Diabetes EDA

May 1, 2025

[1]:
```python
# Setup
import warnings
import pandas as pd
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt

warnings.filterwarnings('ignore', category=FutureWarning)

sns.set_theme(style="whitegrid")
df = pd.read_csv('diabetes_health_indicators.csv')
```

[2]:
```python
# Check for any missing values (NA/NaN)
missing_counts = df.isnull().sum()
total_rows = len(df)
missing_percentages = (missing_counts / total_rows) * 100

print(f"\nMissing values: {missing_counts}")
print(f"\nMissing percentages: {missing_percentages}")
```

```
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

Missing percentages: Diabetes_012            0.0
HighBP                  0.0
HighChol                0.0
CholCheck               0.0
BMI                     0.0
Smoker                  0.0
Stroke                  0.0
HeartDiseaseorAttack    0.0
PhysActivity            0.0
Fruits                  0.0
Veggies                 0.0
HvyAlcoholConsump       0.0
AnyHealthcare           0.0
NoDocbcCost             0.0
GenHlth                 0.0
MentHlth                0.0
PhysHlth                0.0
DiffWalk                0.0
Sex                     0.0
Age                     0.0
Education               0.0
Income                  0.0
dtype: float64
```

[3]:
```
# Convert all floats to ints
for name, values in df.items():
    if name in df.columns:
        df[name] = pd.to_numeric(df[name], errors='coerce')
        df[name] = df[name].astype('Int64')

df.head()
```

[3]:
```
   Diabetes_012  HighBP  HighChol  CholCheck  BMI  Smoker  Stroke  \
0             0       1         1          1   40       1       0
1             0       0         0          0   25       1       0
2             0       1         1          1   28       0       0
3             0       1         0          1   27       0       0
4             0       1         1          1   24       0       0

   HeartDiseaseorAttack  PhysActivity  Fruits  …  AnyHealthcare  \
```

```
0                      0             0        0  ...               1
1                      0             1        0  ...               0
2                      0             0        1  ...               1
3                      0             1        1  ...               1
4                      0             1        1  ...               1

   NoDocbcCost  GenHlth  MentHlth  PhysHlth  DiffWalk  Sex  Age  Education  \
0            0        5        18        15         1    0    9          4
1            1        3         0         0         0    0    7          6
2            1        5        30        30         1    0    9          4
3            0        2         0         0         0    0   11          3
4            0        2         3         0         0    0   11          5

   Income
0       3
1       1
2       8
3       6
4       4

[5 rows x 22 columns]
```

[4]:

```
# Map numerical data to descriptive data
# Create copy for cleaning
df = df.rename(columns={'Diabetes_012': 'Diabetes_Status'})
df_clean = df.copy()

binary_map = {0: 'No', 1: 'Yes', 7: 'Not Sure', 9: 'No Response'}
sex_map = {0: 'Female', 1: 'Male'}

diabetes_map = {
    0: 'No Diabetes',
    1: 'Prediabetes',
    2: 'Diabetes'
}

gen_hlth_map = {
    1: 'Excellent',
    2: 'Very Good',
    3: 'Good',
    4: 'Fair',
    5: 'Poor',
    7: 'Not Sure',
    9: 'No Response'
}

education_map = {
```

```
    1: 'Never attended school',
    2: 'Grades 1-8',
    3: 'Grades 9-11',
    4: 'Grade 12/GED',
    5: 'College 1-3 years',
    6: 'College 4+ years',
    9: 'No Response',
}

income_map = {
    1: '< $10,000',
    2: '$10,000 - $14,999',
    3: '$15,000 - $19,999',
    4: '$20,000 - $24,999',
    5: '$25,000 - $34,999',
    6: '$35,000 - $49,999',
    7: '$50,000 - $74,999',
    8: '>= $75,000',
    77: 'Not Sure',
    99: 'No Response'
}

age_map = {
    1: '18-24', 2: '25-29', 3: '30-34', 4: '35-39', 5: '40-44',
    6: '45-49', 7: '50-54', 8: '55-59', 9: '60-64', 10: '65-69',
    11: '70-74', 12: '75-79', 13: '80+', 14: 'No Response'
}


df_clean['Diabetes_Status'] = df_clean['Diabetes_Status'].map(diabetes_map)

binary_cols = [
    'HighBP', 'HighChol', 'CholCheck', 'Smoker', 'Stroke',
    'HeartDiseaseorAttack', 'PhysActivity', 'Fruits', 'Veggies',
    'HvyAlcoholConsump', 'AnyHealthcare', 'NoDocbcCost', 'DiffWalk'
]

for col in binary_cols:
    if col in df.columns:
        df_clean[col] = df_clean[col].map(binary_map)


df_clean['Sex'] = df_clean['Sex'].map(sex_map)

scale_mappings = {
    'GenHlth': gen_hlth_map,
    'Education': education_map,
```

```
    'Income': income_map,
    'Age': age_map
}

for col, mapping in scale_mappings.items():
    if col in df.columns:
        df_clean[col] = df_clean[col].map(mapping)

df_clean.head()
```

[4]:
```
  Diabetes_Status HighBP HighChol CholCheck  BMI Smoker Stroke  \
0     No Diabetes    Yes      Yes       Yes   40    Yes     No
1     No Diabetes     No       No        No   25    Yes     No
2     No Diabetes    Yes      Yes       Yes   28     No     No
3     No Diabetes    Yes       No       Yes   27     No     No
4     No Diabetes    Yes      Yes       Yes   24     No     No

  HeartDiseaseorAttack PhysActivity Fruits  … AnyHealthcare NoDocbcCost  \
0                   No           No     No  …           Yes          No
1                   No          Yes     No  …            No         Yes
2                   No           No    Yes  …           Yes         Yes
3                   No          Yes    Yes  …           Yes          No
4                   No          Yes    Yes  …           Yes          No

     GenHlth MentHlth PhysHlth  DiffWalk     Sex    Age          Education  \
0       Poor       18       15       Yes  Female  60-64       Grade 12/GED
1       Good        0        0        No  Female  50-54   College 4+ years
2       Poor       30       30       Yes  Female  60-64       Grade 12/GED
3  Very Good        0        0        No  Female  70-74        Grades 9-11
4  Very Good        3        0        No  Female  70-74  College 1-3 years

               Income
0  $15,000 - $19,999
1           < $10,000
2          >= $75,000
3  $35,000 - $49,999
4  $20,000 - $24,999

[5 rows x 22 columns]
```

1 Diabetes Health Indicators Analysis

1.1 Background

Diabetes is a chronic health condition affecting millions of people worldwide. This project analyzes a dataset of health indicators to understand factors associated with diabetes prevalence and risk.

1.2 Problem Definition

This analysis aims to: 1. Identify which health indicators are most strongly associated with diabetes status 2. Examine how demographic factors correlate with diabetes risk 3. Explore relationships between modifiable risk factors and diabetes 4. Suggest potential intervention points for diabetes prevention

