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

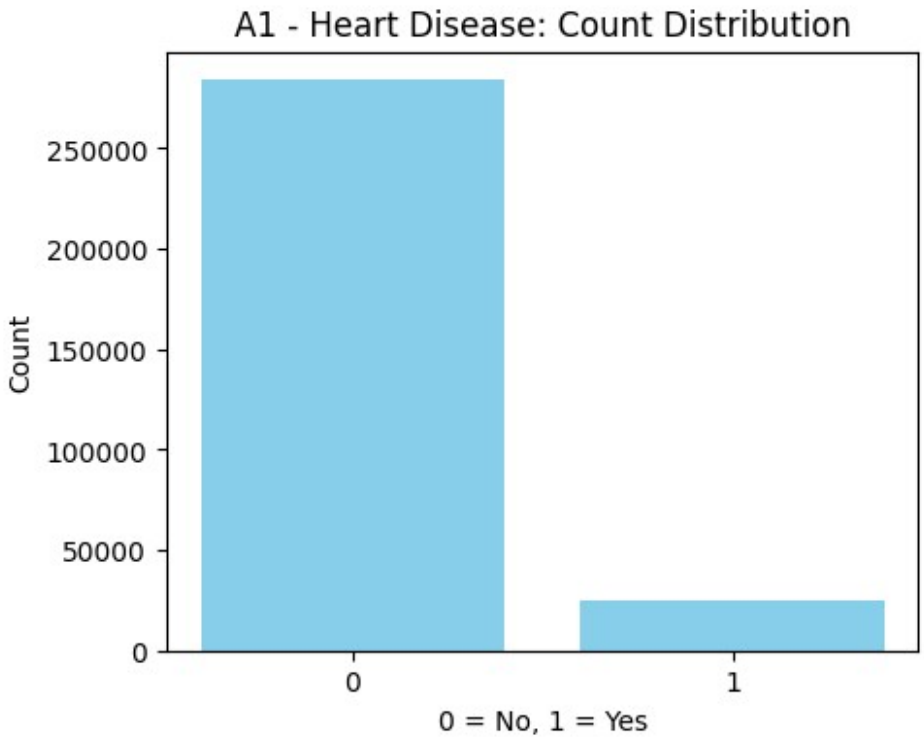
```
<IPython.core.display.HTML object>

Saving full_patient_dataset.csv to full_patient_dataset (1).csv
Saving high_risk_patient_segment.csv to high_risk_patient_segment
(1).csv
```

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


full = pd.read_csv("/content/full_patient_dataset.csv")


print(full.head())
```

```
  General_Health                  Checkup Exercise  Heart_Disease  \
0           Poor  Within the past 2 years       No              0
1      Very Good     Within the past year       No              1
2      Very Good     Within the past year      Yes              0
3           Poor     Within the past year      Yes              1
4           Good     Within the past year       No              0

   Skin_Cancer Other_Cancer  Depression  Diabetes  Arthritis
Sex  \
0            0           No           0       0.0          1  Female

1            0           No           0       1.0          0  Female

2            0           No           0       1.0          0  Female

3            0           No           0       1.0          0    Male

4            0           No           0       0.0          0    Male


  Age_Category  Height_(cm)  Weight_(kg)    BMI Smoking_History  \
0        70-74        150.0        32.66  14.54             Yes
1        70-74        165.0        77.11  28.29              No
2        60-64        163.0        88.45  33.47              No
3        75-79        180.0        93.44  28.73              No
4          80+        191.0        88.45  24.37             Yes

   Alcohol_Consumption  Fruit_Consumption
Green_Vegetables_Consumption  \
0                  0.0               30.0
16.0
1                  0.0               30.0
```

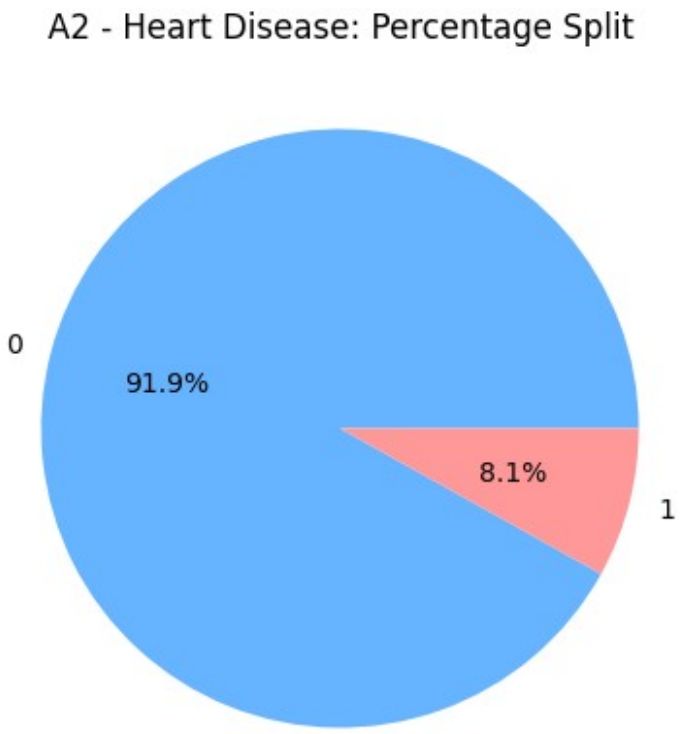
```
0.0
2                    4.0                  12.0
3.0
3                    0.0                  30.0
30.0
4                    0.0                   8.0
4.0

   FriedPotato_Consumption
0                     12.0
1                      4.0
2                     16.0
3                      8.0
4                      0.0
```

```
vc = full["Heart_Disease"].value_counts().sort_index()
fig_a1 = plt.figure(figsize=(5,4))
plt.bar(vc.index.astype(str), vc.values, color="skyblue")
plt.title("A1 - Heart Disease: Count Distribution")
plt.xlabel("0 = No, 1 = Yes")
plt.ylabel("Count")
plt.show()
```

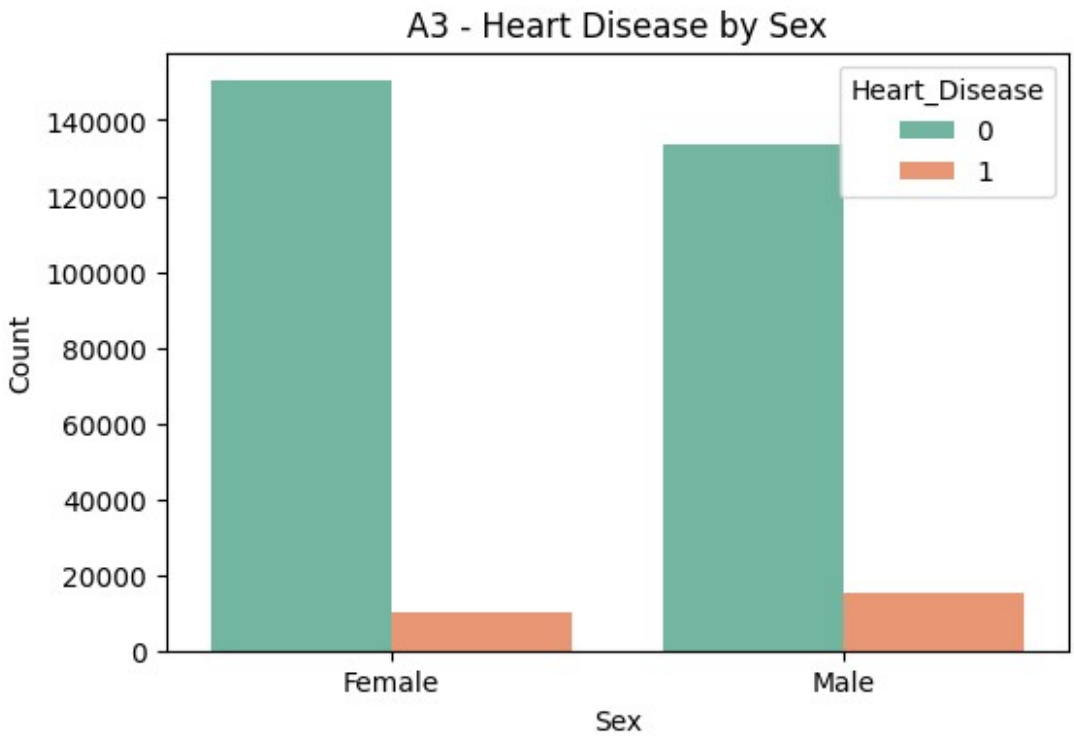



