In [92]:
```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
```

In [106…
```
import pandas as pd
import matplotlib.pyplot as plt

# Load the dataset
file_path = 'C:\\Users\\melha\\Downloads\\updated-hosital-data.csv'
data = pd.read_csv(file_path)

# Print the first few rows of the dataset to verify loading
print(data.head())
```

```
                  date   patient_id patient_gender  patient_age  \
0  2020-03-20 08:47:01  145-39-5406              M           69
1  2020-06-15 11:29:36  316-34-3057              M            4
2  2020-06-20 09:13:13  897-46-3852              F           56
3  2020-02-04 22:34:29  358-31-9711              F           24
4  2020-09-04 17:48:27  289-26-0537              M            5

   patient_sat_score patient_first_inital patient_last_name  \
0               10.0                   H         Glasspool
1                NaN                   X           Methuen
2                9.0                   P         Schubuser
3                8.0                   U          Titcombe
4                NaN                   Y       Gionettitti

                patient_race  patient_admin_flag  patient_waittime  \
0                      White               False                39
1  Native American/Alaska Native               True                27
2           African American               True                55
3  Native American/Alaska Native               True                31
4           African American              False                10

  department_referral  Surgery Duration
0                 NaN               132
1                 NaN               122
2    General Practice                44
3    General Practice               136
4        Orthopedics               101
```

In [112…
```
# Remove missing values
df_clean = df.dropna()

# Remove duplicates
df_clean = df_clean.drop_duplicates()

# Verify if cleaning is successful
df_clean.info()  # Check if there are any missing values left
```

```
<class 'pandas.core.frame.DataFrame'>
Index: 1077 entries, 2 to 9206
Data columns (total 12 columns):
 #   Column               Non-Null Count  Dtype
---  ------               --------------  -----
 0   date                 1077 non-null   object
 1   patient_id           1077 non-null   object
 2   patient_gender       1077 non-null   object
 3   patient_age          1077 non-null   int64
 4   patient_sat_score    1077 non-null   float64
 5   patient_first_inital  1077 non-null  object
 6   patient_last_name    1077 non-null   object
 7   patient_race         1077 non-null   object
 8   patient_admin_flag   1077 non-null   bool
 9   patient_waittime     1077 non-null   int64
 10  department_referral  1077 non-null   object
 11  Surgery Duration     1077 non-null   int64
dtypes: bool(1), float64(1), int64(3), object(7)
memory usage: 102.0+ KB
```

In [108…
```
# Remove duplicates
data.drop_duplicates(inplace=True)

# Remove rows with missing values
data.dropna(inplace=True)
```

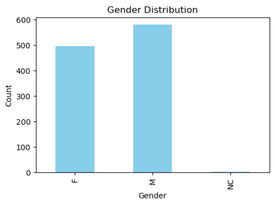
In [110…
```
# Create age categories
age_bins = [0, 18, 35, 50, 65, 80, 100]
age_labels = ['0-18', '19-35', '36-50', '51-65', '66-80', '81-100']
data['age_category'] = pd.cut(data['patient_age'], bins=age_bins, labels=age_labels)

# Group by gender
gender_group = data.groupby('patient_gender').size()

# Group by race
race_group = data.groupby('patient_race').size()

# Group by age category
age_group = data.groupby('age_category').size()

# Print the grouped data to verify
print("Gender Group:")
print(gender_group)
print("\nRace Group:")
print(race_group)
print("\nAge Group:")
print(age_group)
```

```
Gender Group:
patient_gender
F     495
M     580
NC      2
dtype: int64

Race Group:
patient_race
African American                227
Asian                           119
Declined to Identify            118
Native American/Alaska Native    62
Pacific Islander                 54
Two or More Races               178
White                           319
dtype: int64

Age Group:
age_category
0-18      244
19-35     237
36-50     210
51-65     204
66-80     182
81-100      0
dtype: int64
```

```
C:\Users\melha\AppData\Local\Temp\ipykernel_8072\3287377504.py:13: FutureWarning: The default of observed=False is deprecated and will be changed to True in a future version of pandas. Pass observed=False to retain current behavior or observed=True to adopt the future default and silence this warning.
  age_group = data.groupby('age_category').size()
```

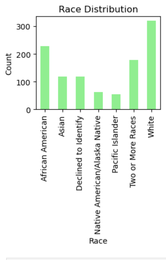
In [102…
```
plt.figure(figsize=(12, 8))

# Gender distribution bar chart
plt.subplot(2, 2, 1)
gender_group.plot(kind='bar', color='skyblue')
plt.title('Gender Distribution')
plt.xlabel('Gender')
plt.ylabel('Count')
```

Out[102…
```
Text(0, 0.5, 'Count')
```



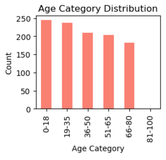
In [74]:
```
# Race distribution bar chart
plt.subplot(2, 2, 2)
race_group.plot(kind='bar', color='lightgreen')
plt.title('Race Distribution')
plt.xlabel('Race')
plt.ylabel('Count')
```

Out[74]:
```
Text(0, 0.5, 'Count')
```

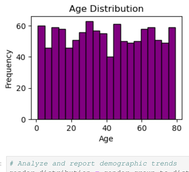


In [76]:
```
# Age category distribution bar chart
plt.subplot(2, 2, 3)
age_group.plot(kind='bar', color='salmon')
plt.title('Age Category Distribution')
plt.xlabel('Age Category')
plt.ylabel('Count')
```

Out[76]:
```
Text(0, 0.5, 'Count')
```



In [78]:
```
# Age histogram
plt.subplot(2, 2, 4)
data['patient_age'].plot(kind='hist', bins=20, color='purple', edgecolor='black')
plt.title('Age Distribution')
plt.xlabel('Age')
plt.ylabel('Frequency')

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



In [80]:
```
# Analyze and report demographic trends
gender_distribution = gender_group.to_dict()
race_distribution = race_group.to_dict()
age_distribution = age_group.to_dict()
average_age = data['patient_age'].mean()

print("Demographic Trends Analysis:")
print(f"Gender Distribution: {gender_distribution}")
print(f"Race Distribution: {race_distribution}")
print(f"Age Distribution: {age_distribution}")
print(f"Average Patient Age: {average_age:.2f} years")
```

```
Demographic Trends Analysis:
Gender Distribution: {'F': 495, 'M': 580, 'NC': 2}
```

Race Distribution: {'African American': 227, 'Asian': 119, 'Declined to Identify': 118, 'Native American/Alaska Native': 62, 'Pacific Islander': 54, 'Two or More Races': 178, 'White': 319}
Age Distribution: {'0-18': 244, '19-35': 237, '36-50': 210, '51-65': 204, '66-80': 182, '81-100': 0}
Average Patient Age: 39.81 years