task1

December 1, 2024

[57]:

```
# Importing Libraries
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
```

[59]:

```
# Step 1: Load Dataset
# Load the dataset
data = pd.read_csv('task1.csv')

# Display the first few rows of the dataset
print("Dataset Preview:")
display(data.head())
```

```
Dataset Preview:

   Age  Gender Income_Level Education_Level Employment_Status Marital_Status  \
0   56  Female       Medium        Bachelor          Employed         Single
1   69  Female         High     High School     Self-Employed         Single
2   46  Female       Medium     High School          Employed         Single
3   32    Male       Medium          Master        Unemployed         Single
4   60  Female          Low        Bachelor          Employed         Single

   Number_of_Children Housing_Type  Monthly_Expenditure Health_Condition  \
0                   1       Rented                 3219        Excellent
1                   0        Owned                 4008             Good
2                   3        Owned                 4241             Good
3                   1        Owned                 2074             Good
4                   0        Owned                 4498             Good

  Favorite_Hobby
0        Reading
1          Music
2      Traveling
3         Gaming
4         Gaming
```

[61]:

```
# Step 2: Summary Statistics and Initial Insights
print("\nSummary Statistics:")
display(data.describe(include='all'))
```

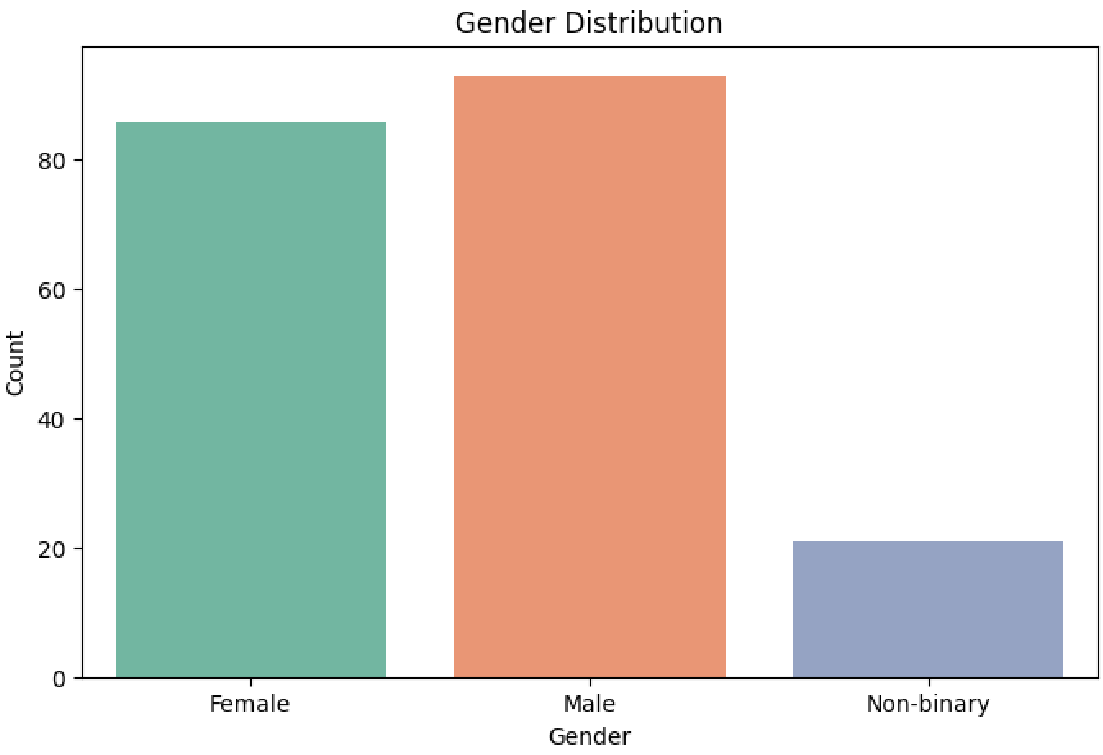
```python
print("\nDataset Information:")
data.info()
```

```
Summary Statistics:
               Age Gender Income_Level Education_Level Employment_Status  \
count   200.000000    200          200             200               200
unique         NaN      3            3               4                 3
top            NaN   Male       Medium     High School          Employed
freq           NaN     93           92              78               114
mean     49.590000    NaN          NaN             NaN               NaN
std      18.982189    NaN          NaN             NaN               NaN
min      18.000000    NaN          NaN             NaN               NaN
25%      32.000000    NaN          NaN             NaN               NaN
50%      50.000000    NaN          NaN             NaN               NaN
75%      65.250000    NaN          NaN             NaN               NaN
max      80.000000    NaN          NaN             NaN               NaN

       Marital_Status  Number_of_Children Housing_Type  Monthly_Expenditure  \
count             200          200.000000          200           200.000000
unique              4                 NaN            3                  NaN
top            Single                 NaN        Owned                  NaN
freq               89                 NaN           92                  NaN
mean              NaN            1.875000          NaN          2640.890000
std               NaN            1.445622          NaN          1309.149326
min               NaN            0.000000          NaN           501.000000
25%               NaN            1.000000          NaN          1414.250000
50%               NaN            2.000000          NaN          2574.000000
75%               NaN            3.000000          NaN          3862.000000
max               NaN            4.000000          NaN          4973.000000

       Health_Condition Favorite_Hobby
count               200            200
unique                4              5
top                Good      Traveling
freq                101             59
mean                NaN            NaN
std                 NaN            NaN
min                 NaN            NaN
25%                 NaN            NaN
50%                 NaN            NaN
75%                 NaN            NaN
max                 NaN            NaN

Dataset Information:
<class 'pandas.core.frame.DataFrame'>
```

```
RangeIndex: 200 entries, 0 to 199
Data columns (total 11 columns):
 #   Column              Non-Null Count  Dtype 
---  ------              --------------  ----- 
 0   Age                 200 non-null    int64 
 1   Gender              200 non-null    object
 2   Income_Level        200 non-null    object
 3   Education_Level     200 non-null    object
 4   Employment_Status   200 non-null    object
 5   Marital_Status      200 non-null    object
 6   Number_of_Children  200 non-null    int64 
 7   Housing_Type        200 non-null    object
 8   Monthly_Expenditure  200 non-null    int64 
 9   Health_Condition    200 non-null    object
 10  Favorite_Hobby      200 non-null    object
dtypes: int64(3), object(8)
memory usage: 17.3+ KB
```