```
vc = full["Heart_Disease"].value_counts().sort_index()
fig_a2 = plt.figure(figsize=(5,5))
plt.pie(vc.values, labels=vc.index.astype(str), autopct="%1.1f%%",
```

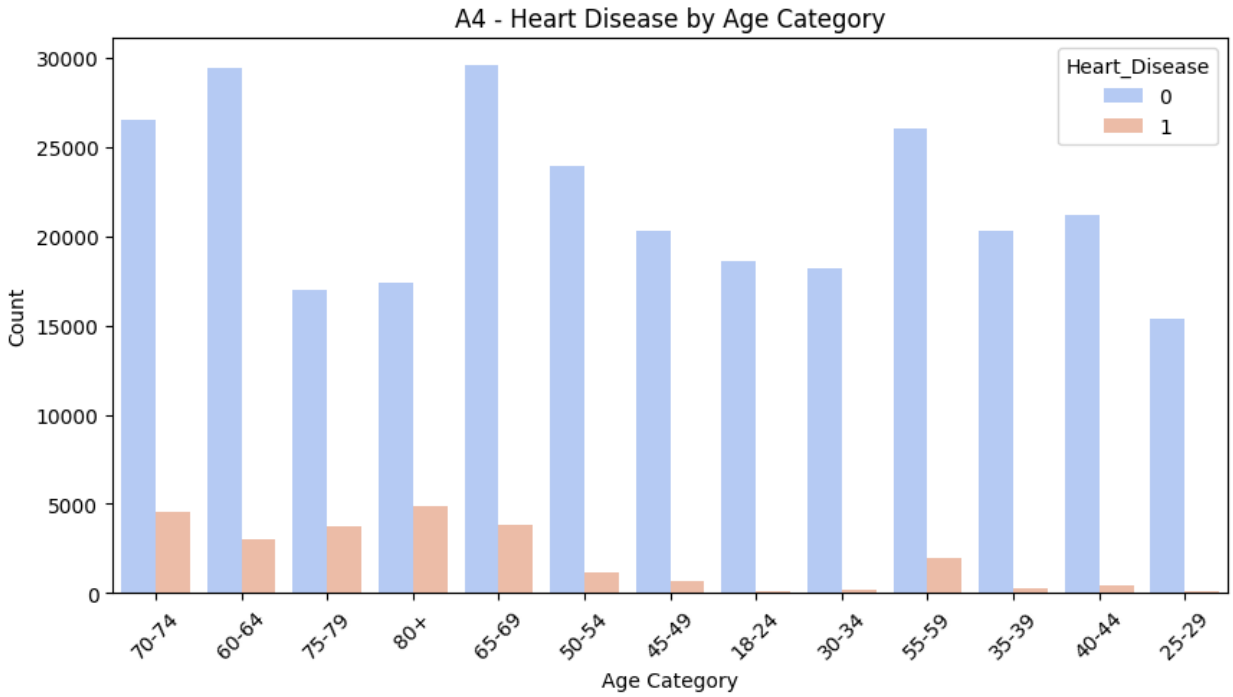
```
colors=["#66b3ff","#ff9999"])
plt.title("A2 - Heart Disease: Percentage Split")
plt.show()
```



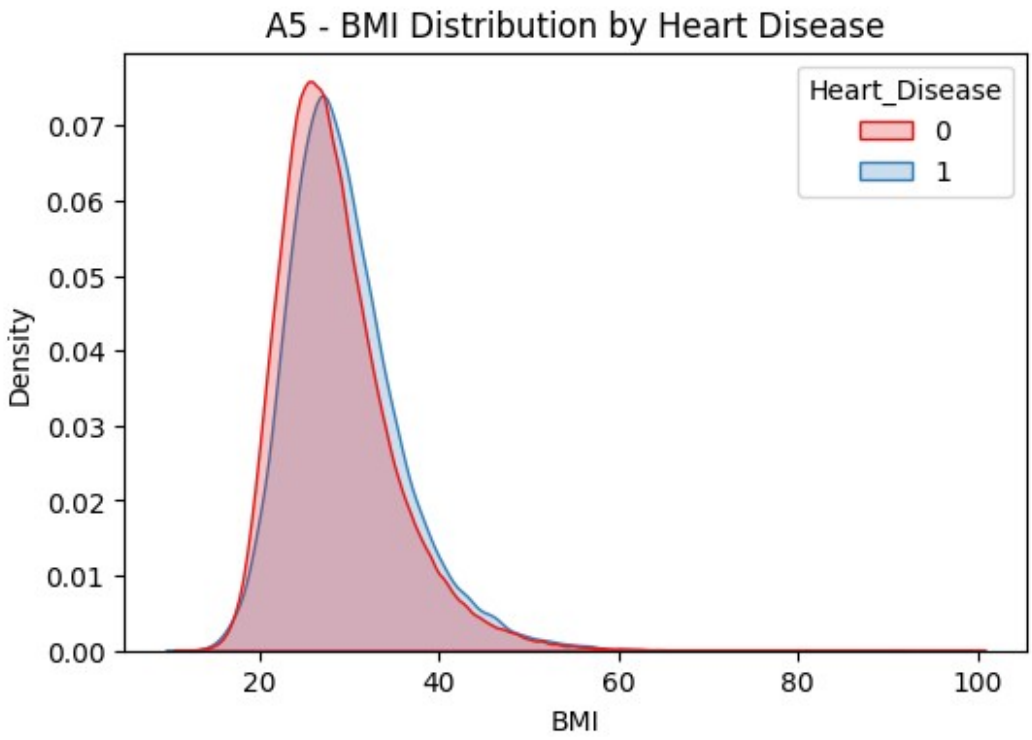

```
fig_a3 = plt.figure(figsize=(6,4))
sns.countplot(data=full, x="Sex", hue="Heart_Disease", palette="Set2")
plt.title("A3 - Heart Disease by Sex")
plt.xlabel("Sex")
plt.ylabel("Count")
plt.show()
```