[5]:
```python
# Basic dataset statistics
print("Dataset overview:")
print(f"Total records: {len(df_clean)}")
print(f"Features: {df_clean.shape[1]}")

# Distribution of diabetes status
diabetes_counts = df_clean['Diabetes_Status'].value_counts()
print(f"\nDistribution of diabetes status:\n{diabetes_counts}")
print(f"Percentage:\n{round(diabetes_counts / len(df_clean) * 100, 2)}%")

# Analyze key health indicators by diabetes status
print("\n--- Key Health Indicators by Diabetes Status ---")
for column in ['HighBP', 'HighChol', 'BMI', 'GenHlth', 'Age']:
    print(f"\n{column} by Diabetes Status:")
    cross_tab = pd.crosstab(df_clean['Diabetes_Status'], df_clean[column])
    percentage = pd.crosstab(df_clean['Diabetes_Status'], df_clean[column],
                            normalize='index').round(3) * 100
    print(f"Counts:\n{cross_tab}")
    print(f"Percentage:\n{percentage}")

# Analyze demographic factors
print("\n--- Demographic Analysis ---")
for column in ['Sex', 'Age', 'Education', 'Income']:
    print(f"\nDiabetes Status by {column}:")
    demo_cross = pd.crosstab(df_clean[column], df_clean['Diabetes_Status'],
                            normalize='index').round(3) * 100
    print(demo_cross)

# Summary statistics by diabetes status
print("\n--- Summary Statistics by Diabetes Status ---")
numeric_cols = ['BMI', 'PhysHlth', 'MentHlth']
for status in df_clean['Diabetes_Status'].unique():
    subset = df[df_clean['Diabetes_Status'] == status]
    print(f"\nFor {status}:")
    print(subset[numeric_cols].describe().round(2))
```

```
Dataset overview:
Total records: 253680
Features: 22

Distribution of diabetes status:
Diabetes_Status
```

```
No Diabetes    213703
Diabetes        35346
Prediabetes      4631
Name: count, dtype: int64
Percentage:
Diabetes_Status
No Diabetes    84.24
Diabetes       13.93
Prediabetes     1.83
Name: count, dtype: float64%

--- Key Health Indicators by Diabetes Status ---

HighBP by Diabetes Status:
Counts:
HighBP               No    Yes
Diabetes_Status
Diabetes           8742  26604
No Diabetes      134391  79312
Prediabetes        1718   2913
Percentage:
HighBP             No   Yes
Diabetes_Status
Diabetes         24.7  75.3
No Diabetes      62.9  37.1
Prediabetes      37.1  62.9

HighChol by Diabetes Status:
Counts:
HighChol             No    Yes
Diabetes_Status
Diabetes          11660  23686
No Diabetes      132673  81030
Prediabetes        1756   2875
Percentage:
HighChol           No   Yes
Diabetes_Status
Diabetes         33.0  67.0
No Diabetes      62.1  37.9
Prediabetes      37.9  62.1

BMI by Diabetes Status:
Counts:
BMI              12  13  14   15   16   17    18    19    20    21  …  86  \
Diabetes_Status                                                     …
Diabetes          0   2   4   12   20   48    83   135   241   479  …   0
No Diabetes       6  18  36  120  326  719  1705  3795  6039  9301  …   1
Prediabetes       0   1   1    0    2    9    15    38    47    75  …   0
```

```
BMI               87  88  89  90  91  92  95  96  98
Diabetes_Status
Diabetes           9   0   3   0   0   5   1   0   3
No Diabetes       52   2  25   1   1  27  11   0   4
Prediabetes        0   0   0   0   0   0   0   1   0

[3 rows x 84 columns]
Percentage:
BMI                12   13   14   15   16   17   18   19   20   21  …   86  \
Diabetes_Status                                                      …
Diabetes          0.0  0.0  0.0  0.0  0.1  0.1  0.2  0.4  0.7  1.4  …  0.0
No Diabetes       0.0  0.0  0.0  0.1  0.2  0.3  0.8  1.8  2.8  4.4  …  0.0
Prediabetes       0.0  0.0  0.0  0.0  0.0  0.2  0.3  0.8  1.0  1.6  …  0.0

BMI                87   88   89   90   91   92   95   96   98
Diabetes_Status
Diabetes          0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0
No Diabetes       0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0
Prediabetes       0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0

[3 rows x 84 columns]

GenHlth by Diabetes Status:
Counts:
GenHlth           Excellent   Fair   Good  Poor  Very Good
Diabetes_Status
Diabetes               1140   9790  13457  4578       6381
No Diabetes           43846  20755  60461  7152      81489
Prediabetes             313   1025   1728   351       1214
Percentage:
GenHlth           Excellent  Fair  Good  Poor  Very Good
Diabetes_Status
Diabetes                3.2  27.7  38.1  13.0       18.1
No Diabetes            20.5   9.7  28.3   3.3       38.1
Prediabetes             6.8  22.1  37.3   7.6       26.2

Age by Diabetes Status:
Counts:
Age               18-24  25-29  30-34  35-39  40-44  45-49  50-54  55-59  \
Diabetes_Status
Diabetes             78    140    314    626   1051   1742   3088   4263
No Diabetes        5601   7404  10737  13055  14943  17765  22808  26019
Prediabetes          21     54     72    142    163    312    418    550

Age               60-64  65-69  70-74  75-79    80+
Diabetes_Status
Diabetes           5733   6558   5141   3403   3209
```

```
No Diabetes      26809  24939  17790  12132  13701
Prediabetes        702    697    602    445    453
Percentage:
Age              18-24  25-29  30-34  35-39  40-44  45-49  50-54  55-59  \
Diabetes_Status
Diabetes           0.2    0.4    0.9    1.8    3.0    4.9    8.7   12.1
No Diabetes        2.6    3.5    5.0    6.1    7.0    8.3   10.7   12.2
Prediabetes        0.5    1.2    1.6    3.1    3.5    6.7    9.0   11.9

Age              60-64  65-69  70-74  75-79  80+
Diabetes_Status
Diabetes          16.2   18.6   14.5    9.6  9.1
No Diabetes       12.5   11.7    8.3    5.7  6.4
Prediabetes       15.2   15.1   13.0    9.6  9.8

--- Demographic Analysis ---

Diabetes Status by Sex:
Diabetes_Status  Diabetes  No Diabetes  Prediabetes
Sex
Female               13.0         85.2          1.8
Male                 15.2         83.0          1.8

Diabetes Status by Age:
Diabetes_Status  Diabetes  No Diabetes  Prediabetes
Age
18-24                 1.4         98.3          0.4
25-29                 1.8         97.4          0.7
30-34                 2.8         96.5          0.6
35-39                 4.5         94.4          1.0
40-44                 6.5         92.5          1.0
45-49                 8.8         89.6          1.6
50-54                11.7         86.7          1.6
55-59                13.8         84.4          1.8
60-64                17.2         80.6          2.1
65-69                20.4         77.5          2.2
70-74                21.8         75.6          2.6
75-79                21.3         75.9          2.8
80+                  18.5         78.9          2.6

Diabetes Status by Education:
Diabetes_Status        Diabetes  No Diabetes  Prediabetes
Education
College 1-3 years          14.8         83.3          1.9
College 4+ years            9.7         88.9          1.4
Grade 12/GED               17.6         80.2          2.2
Grades 1-8                 29.3         66.8          4.0
Grades 9-11                24.2         72.5          3.3
```

Never attended school	27.0	71.8	1.1

Diabetes Status by Income:

Diabetes_Status	Diabetes	No Diabetes	Prediabetes
Income			
$10,000 - $14,999	26.2	70.8	3.0
$15,000 - $19,999	22.3	75.1	2.6
$20,000 - $24,999	20.1	77.6	2.3
$25,000 - $34,999	17.4	80.3	2.3
$35,000 - $49,999	14.5	83.4	2.1
$50,000 - $74,999	12.2	86.1	1.7
< $10,000	24.3	72.5	3.2
>= $75,000	8.0	90.9	1.1

--- Summary Statistics by Diabetes Status ---

For No Diabetes:

	BMI	PhysHlth	MentHlth
count	213703.0	213703.0	213703.0
mean	27.74	3.58	2.94
std	6.26	8.0	7.06
min	12.0	0.0	0.0
25%	24.0	0.0	0.0
50%	27.0	0.0	0.0
75%	30.0	2.0	2.0
max	98.0	30.0	30.0

For Diabetes:

	BMI	PhysHlth	MentHlth
count	35346.0	35346.0	35346.0
mean	31.94	7.95	4.46
std	7.36	11.3	8.95
min	13.0	0.0	0.0
25%	27.0	0.0	0.0
50%	31.0	1.0	0.0
75%	35.0	15.0	3.0
max	98.0	30.0	30.0

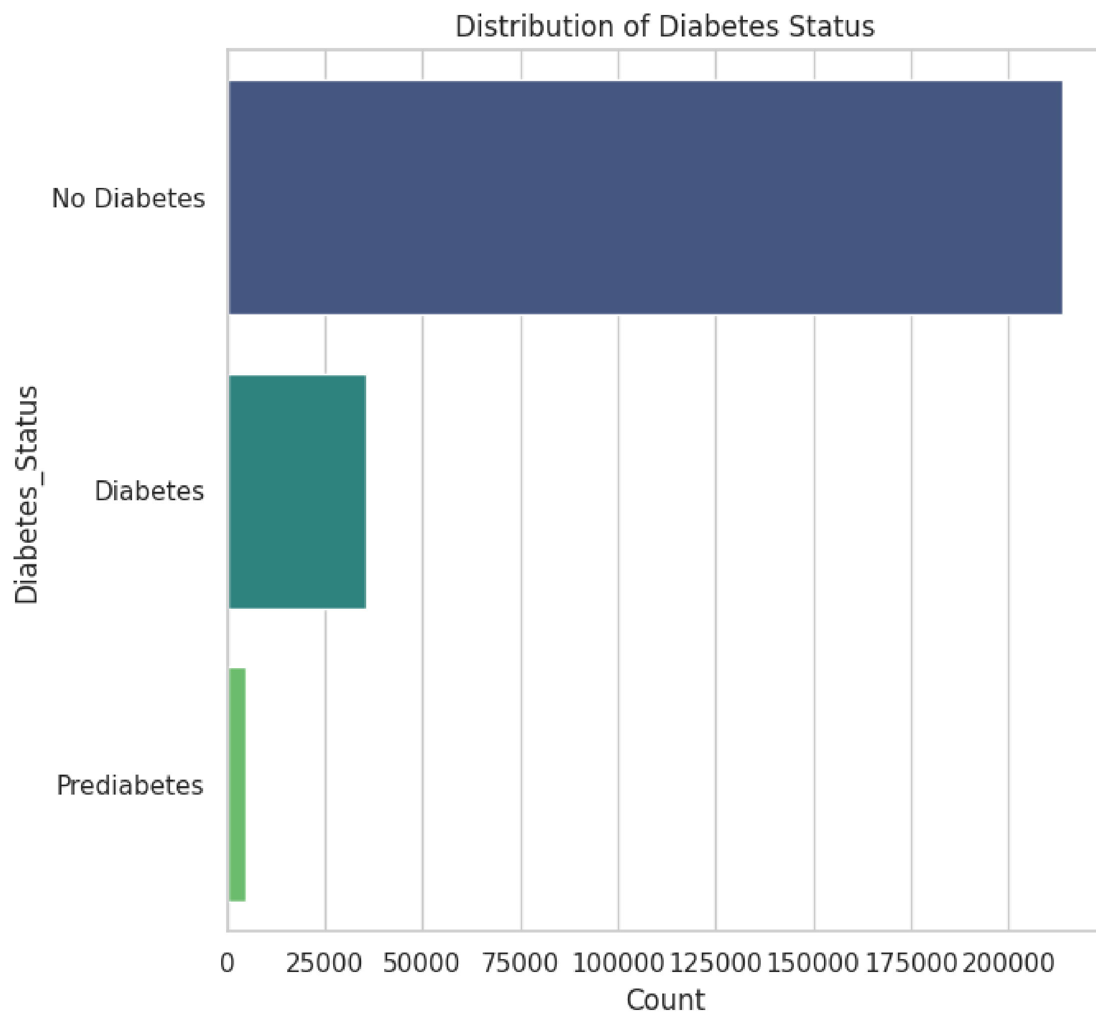
For Prediabetes:

