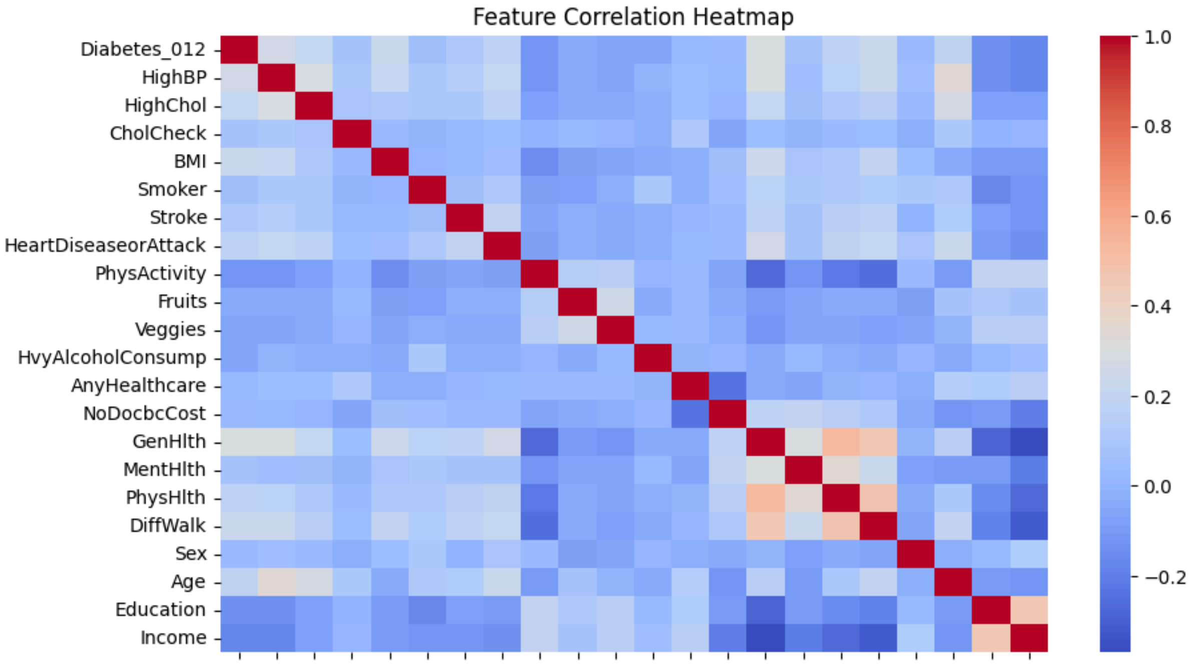
```
# Categorical Feature Distributions
fig, axes = plt.subplots(1, 2, figsize=(12, 5))

# Use hue and dodge=False for proper palette application
sns.countplot(x='Sex', data=df, ax=axes[0], palette="viridis", hue='Sex', dodge=False)
axes[0].set_title("Sex Distribution")
axes[0].legend([],[], frameon=False) # Remove legend

# Use hue and dodge=False for proper palette application
sns.countplot(x='Diabetes_012', data=df, ax=axes[1], palette="viridis", hue='Diabetes_012', dodge=False)
axes[1].set_title("Diabetes Classification Distribution")
axes[1].legend([],[], frameon=False) # Remove legend


plt.tight_layout()
plt.savefig("categorical_distributions.png")
plt.show()
```



```
plt.figure(figsize=(8,5))
df['Age'].hist(bins=10, color='skyblue', edgecolor='black')
plt.title("Age Distribution")
plt.xlabel("Age")
plt.ylabel("Frequency")
plt.show()
```

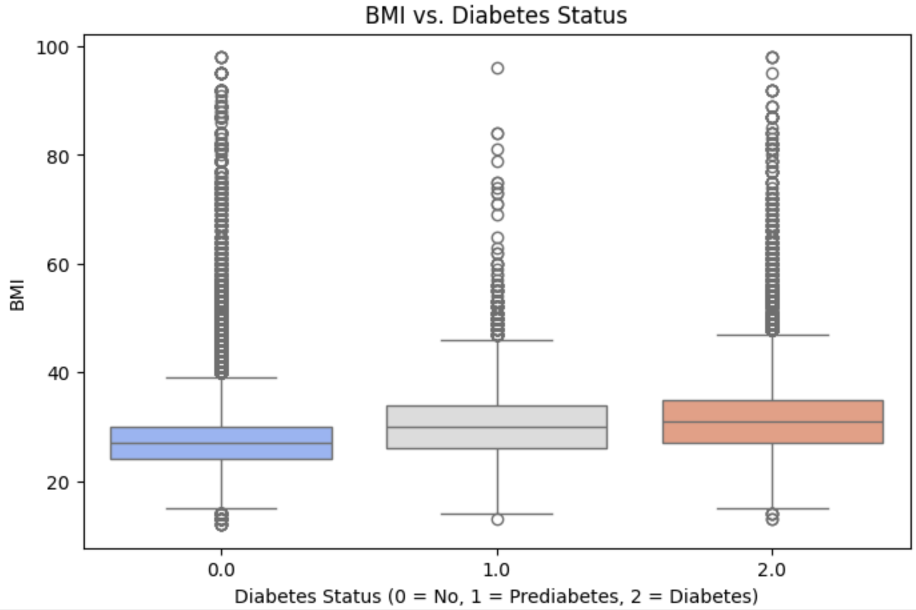


```
plt.figure(figsize=(10,6))
sns.heatmap(df.corr(), cmap="coolwarm", annot=False)
plt.title("Feature Correlation Heatmap")
plt.show()
```



```
plt.figure(figsize=(8,5))
sns.boxplot(x=df["Diabetes_012"], y=df["BMI"], palette="coolwarm")
plt.title("BMI vs. Diabetes Status")
plt.xlabel("Diabetes Status (0 = No, 1 = Prediabetes, 2 = Diabetes)")
plt.ylabel("BMI")
plt.show()
```

```
<ipython-input-56-da104d666e70>:2: 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

  sns.boxplot(x=df["Diabetes_012"], y=df["BMI"], palette="coolwarm")
```


BMI vs. Diabetes Status
BMI
Diabetes Status (0 = No, 1 = Prediabetes, 2 = Diabetes)