```
fig_a4 = plt.figure(figsize=(10,5))
sns.countplot(data=full, x="Age_Category", hue="Heart_Disease",
palette="coolwarm")
plt.title("A4 - Heart Disease by Age Category")
plt.xticks(rotation=45)
plt.xlabel("Age Category")
plt.ylabel("Count")
plt.show()
```

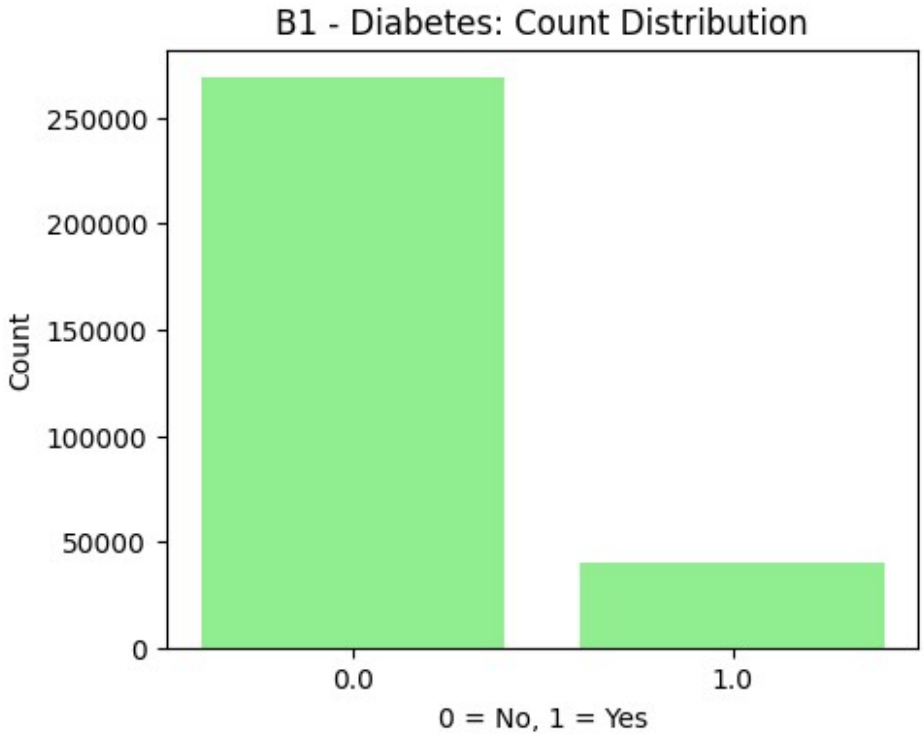


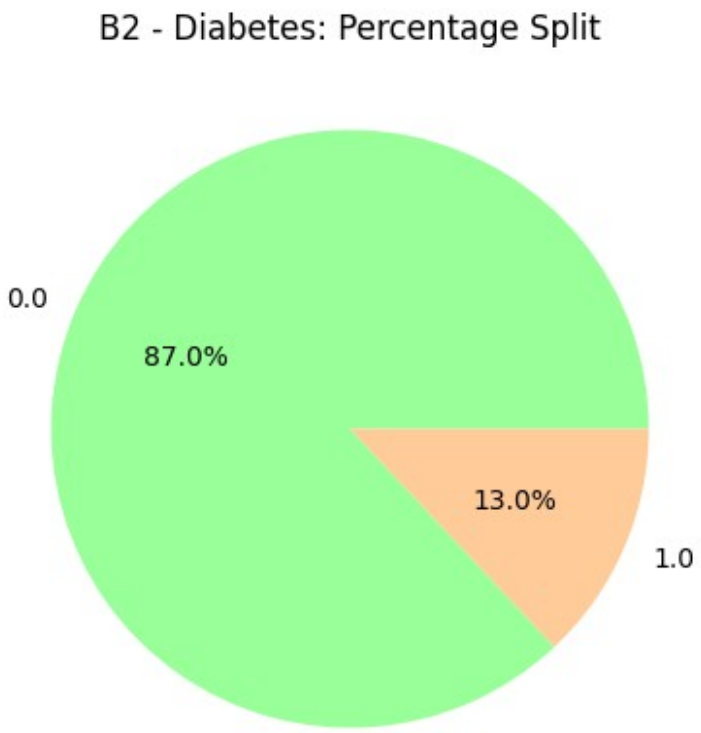

```
fig_a5 = plt.figure(figsize=(6,4))
sns.kdeplot(data=full, x="BMI", hue="Heart_Disease", fill=True,
common_norm=False, palette="Set1")
plt.title("A5 - BMI Distribution by Heart Disease")
plt.xlabel("BMI")
plt.ylabel("Density")
plt.show()
```

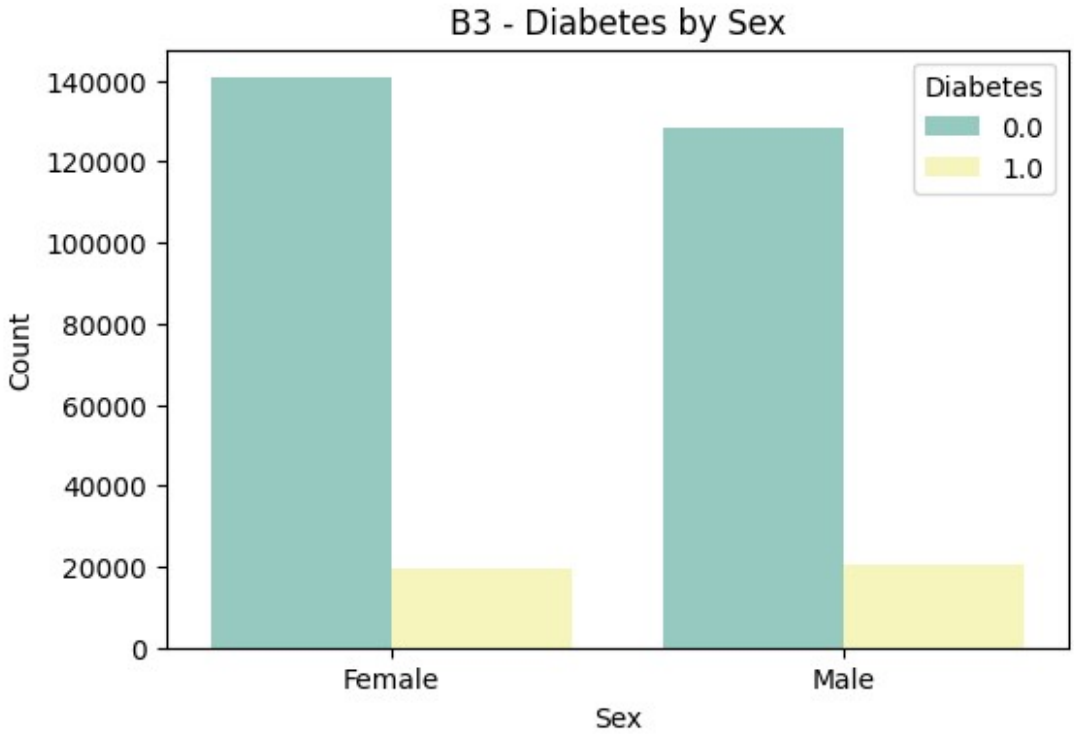


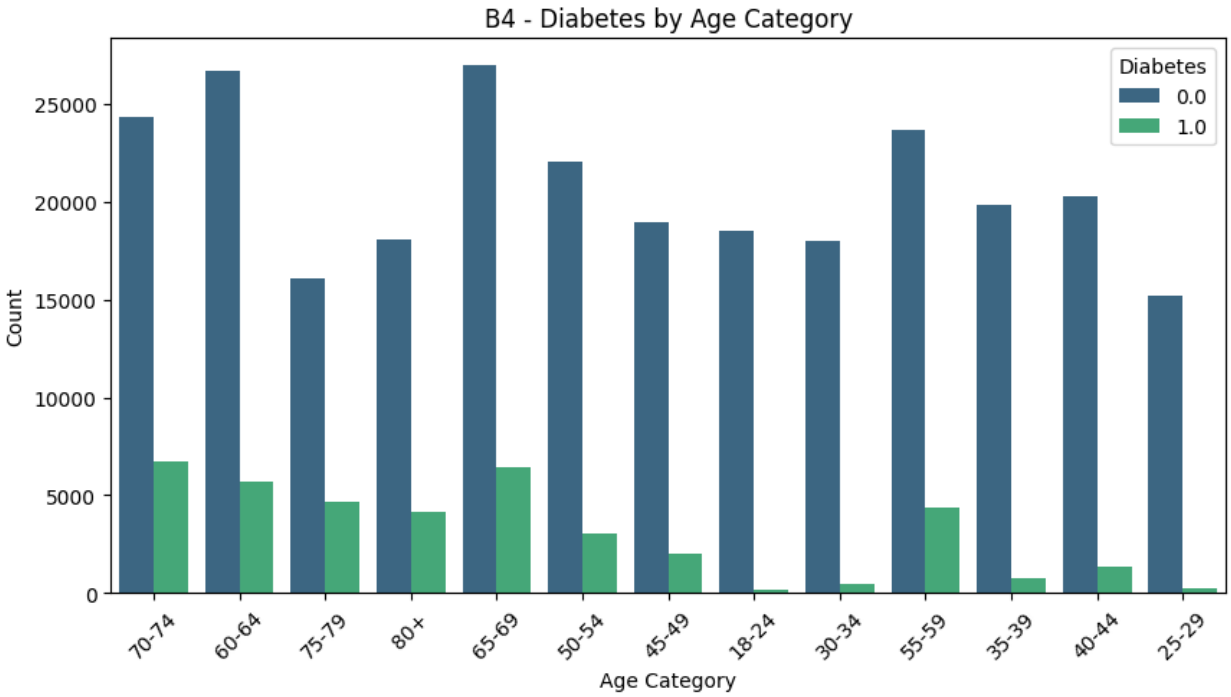