	BMI	PhysHlth	MentHlth
count	4631.0	4631.0	4631.0
mean	30.72	6.35	4.53
std	6.96	10.3	8.9
min	13.0	0.0	0.0
25%	26.0	0.0	0.0
50%	30.0	0.0	0.0
75%	34.0	8.0	4.0
max	96.0	30.0	30.0

The dataset reveals a significant disparity in diabetes prevalence among the studied population. The majority of subjects (approximately 85%) have no diabetes, while only about 13% have diabetes and a mere 2% have prediabetes. This distribution highlights that while diabetes affects a minority of the population, it still represents a substantial health burden given the sample size. The stark contrast between these groups provides a strong basis for comparative analysis of risk factors and demographic patterns associated with the condition.

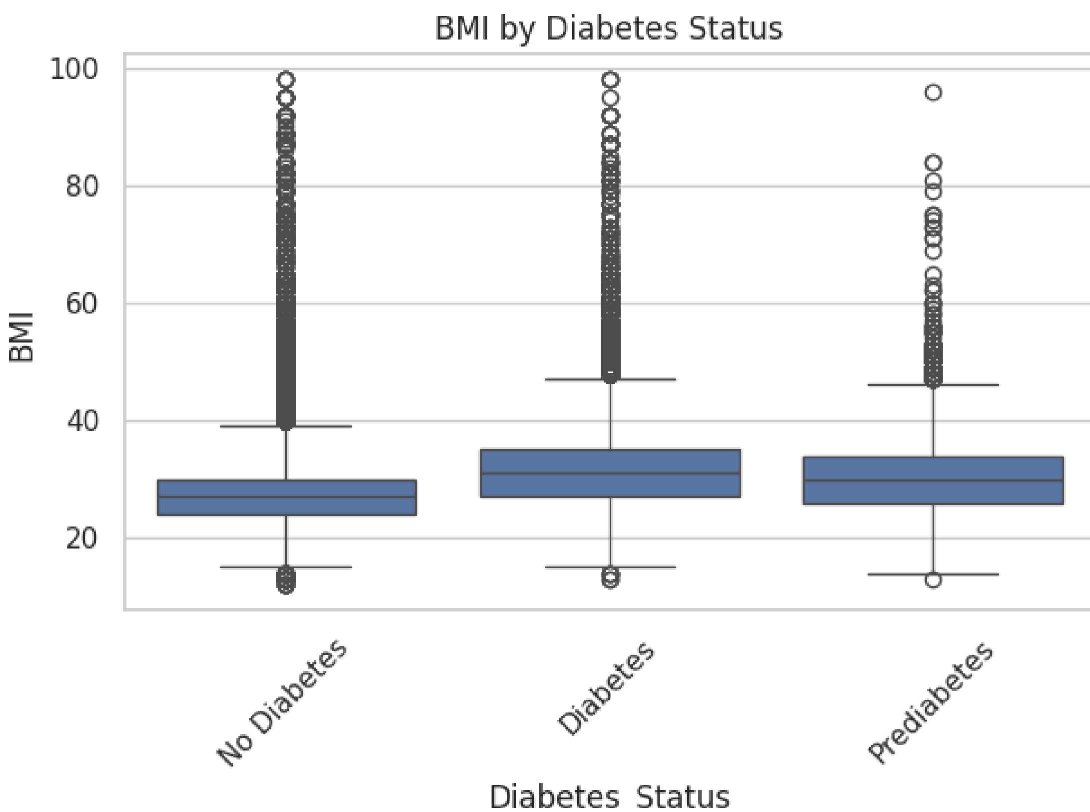
[6]:
```
# Create a dashboard of visualizations
plt.figure(figsize=(18, 12))

# Diabetes Status Distribution
plt.subplot(2, 3, 1)
sns.countplot(y='Diabetes_Status', data=df_clean, palette='viridis')
plt.title('Distribution of Diabetes Status')
plt.xlabel('Count')
plt.tight_layout()
```



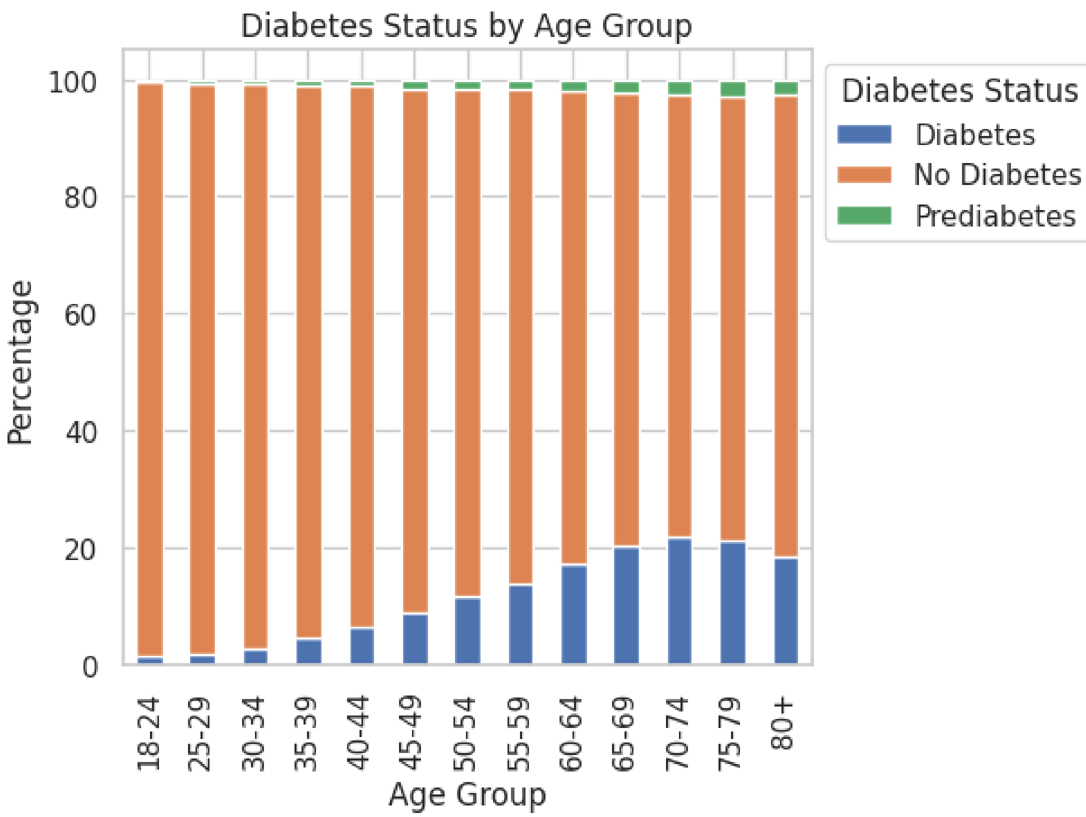
BMI distributions show a clear relationship with diabetes status. Individuals with diabetes and prediabetes demonstrate notably higher median BMI values (approximately 30 and 29, respectively) compared to those without diabetes (approximately 25). The density plot reveals that people without diabetes have a peak BMI distribution around 25, while those with diabetes show a broader distribution with higher concentrations in the overweight (BMI 25-30) and obese (BMI >30) ranges. This visualization confirms BMI as a significant risk factor, with higher values strongly associated with diabetes diagnosis.

[7]:
```python
# BMI vs Diabetes Status
sns.boxplot(x='Diabetes_Status', y='BMI', data=df_clean)
plt.title('BMI by Diabetes Status')
plt.xticks(rotation=45)
plt.tight_layout()
```



Age emerges as a critical factor in diabetes prevalence, with a dramatic increase observed in older age groups. The below graph demonstrates that diabetes rates begin climbing noticeably after age 45, with the steepest increases in the 65-69, 70-74, 75-79, and 80+ age brackets. The visualization reveals that while diabetes affects less than 10% of adults under 45, this rate more than doubles to over 20% in the elderly population. This clear age-related progression suggests that age-appropriate screening and intervention strategies should be prioritized, particularly for individuals entering middle age and beyond.