```
# ========================
# B - DIABETES
# ========================

# B1 - Bar Chart
fig_b1 = plt.figure(figsize=(5,4))
vc = full["Diabetes"].value_counts().sort_index()
plt.bar(vc.index.astype(str), vc.values, color="lightgreen")
plt.title("B1 - Diabetes: Count Distribution")
plt.xlabel("0 = No, 1 = Yes")
plt.ylabel("Count")
plt.show()

# B2 - Pie Chart
fig_b2 = plt.figure(figsize=(5,5))
vc = full["Diabetes"].value_counts().sort_index()
plt.pie(vc.values, labels=vc.index.astype(str), autopct="%1.1f%%",
colors=["#99ff99","#ffcc99"])
plt.title("B2 - Diabetes: Percentage Split")
plt.show()

# B3 - By Sex
fig_b3 = plt.figure(figsize=(6,4))
sns.countplot(data=full, x="Sex", hue="Diabetes", palette="Set3")
plt.title("B3 - Diabetes by Sex")
plt.xlabel("Sex")
plt.ylabel("Count")
```

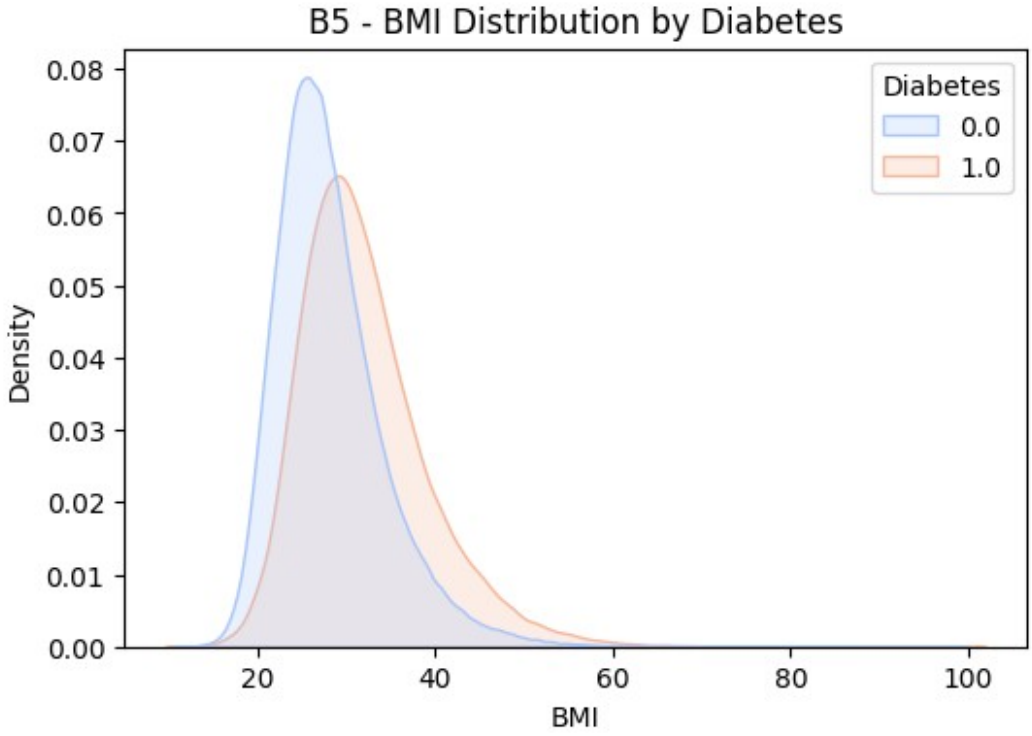
```
plt.show()

# B4 - By Age Category
fig_b4 = plt.figure(figsize=(10,5))
sns.countplot(data=full, x="Age_Category", hue="Diabetes",
palette="viridis")
plt.title("B4 - Diabetes by Age Category")
plt.xticks(rotation=45)
plt.xlabel("Age Category")
plt.ylabel("Count")
plt.show()