[8]:
```
# Age Distribution by Diabetes
age_diabetes = pd.crosstab(df_clean['Age'], df_clean['Diabetes_Status'],␣
↪normalize='index') * 100
age_diabetes.plot(kind='bar', stacked=True)
plt.title('Diabetes Status by Age Group')
plt.xlabel('Age Group')
plt.ylabel('Percentage')
plt.xticks(rotation=90)
plt.legend(title='Diabetes Status', loc='upper left', bbox_to_anchor=(1, 1))
plt.tight_layout()
```



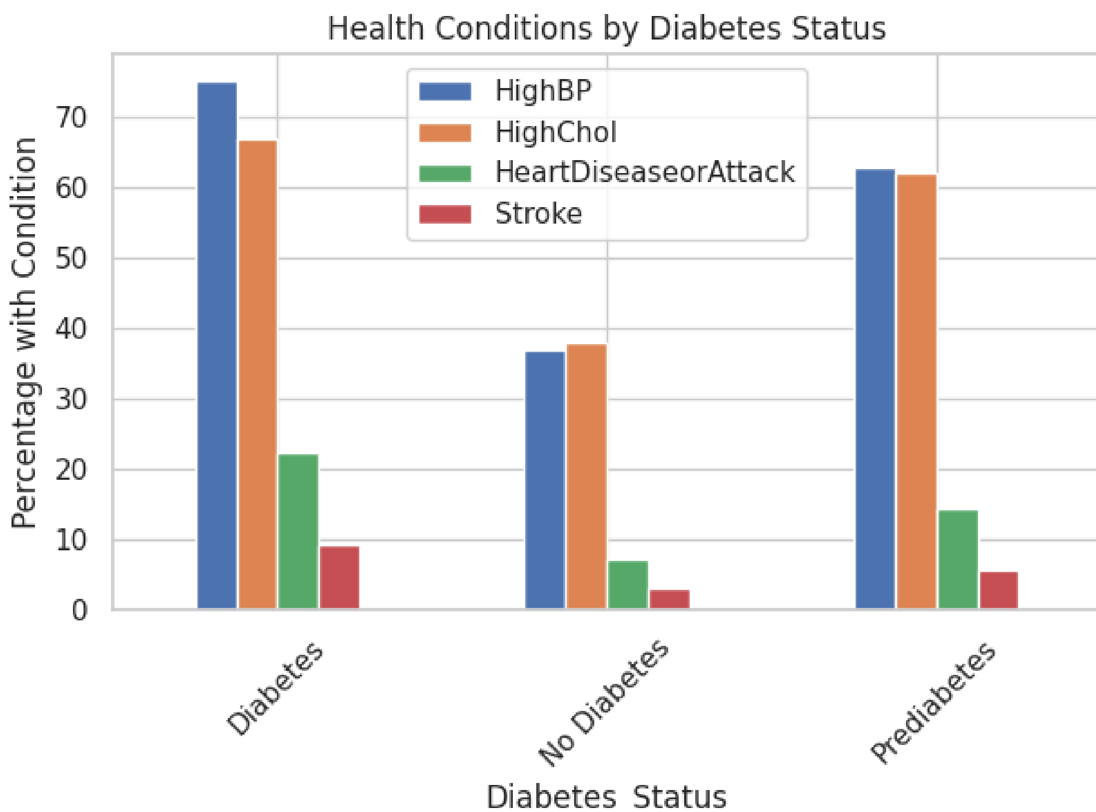
Individuals with diabetes show substantially higher rates of comorbid conditions. Those with diabetes have markedly elevated rates of high blood pressure (73%), high cholesterol (67%), and heart disease (22%) compared to non-diabetic individuals (38%, 35%, and 7% respectively). Interestingly, prediabetic individuals also show higher comorbidity rates than the non-diabetic group, suggesting that these conditions may develop along a continuum with prediabetes representing an intermediate risk state. These patterns underscore the interconnected nature of metabolic and cardiovascular conditions, highlighting the importance of comprehensive care approaches.

[9]:
```
# Health Metrics Comparison
health_vars = ['HighBP', 'HighChol', 'HeartDiseaseorAttack', 'Stroke']
```

```
yes_percentages = {}

for var in health_vars:
    yes_percentages[var] = pd.crosstab(df_clean['Diabetes_Status'],␣
↪df_clean[var])['Yes'] / \
                              df_clean['Diabetes_Status'].value_counts() * 100

pd.DataFrame(yes_percentages).plot(kind='bar')
plt.title('Health Conditions by Diabetes Status')
plt.ylabel('Percentage with Condition')
plt.xticks(rotation=45)
plt.tight_layout()
```

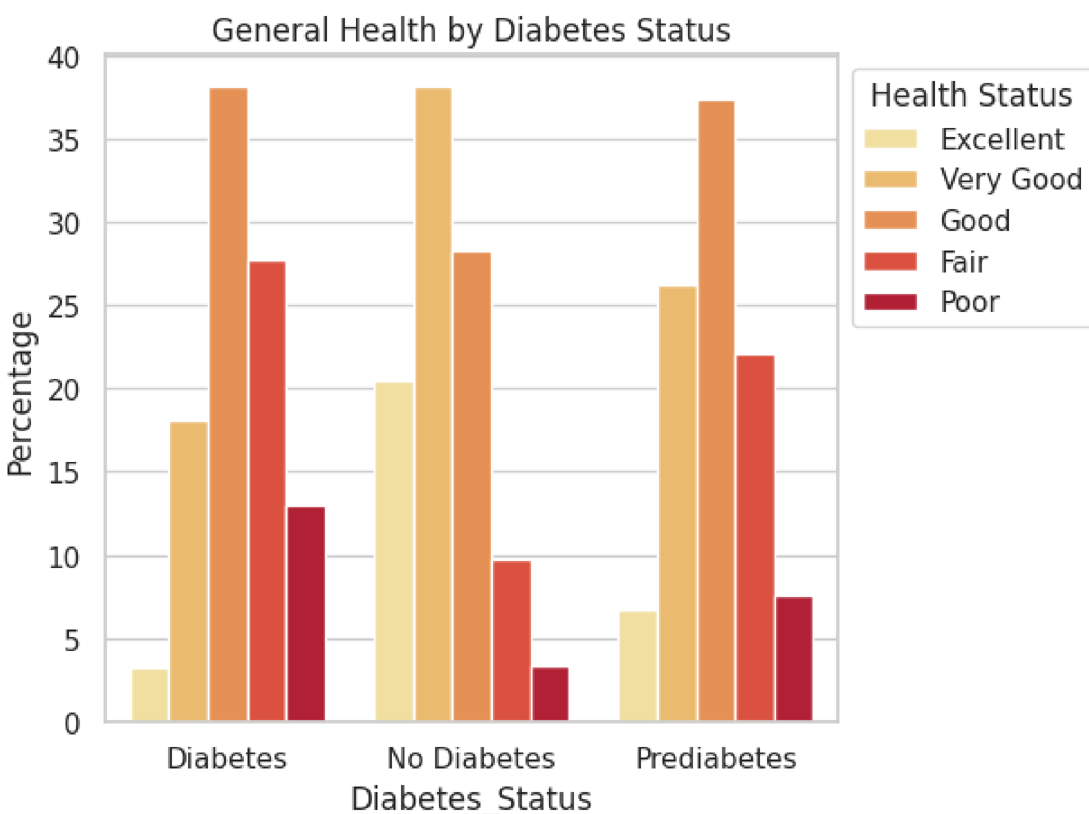


Self-reported general health shows a striking correlation with diabetes status. The graph below reveals that individuals without diabetes most frequently report "very good" health (38%), while those with diabetes more commonly report only "good" (27%) or "fair" (28%) health, with very few reporting "excellent" health (3%). This suggests that diabetes significantly impacts perceived well-being and quality of life. Prediabetic individuals show an intermediate pattern, with health ratings falling between the other two groups, further supporting the concept of prediabetes as a transitional state in terms of both physical health and subjective well-being.

[10]:
```
# General Health by Diabetes
health_order = ['Excellent', 'Very Good', 'Good', 'Fair', 'Poor']
gen_health = pd.crosstab(df_clean['Diabetes_Status'], df_clean['GenHlth'])
gen_health_pct = gen_health.div(gen_health.sum(axis=1), axis=0) * 100

# Select only the ordered health categories and convert to DataFrame for␣
↪plotting
gen_health_pct_ordered = gen_health_pct[health_order].reset_index()
gen_health_pct_ordered = pd.melt(gen_health_pct_ordered,␣
↪id_vars=['Diabetes_Status'],
                                value_vars=health_order)

sns.barplot(x='Diabetes_Status', y='value', hue='GenHlth',␣
↪data=gen_health_pct_ordered,
           hue_order=health_order, palette='YlOrRd')
plt.title('General Health by Diabetes Status')
plt.ylabel('Percentage')
plt.legend(title='Health Status', loc='upper left', bbox_to_anchor=(1, 1))
plt.tight_layout()
```