# B5 - BMI Distribution
fig_b5 = plt.figure(figsize=(6,4))
sns.kdeplot(data=full, x="BMI", hue="Diabetes", fill=True,
common_norm=False, palette="coolwarm")
plt.title("B5 - BMI Distribution by Diabetes")
plt.xlabel("BMI")
plt.ylabel("Density")
plt.show()
```




B2 - Diabetes: Percentage Split
0.0
87.0%
13.0%
1.0

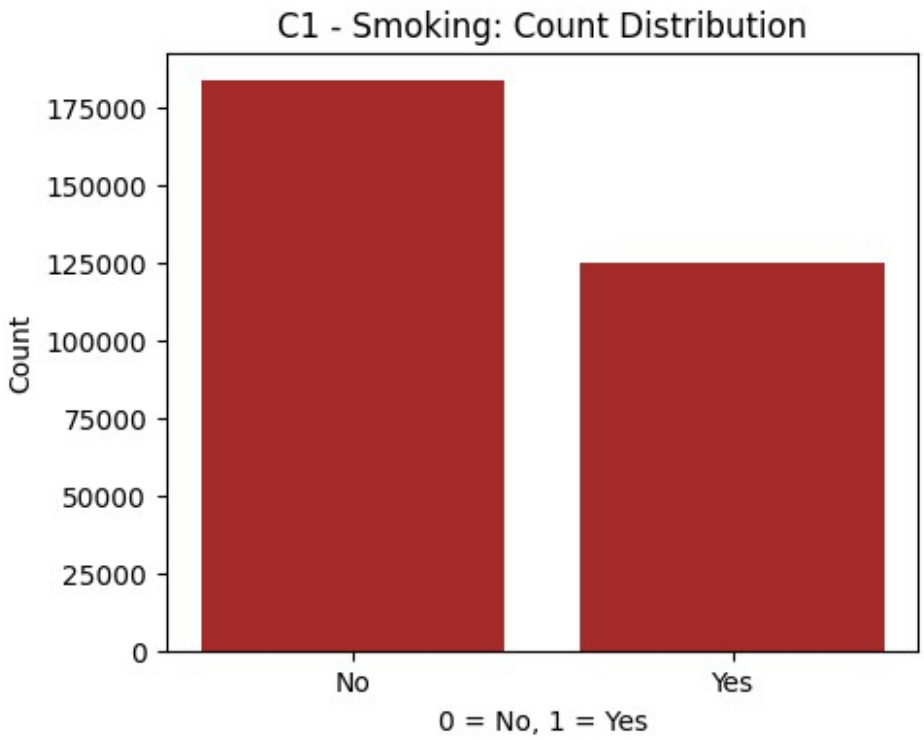

B3 - Diabetes by Sex
Diabetes
0.0
1.0
Count
140000
120000
100000
80000
60000
40000
20000
0
Female
Male
Sex

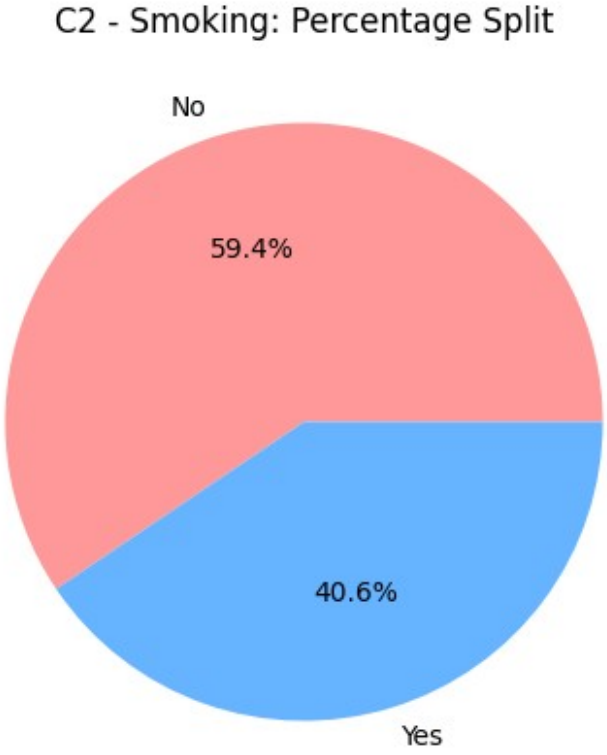


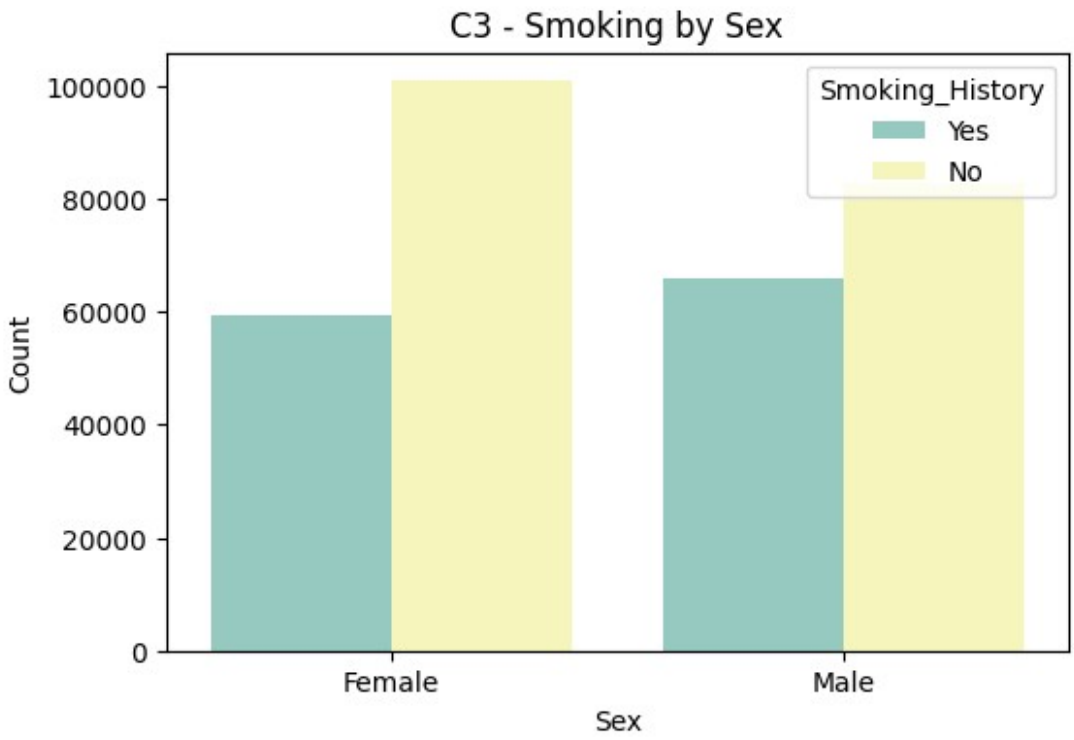


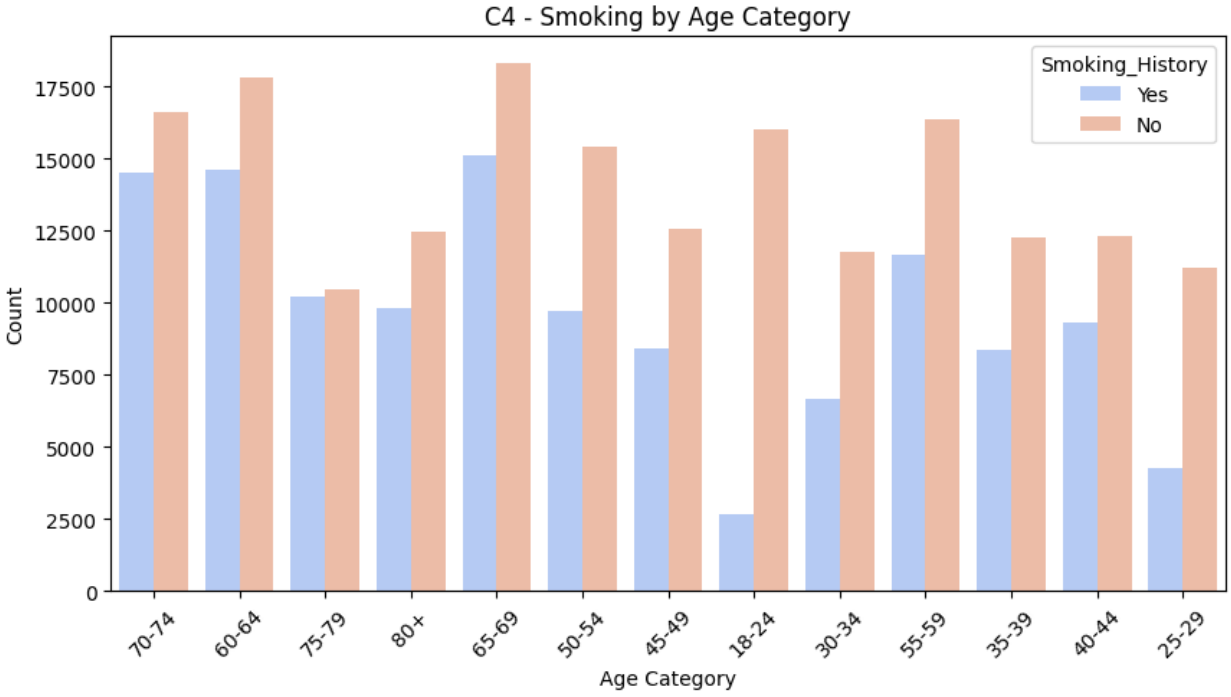

```
# ========================
# C - SMOKING
# ========================

# C1 - Bar Chart
```

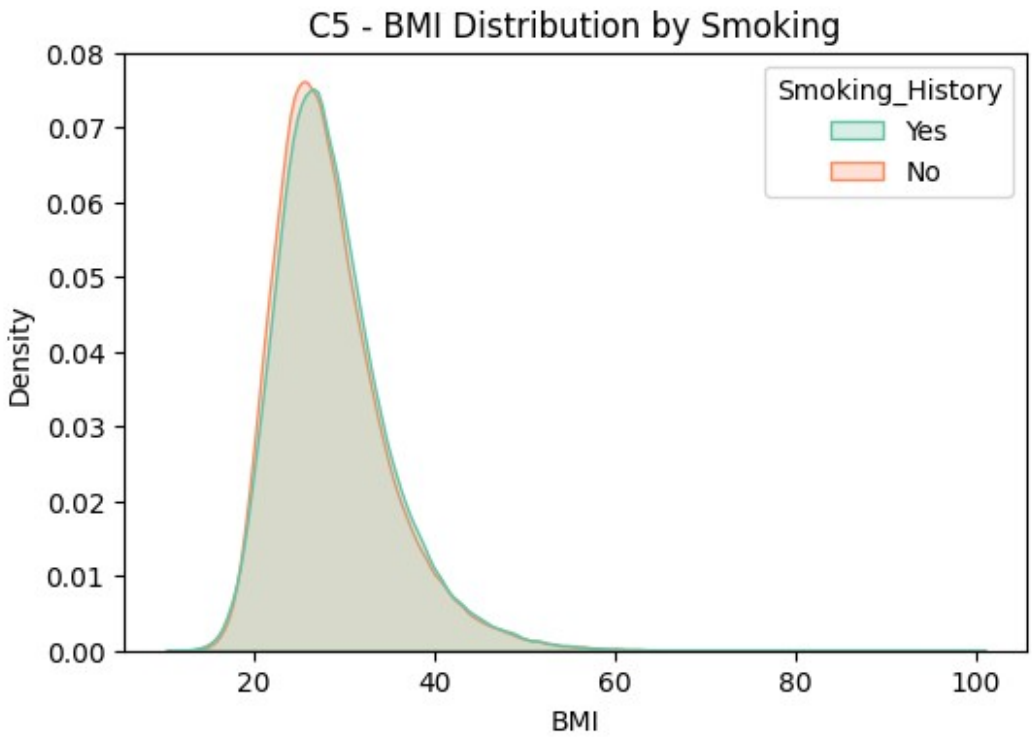
```
fig_c1 = plt.figure(figsize=(5,4))
vc = full["Smoking_History"].value_counts().sort_index()
plt.bar(vc.index.astype(str), vc.values, color="brown")
plt.title("C1 - Smoking: Count Distribution")
plt.xlabel("0 = No, 1 = Yes")
plt.ylabel("Count")
plt.show()

# C2 - Pie Chart
fig_c2 = plt.figure(figsize=(5,5))
vc = full["Smoking_History"].value_counts().sort_index()
plt.pie(vc.values, labels=vc.index.astype(str), autopct="%1.1f%%",
colors=["#ff9999","#66b3ff"])
plt.title("C2 - Smoking: Percentage Split")
plt.show()

# C3 - By Sex
fig_c3 = plt.figure(figsize=(6,4))
sns.countplot(data=full, x="Sex", hue="Smoking_History",
palette="Set3")
plt.title("C3 - Smoking by Sex")
plt.xlabel("Sex")
plt.ylabel("Count")
plt.show()

# C4 - By Age Category
fig_c4 = plt.figure(figsize=(10,5))
sns.countplot(data=full, x="Age_Category", hue="Smoking_History",
palette="coolwarm")
plt.title("C4 - Smoking by Age Category")
plt.xticks(rotation=45)
plt.xlabel("Age Category")
plt.ylabel("Count")
plt.show()

# C5 - BMI Distribution
fig_c5 = plt.figure(figsize=(6,4))
sns.kdeplot(data=full, x="BMI", hue="Smoking_History", fill=True,
common_norm=False, palette="Set2")
plt.title("C5 - BMI Distribution by Smoking")
plt.xlabel("BMI")
plt.ylabel("Density")
plt.show()
```


C1 - Smoking: Count Distribution
175000
150000
125000
100000
75000
50000
25000
0
Count
No
Yes
0 = No, 1 = Yes


C2 - Smoking: Percentage Split
No
59.4%
40.6%
Yes

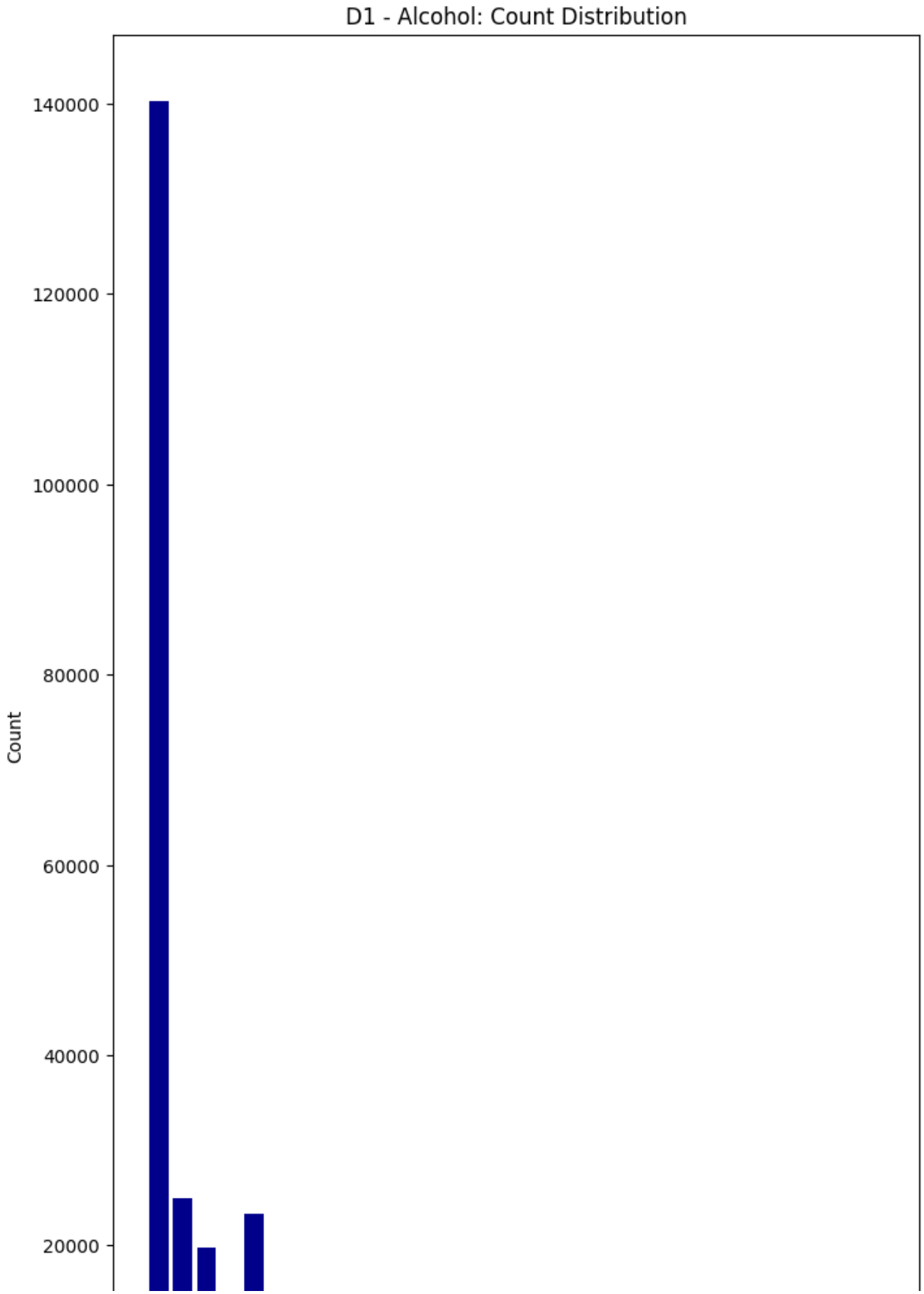

C3 - Smoking by Sex
Smoking_History
Yes
No
100000
80000
60000
40000
20000
0
Count
Female
Male
Sex

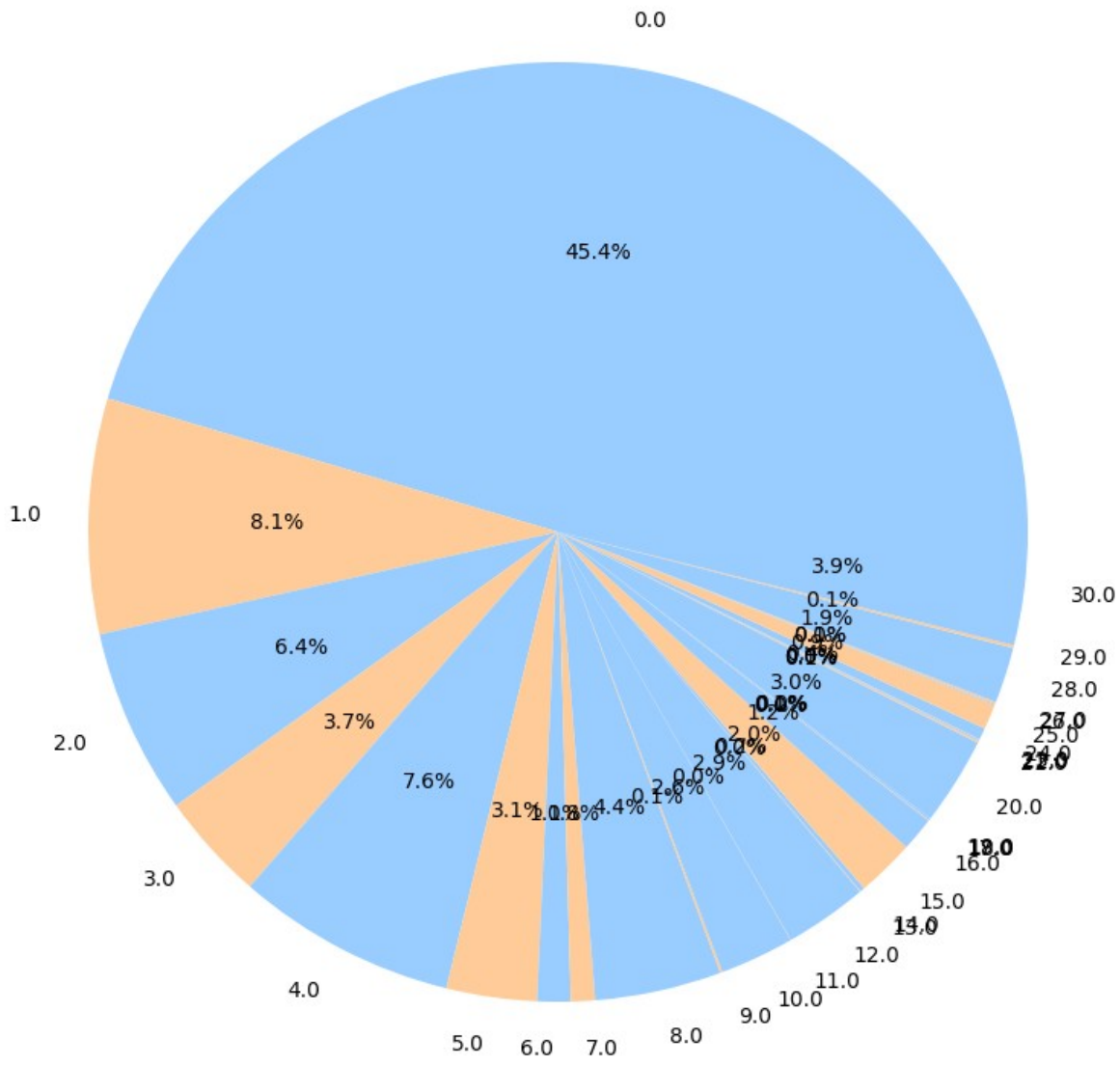

C4 - Smoking by Age Category
Smoking_History
Yes
No
17500
15000
12500
10000
7500
5000
2500
0
Count
70-74
60-64
75-79
80+
65-69
50-54
45-49
18-24
30-34
55-59
35-39
40-44
25-29
Age Category

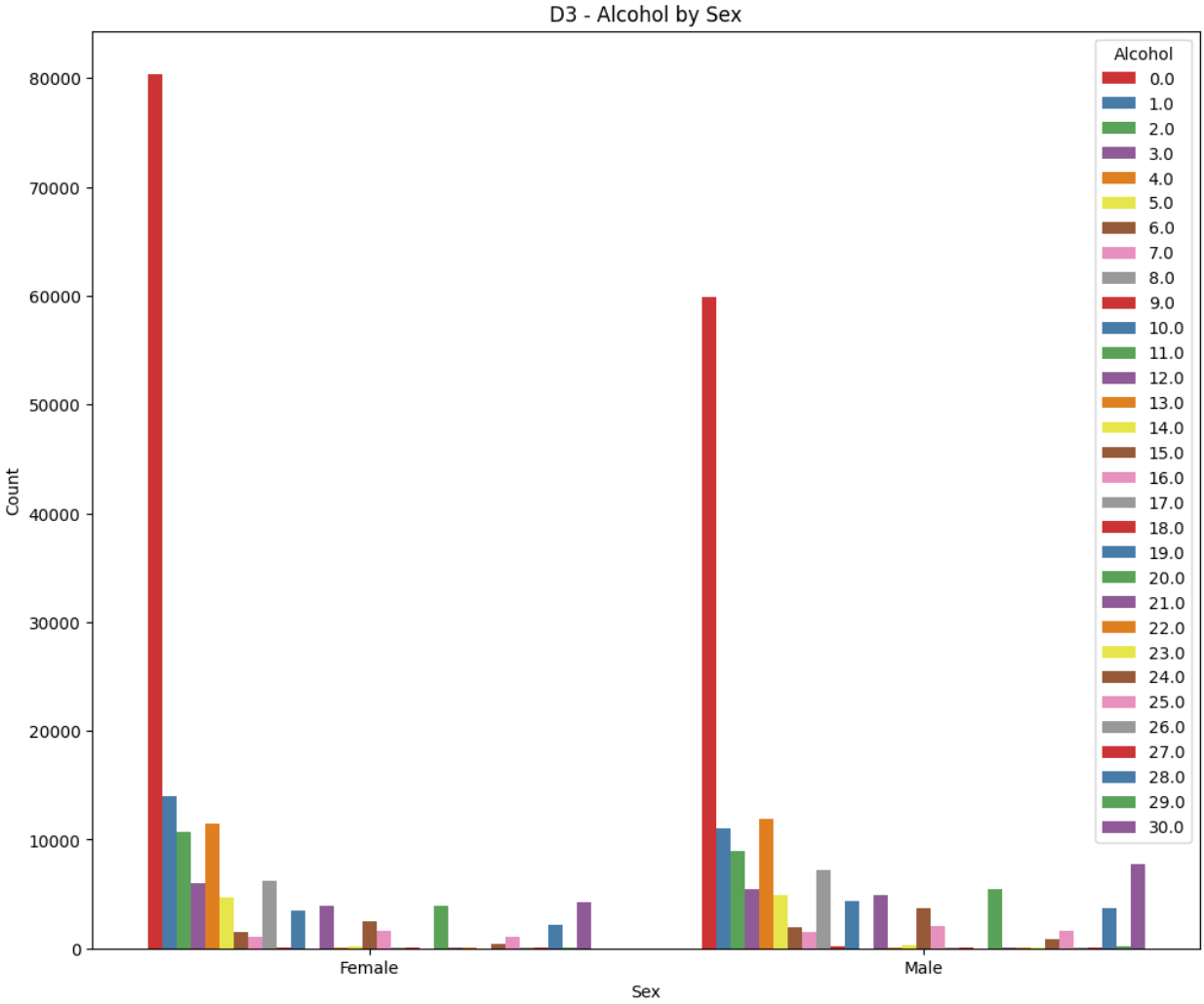


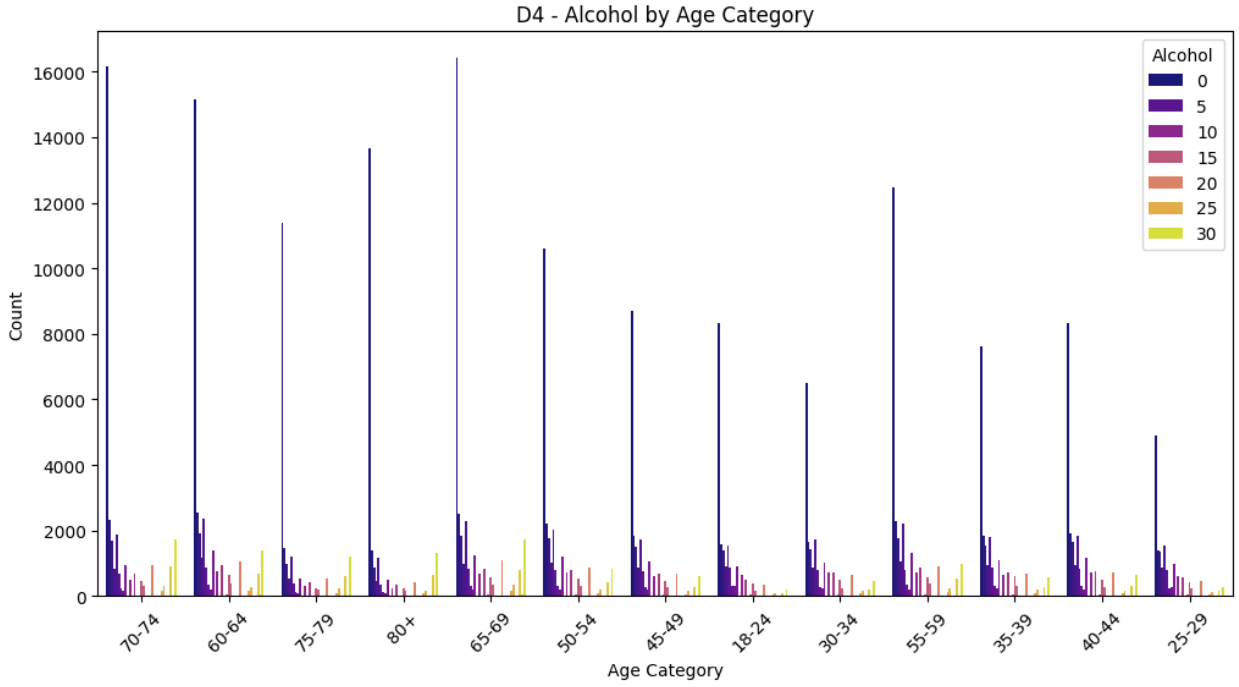

```
# ========================
# D - ALCOHOL (Resized D3 & D5)
# ========================