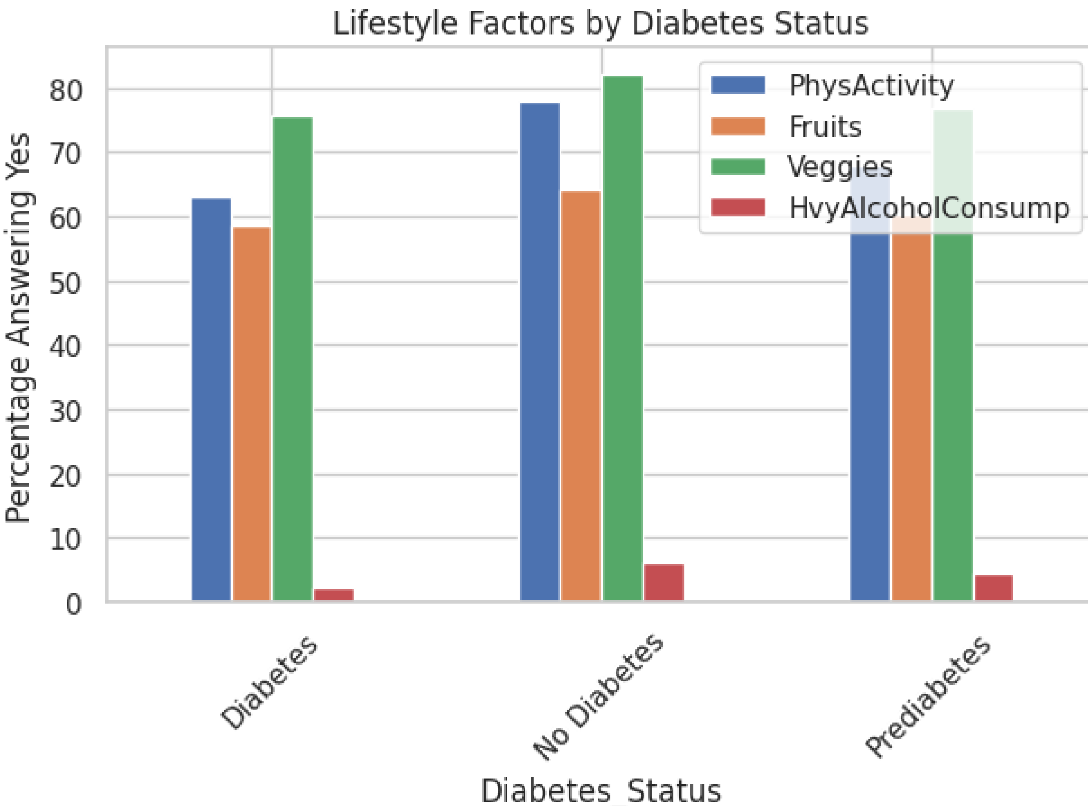
The analysis of lifestyle behaviors in the bar chart below reveals meaningful differences across diabetes status groups. People without diabetes show higher rates of positive health behaviors: physical activity (78%), fruit consumption (63%), and vegetable intake (81%) compared to diabetic individuals (63%, 60%, and 75% respectively). Heavy alcohol consumption is generally low across all groups but slightly higher in those without diabetes. These patterns suggest that lifestyle modifications might be both preventive for those at risk and therapeutic for those already diagnosed with diabetes or prediabetes, with particular emphasis on increasing physical activity.

[11]:
```
# Lifestyle Factors
lifestyle = ['PhysActivity', 'Fruits', 'Veggies', 'HvyAlcoholConsump']
lifestyle_yes = {}

for var in lifestyle:
    lifestyle_yes[var] = pd.crosstab(df_clean['Diabetes_Status'],␣
 ↪df_clean[var])['Yes'] / \
                        df_clean['Diabetes_Status'].value_counts() * 100

pd.DataFrame(lifestyle_yes).plot(kind='bar')
plt.title('Lifestyle Factors by Diabetes Status')
plt.ylabel('Percentage Answering Yes')
plt.xticks(rotation=45)
plt.tight_layout()

plt.subplots_adjust(wspace=0.3, hspace=0.4)
plt.tight_layout()
plt.show()
```

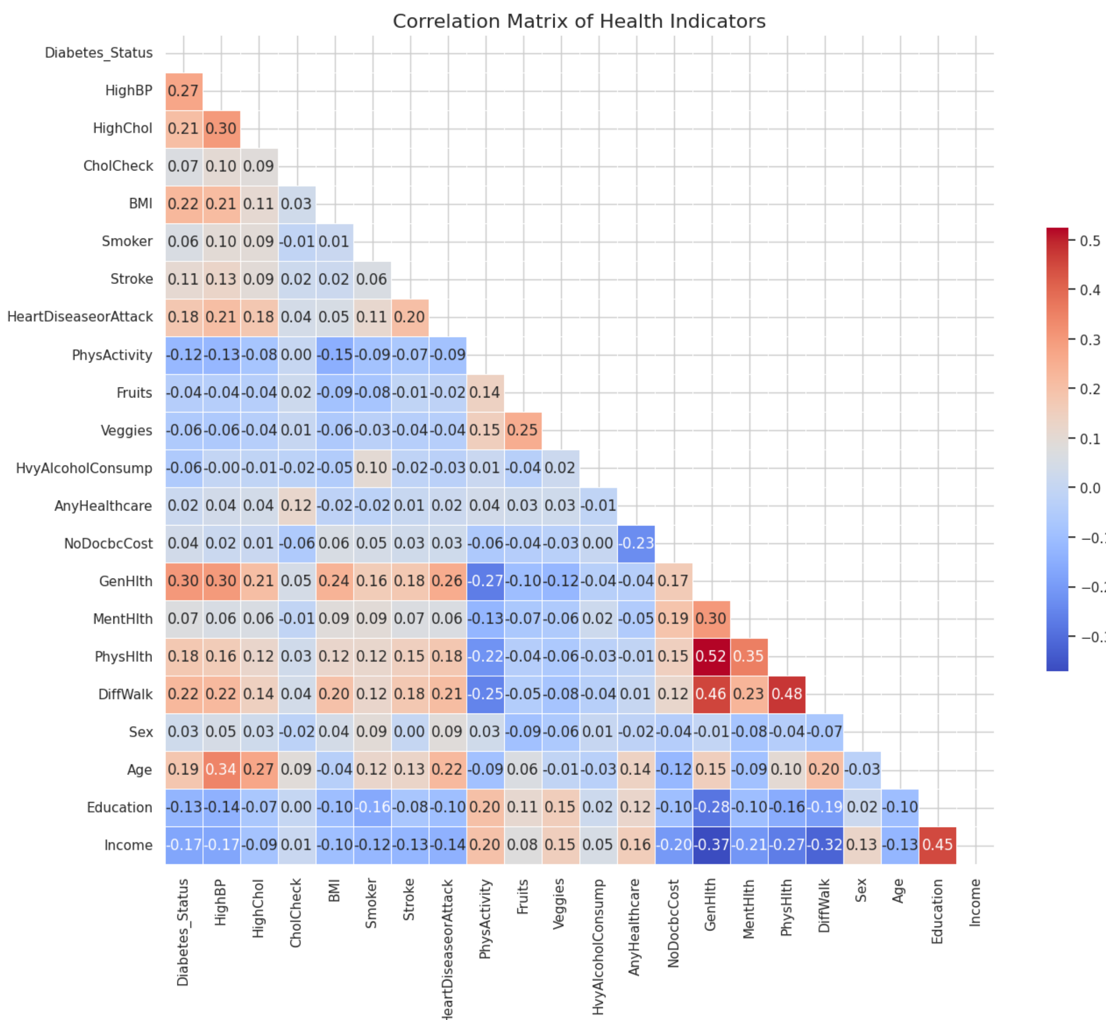


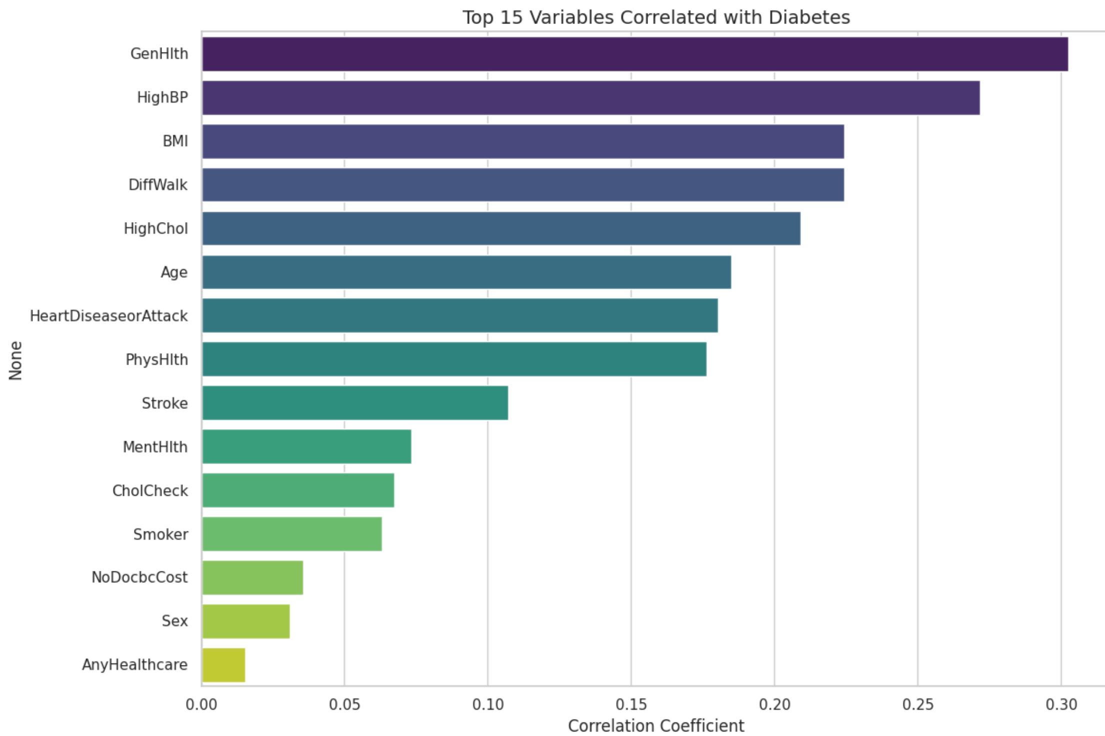
The correlation matrix provides a comprehensive view of the relationships between health variables. Diabetes status shows the strongest positive correlations with general health (0.30), high blood pressure (0.27), BMI (0.22), and difficulty walking (0.22). Prior figures confirms these as the top factors associated with diabetes. Physical activity shows a negative correlation (-0.12), indicating its protective effect. This multifactorial correlation analysis reinforces the complex, interconnected nature of diabetes with various physiological, behavioral, and demographic factors, suggesting that comprehensive assessment and intervention approaches are necessary.

[12]:
```
# Correlation analysis using original numeric data
numeric_df = df.select_dtypes(include=['int64', 'float64'])
corr = numeric_df.corr()

# Plot correlation heatmap
plt.figure(figsize=(14, 12))
mask = np.triu(np.ones_like(corr, dtype=bool))
sns.heatmap(corr, mask=mask, annot=True, fmt=".2f", cmap="coolwarm",
            square=True, linewidths=.5, cbar_kws={"shrink": .5})
plt.title('Correlation Matrix of Health Indicators', fontsize=16)
plt.tight_layout()
plt.show()
```

```python
# Extract and plot top correlations with Diabetes_Stats
diabetes_corr = corr['Diabetes_Status'].sort_values(ascending=False)
plt.figure(figsize=(12, 8))
sns.barplot(x=diabetes_corr.values[1:16], y=diabetes_corr.index[1:16], 
 hue=diabetes_corr
            .index[1:16], palette='viridis', legend=False)
plt.title('Top 15 Variables Correlated with Diabetes', fontsize=14)
plt.xlabel('Correlation Coefficient')
plt.tight_layout()
plt.show()
```



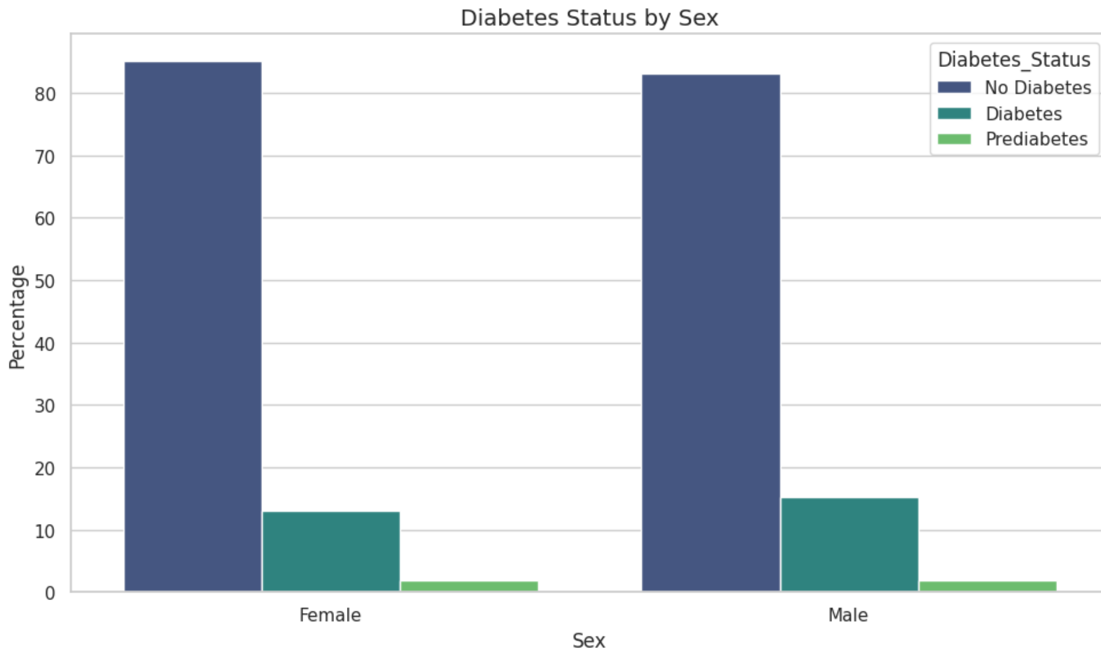


The prevalence of diabetes shows minimal variation between sexes. Both females and males exhibit similar patterns with approximately 13-15% having diabetes and 1-2% having prediabetes. This suggests that biological sex alone may not be a strong independent risk factor for diabetes, though interactions between sex and other risk factors could still be clinically relevant. The comparable rates across sexes indicate that diabetes prevention and management strategies should target both men and women, with perhaps more emphasis on risk factors that transcend sex differences.

[13]:
```python
# Sex and Diabetes
plt.figure(figsize=(10, 6))
sex_data = []
for sex in df_clean['Sex'].unique():
    for status in df_clean['Diabetes_Status'].unique():
        subset = df_clean[(df_clean['Sex'] == sex)]
        pct = (subset['Diabetes_Status'] == status).mean() * 100
        sex_data.append({'Sex': sex, 'Diabetes_Status': status, 'Percentage':␣
↪pct})

sex_df = pd.DataFrame(sex_data)
sns.barplot(x='Sex', y='Percentage', hue='Diabetes_Status', data=sex_df,␣
↪palette='viridis')
plt.title('Diabetes Status by Sex', fontsize=14)
plt.ylabel('Percentage')
plt.tight_layout()
```

```
plt.show()
```

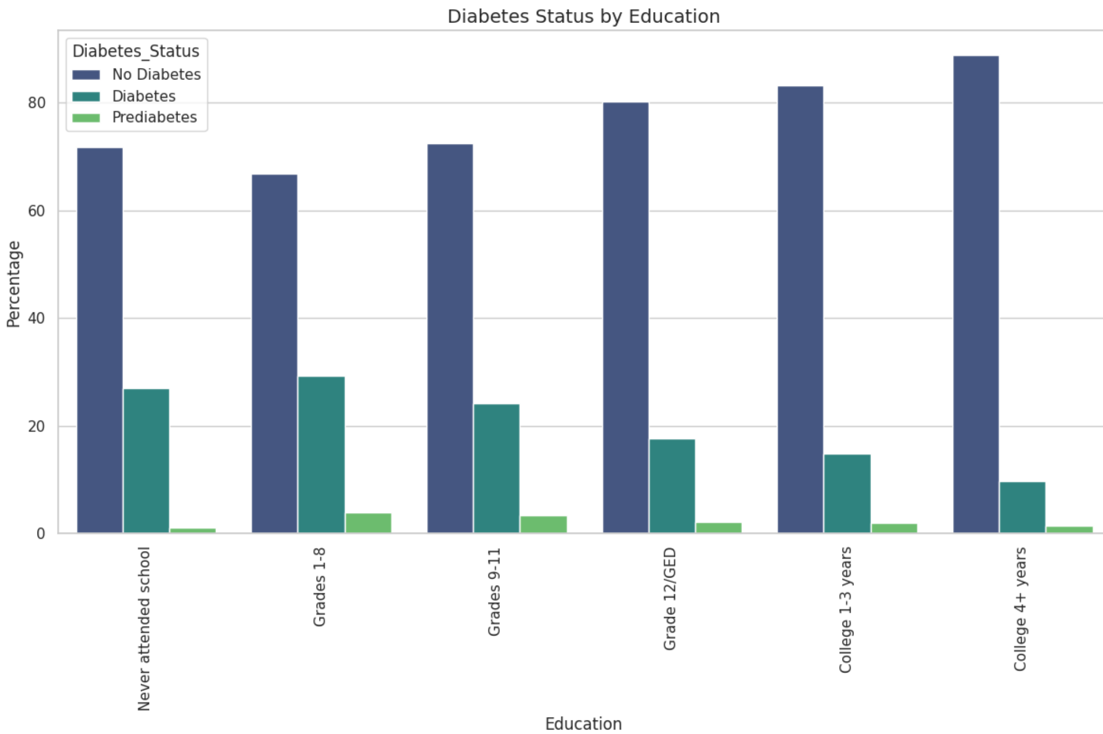


Education level demonstrates a clear inverse relationship with diabetes prevalence. The comparison below shows that individuals with higher education levels (college education) have significantly lower diabetes rates (10% for those with 4+ years of college) compared to those with less education (27% for those who never attended school). Similarly, reveals that higher income levels are associated with lower diabetes prevalence, with rates decreasing from 25% in the lowest income bracket to just 8% in the highest. These socioeconomic gradients highlight the social determinants of health and suggest that educational initiatives and economic policies could indirectly impact diabetes prevalence by addressing these underlying disparities.