# D1 - Bar Chart
fig_d1 = plt.figure(figsize=(8,14))
vc = full["Alcohol_Consumption"].value_counts().sort_index()
plt.bar(vc.index.astype(str), vc.values, color="darkblue")
plt.title("D1 - Alcohol: Count Distribution")
plt.xlabel("0 = No, 1 = Yes")
plt.ylabel("Count")
plt.show()

# D2 - Pie Chart
fig_d2 = plt.figure(figsize=(10,10))
vc = full["Alcohol_Consumption"].value_counts().sort_index()
plt.pie(vc.values, labels=vc.index.astype(str), autopct="%1.1f%%",
colors=["#99ccff","#ffcc99"])
plt.title("D2 - Alcohol: Percentage Split")
plt.show()

# D3 - By Sex (Bigger)
fig_d3 = plt.figure(figsize=(12,10))  # Resize for better readability
sns.countplot(data=full, x="Sex", hue="Alcohol_Consumption",
palette="Set1")
plt.title("D3 - Alcohol by Sex")
plt.xlabel("Sex")
```

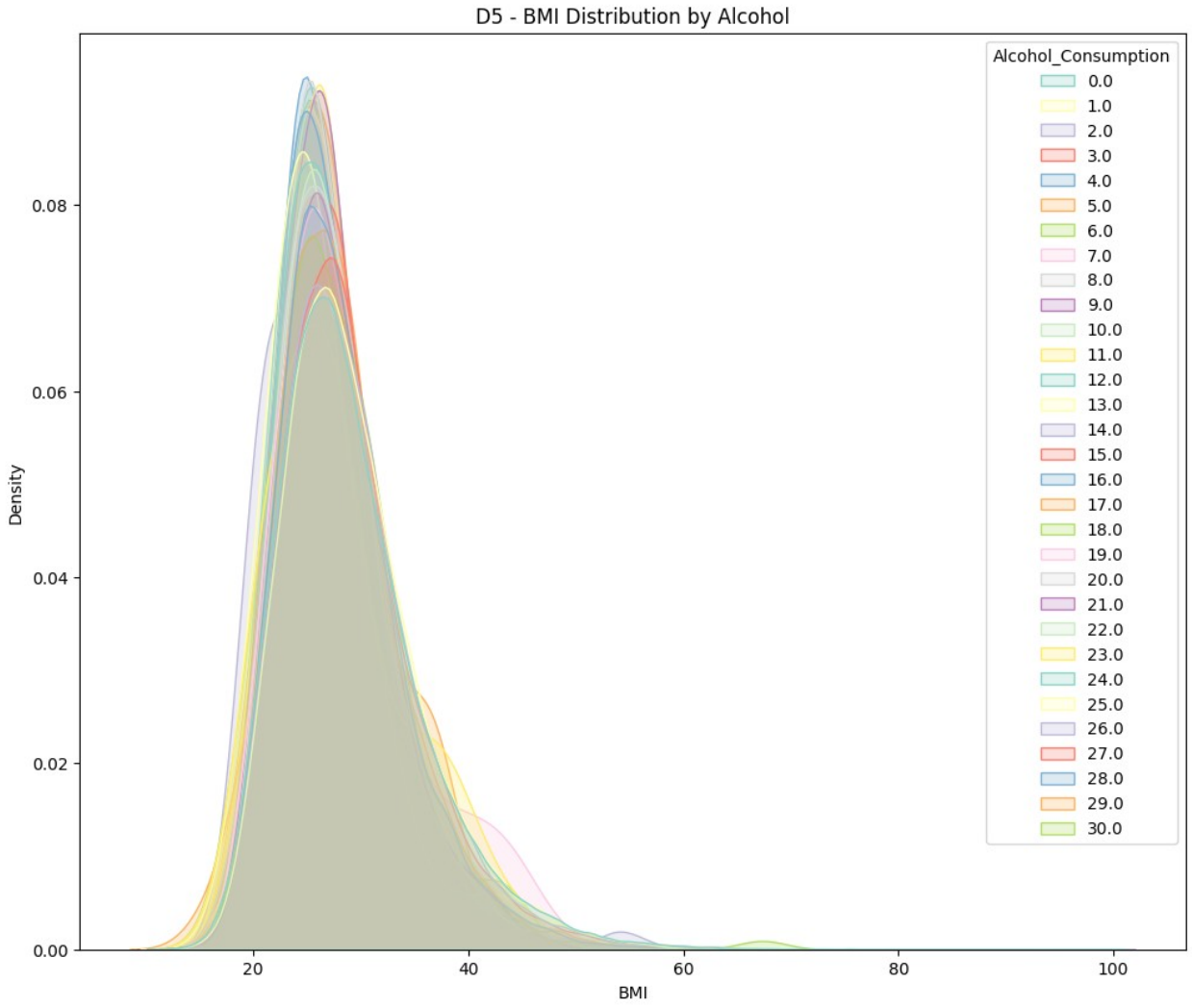
```
plt.ylabel("Count")
plt.legend(title="Alcohol")
plt.show()

# D4 - By Age Category
fig_d4 = plt.figure(figsize=(12,6))
sns.countplot(data=full, x="Age_Category", hue="Alcohol_Consumption",
palette="plasma")
plt.title("D4 - Alcohol by Age Category")
plt.xticks(rotation=45)
plt.xlabel("Age Category")
plt.ylabel("Count")
plt.legend(title="Alcohol")
plt.show()

# D5 - BMI Distribution (Bigger)
fig_d5 = plt.figure(figsize=(12,10))  # Resize for better readability
sns.kdeplot(data=full, x="BMI", hue="Alcohol_Consumption", fill=True,
common_norm=False, palette="Set3")
plt.title("D5 - BMI Distribution by Alcohol")
plt.xlabel("BMI")
plt.ylabel("Density")
plt.show()
```


D1 - Alcohol: Count Distribution
Count
140000
120000
100000
80000
60000
40000
20000


D2 - Alcohol: Percentage Split
0.0
45.4%
1.0
8.1%
2.0
6.4%
3.0
3.7%
4.0
7.6%
5.0
3.1%
6.0
7.0
4.4%
8.0
9.0
10.0
11.0
12.0
15.0
16.0
20.0
25.0
28.0
29.0
30.0
3.9%
0.1%
1.9%
3.0%
2.0%
2.9%
0.1%


D3 - Alcohol by Sex
Count
Sex
Female
Male
Alcohol
0.0
1.0
2.0
3.0
4.0
5.0
6.0
7.0
8.0
9.0
10.0
11.0
12.0
13.0
14.0
15.0
16.0
17.0
18.0
19.0
20.0
21.0
22.0
23.0
24.0
25.0
26.0
27.0
28.0
29.0
30.0


D4 - Alcohol by Age Category
Alcohol
0
5
10
15
20
25
30
Count
16000
14000
12000
10000
8000
6000
4000
2000
0
70-74
60-64
75-79
80+
65-69
50-54
45-49
18-24
30-34
55-59
35-39
40-44
25-29
Age Category




```
full.to_csv("processed_data.csv", index=False)


print(" EDA Complete")
print(f"Total rows: {len(full)}")
print(f"Columns: {full.columns.tolist()}")
print("Missing values per column:")
print(full.isnull().sum())
```

```
 EDA Complete
Total rows: 308854
Columns: ['General_Health', 'Checkup', 'Exercise', 'Heart_Disease',
'Skin_Cancer', 'Other_Cancer', 'Depression', 'Diabetes', 'Arthritis',
'Sex', 'Age_Category', 'Height_(cm)', 'Weight_(kg)', 'BMI',
'Smoking_History', 'Alcohol_Consumption', 'Fruit_Consumption',
'Green_Vegetables_Consumption', 'FriedPotato_Consumption']
Missing values per column:
```

```
General_Health                  0
Checkup                         0
Exercise                        0
Heart_Disease                   0
Skin_Cancer                     0
Other_Cancer                    0
Depression                      0
Diabetes                        0
Arthritis                       0
Sex                             0
Age_Category                    0
Height_(cm)                     0
Weight_(kg)                     0
BMI                             0
Smoking_History                 0
Alcohol_Consumption             0
Fruit_Consumption               0
Green_Vegetables_Consumption    0
FriedPotato_Consumption         0
dtype: int64
```