[14]:
```
# Education Factors
plt.figure(figsize=(12, 8))
edu_order = [v for k, v in sorted(education_map.items()) if v not in ['No␣
↪Response']]
edu_data = []
for edu in edu_order:
    if edu in df_clean['Education'].values:
        for status in df_clean['Diabetes_Status'].unique():
            subset = df_clean[(df_clean['Education'] == edu)]
            if len(subset) > 0:
                pct = (subset['Diabetes_Status'] == status).mean() * 100
                edu_data.append({'Education': edu, 'Diabetes_Status': status,␣
↪'Percentage': pct})

edu_df = pd.DataFrame(edu_data)
```

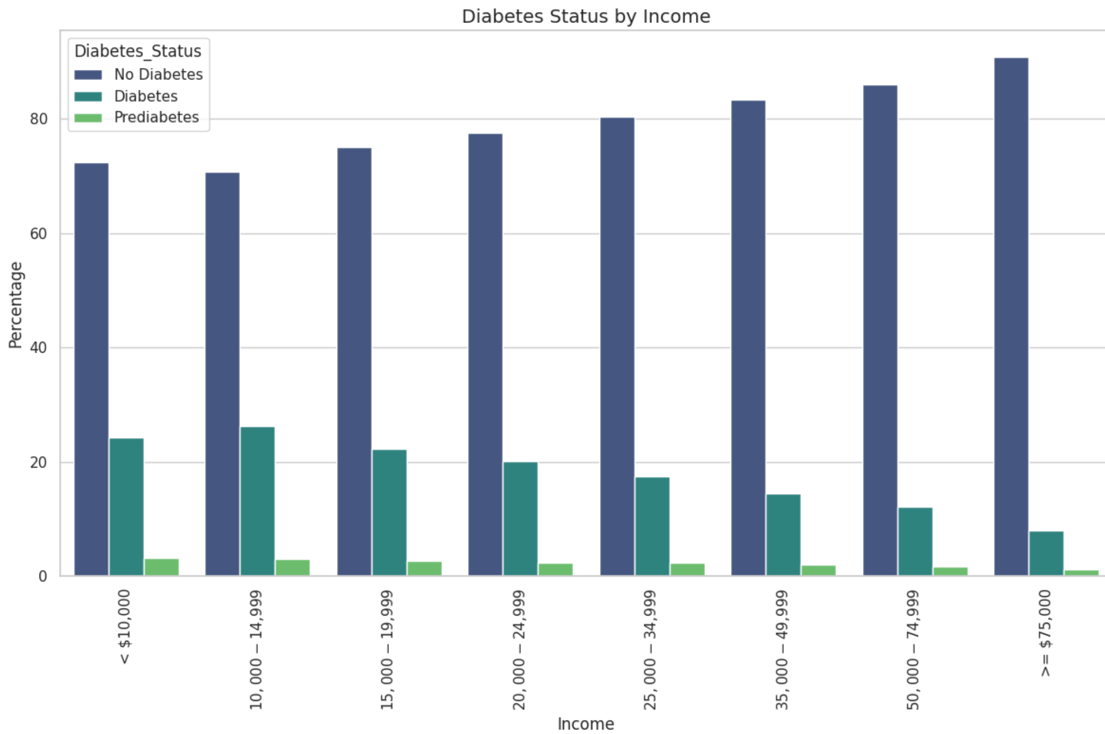
```
sns.barplot(x='Education', y='Percentage', hue='Diabetes_Status', data=edu_df,␣
↪palette='viridis')
plt.title('Diabetes Status by Education', fontsize=14)
plt.ylabel('Percentage')
plt.xticks(rotation=90)
plt.tight_layout()
plt.show()
```



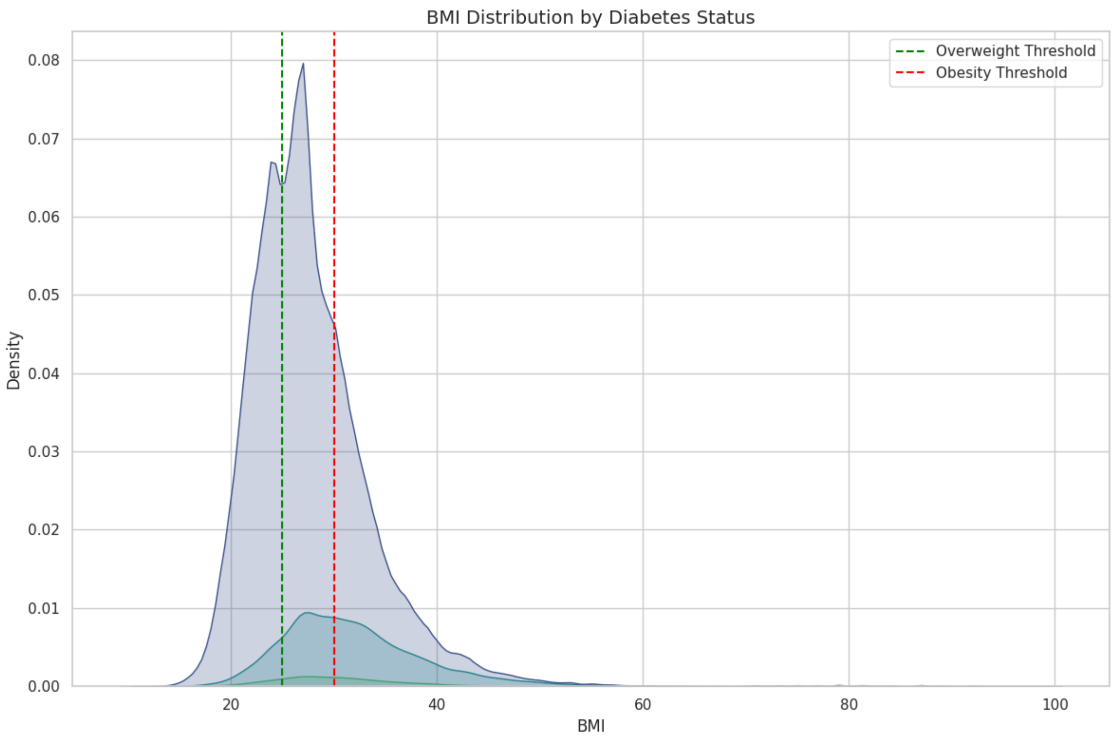
[15]:
```
# Income and Diabetes
plt.figure(figsize=(12, 8))
income_order = [v for k, v in sorted(income_map.items()) if v not in ['Not␣
↪Sure', 'No Response']]
income_data = []
for inc in income_order:
    if inc in df_clean['Income'].values:
        for status in df_clean['Diabetes_Status'].unique():
            subset = df_clean[(df_clean['Income'] == inc)]
            if len(subset) > 0:
                pct = (subset['Diabetes_Status'] == status).mean() * 100
                income_data.append({'Income': inc, 'Diabetes_Status': status,␣
↪'Percentage': pct})

income_df = pd.DataFrame(income_data)
```

```python
sns.barplot(x='Income', y='Percentage', hue='Diabetes_Status', data=income_df,␣
↪palette='viridis')
plt.title('Diabetes Status by Income', fontsize=14)
plt.ylabel('Percentage')
plt.xticks(rotation=90)
plt.tight_layout()
plt.show()
```



[16]:
```python
# BMI Distribution by Diabetes Status
plt.figure(figsize=(12, 8))
sns.kdeplot(data=df_clean, x='BMI', hue='Diabetes_Status', palette='viridis',␣
↪fill=True)
plt.title('BMI Distribution by Diabetes Status', fontsize=14)
plt.xlabel('BMI')
plt.ylabel('Density')
plt.axvline(x=25, color='green', linestyle='--', label='Overweight Threshold')
plt.axvline(x=30, color='red', linestyle='--', label='Obesity Threshold')
plt.legend()
plt.tight_layout()
plt.show()
```

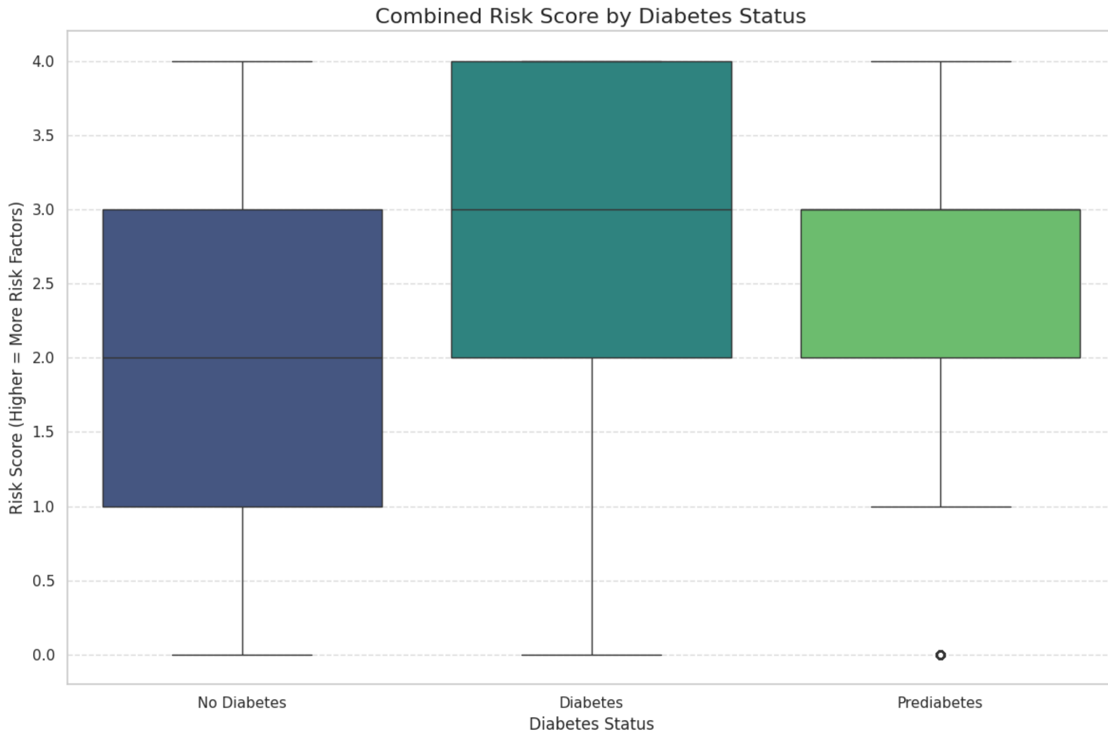


The combined risk score analysis below illustrates how risk factors accumulate differently across diabetes status groups. Individuals with diabetes show significantly higher median risk scores and wider variability in their risk profiles compared to those without diabetes. This analysis suggests that diabetes is often accompanied by a constellation of risk factors rather than isolated abnormalities. The violin plot of BMI by physical activity further demonstrates how lifestyle factors interact with metabolic parameters across diabetes status groups, with physical activity associated with lower BMI distributions regardless of diabetes status.

[17]:
```
# Risk Factor Analysis
plt.figure(figsize=(12, 8))
risk_df = df.copy()
risk_df['BMI_Risk'] = risk_df['BMI'].apply(lambda x: 0 if x < 25 else (1 if x <␣
↪30 else 2))
risk_df['Total_Risk'] = risk_df['HighBP'] + risk_df['HighChol'] +␣
↪risk_df['BMI_Risk']
risk_df['Diabetes_Status'] = df_clean['Diabetes_Status']

sns.boxplot(x='Diabetes_Status', y='Total_Risk', data=risk_df,␣
↪palette='viridis')
plt.title('Combined Risk Score by Diabetes Status', fontsize=16)
plt.xlabel('Diabetes Status')
plt.ylabel('Risk Score (Higher = More Risk Factors)')
plt.grid(axis='y', linestyle='--', alpha=0.7)
```

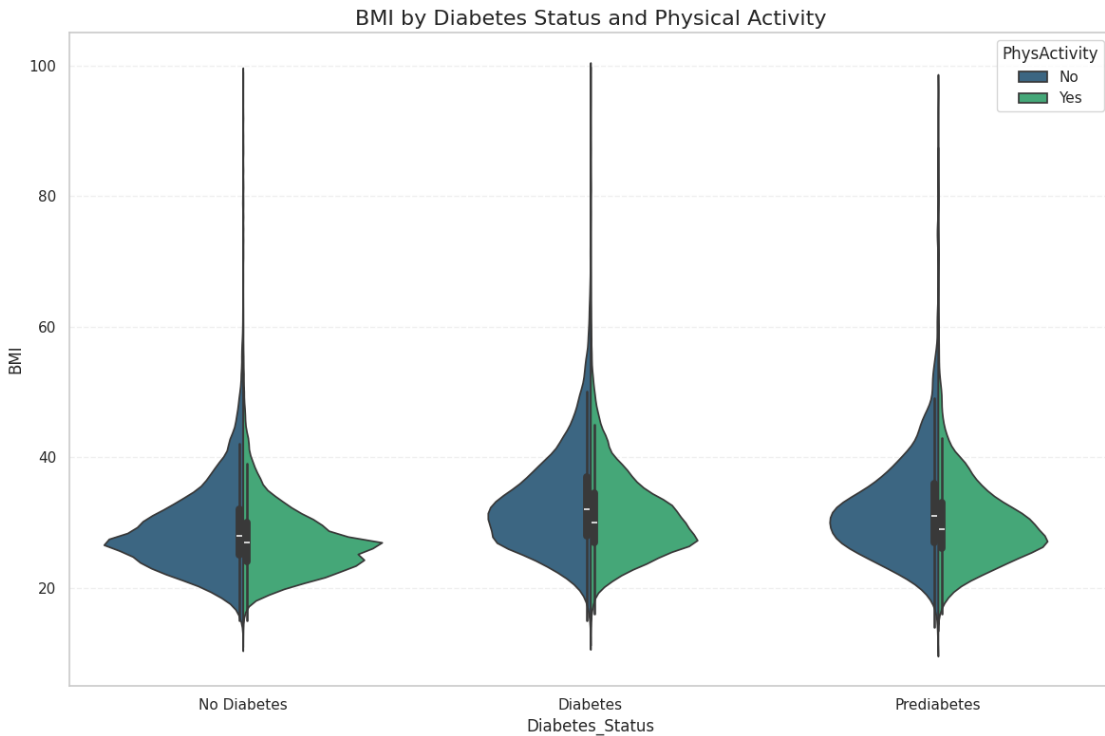
```
plt.tight_layout()
plt.show()
```



The scatter plot exploring the relationship between age, BMI, and diabetes status reveals complex interactions between these variables. While higher BMI values are more frequently associated with diabetes regardless of age, the distribution of points suggests that the BMI threshold for diabetes risk may vary across age groups. This visualization helps identify particularly vulnerable populations—those with both advanced age and elevated BMI—who may benefit most from targeted screening and intervention efforts.

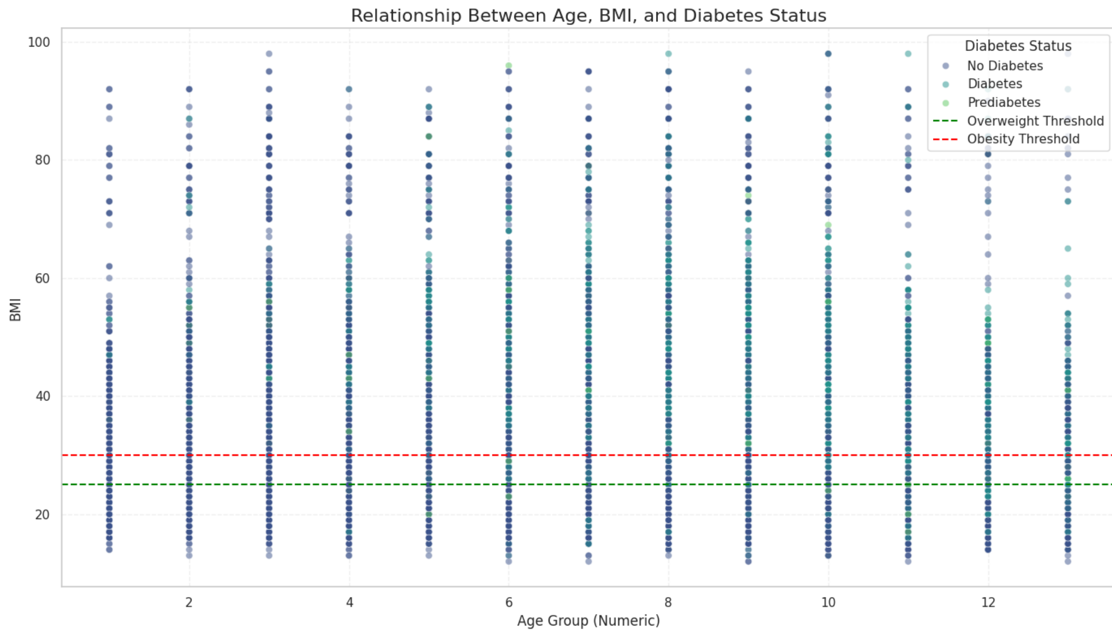
[18]:
```
# BMI and Physical Activity Relationship
plt.figure(figsize=(12, 8))
physical_act_data = pd.DataFrame({
    'BMI': df['BMI'],
    'PhysActivity': df_clean['PhysActivity'],
    'Diabetes_Status': df_clean['Diabetes_Status']
})

sns.violinplot(x='Diabetes_Status', y='BMI', hue='PhysActivity',
               data=physical_act_data, palette='viridis', split=True)
plt.title('BMI by Diabetes Status and Physical Activity', fontsize=16)
plt.grid(axis='y', linestyle='--', alpha=0.3)
plt.tight_layout()
plt.show()
```



[19]:
```
# Age, BMI, and Diabetes Status
plt.figure(figsize=(14, 8))
age_bmi_data = pd.DataFrame({
    'Age_Numeric': df['Age'],
    'BMI': df['BMI'],
    'Diabetes_Status': df_clean['Diabetes_Status']
})

sns.scatterplot(x='Age_Numeric', y='BMI', hue='Diabetes_Status',
               data=age_bmi_data, palette='viridis', alpha=0.5)
plt.title('Relationship Between Age, BMI, and Diabetes Status', fontsize=16)
plt.xlabel('Age Group (Numeric)')
plt.ylabel('BMI')
plt.axhline(y=25, color='green', linestyle='--', label='Overweight Threshold')
plt.axhline(y=30, color='red', linestyle='--', label='Obesity Threshold')
plt.legend(title='Diabetes Status')
plt.grid(True, linestyle='--', alpha=0.3)
plt.tight_layout()
plt.show()
```



2 Conclusion

This exploratory data analysis reveals diabetes as a complex condition with multiple interrelated risk factors spanning demographics, lifestyle behaviors, comorbidities, and socioeconomic indicators. The clear patterns observed across BMI distributions, age groups, comorbidity rates, and socioeconomic gradients provide valuable insights for developing targeted prevention strategies and personalized interventions. The analysis particularly highlights the importance of addressing modifiable factors such as physical activity and diet, while recognizing the influence of social determinants like education and income on diabetes risk. These findings can inform both clinical approaches to diabetes management and public health policies aimed at reducing diabetes burden in the population.