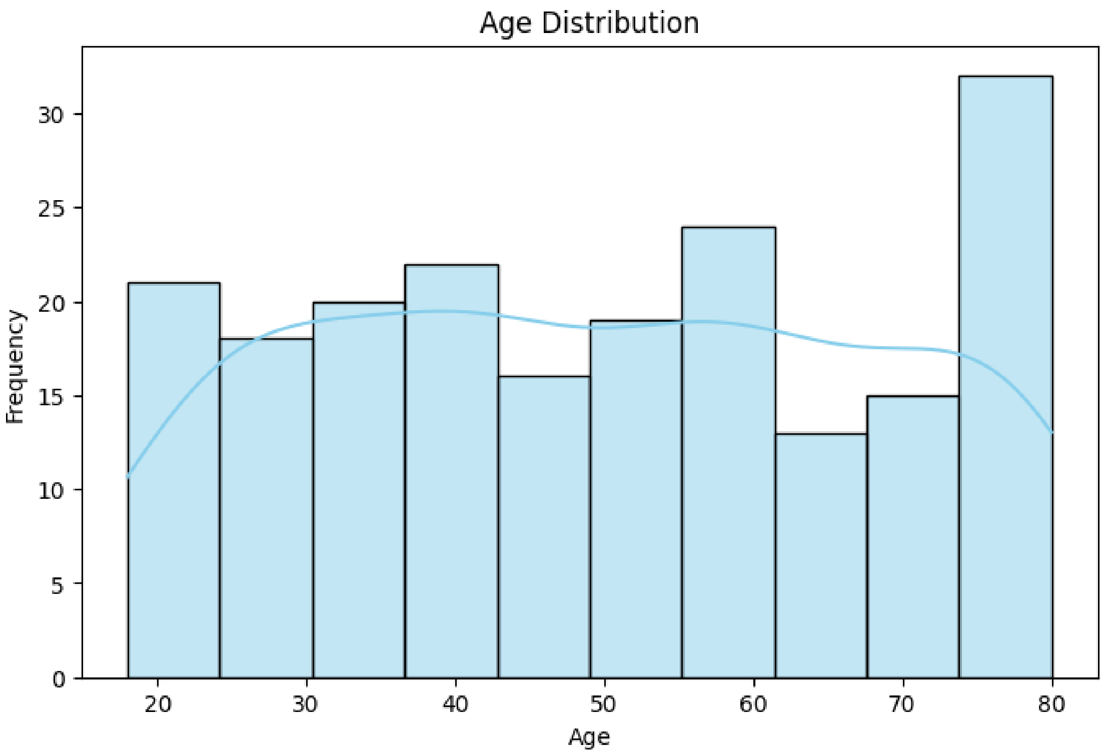
[68]:

```
# Step 3: Bar Chart for Categorical Variable (Gender Distribution)
plt.figure(figsize=(8, 5))
sns.countplot(data=data, x='Gender', hue='Gender', dodge=False, palette='Set2',␣
  ↪legend=False)
plt.title('Gender Distribution')
plt.xlabel('Gender')
plt.ylabel('Count')
plt.show()

# Insight: Gender distribution shows whether the data is balanced or skewed␣
  ↪across categories.
```



0.1 Conclusion: Gender Distribution: A fairly balanced distribution across Male, Female, and Non-binary.

[70]:
```
# Step 4: Histogram for Continuous Variable (Age Distribution)
plt.figure(figsize=(8, 5))
sns.histplot(data=data, x='Age', bins=10, kde=True, color='skyblue')
plt.title('Age Distribution')
plt.xlabel('Age')
plt.ylabel('Frequency')
plt.show()

# Insight: Age distribution highlights the age group concentration in the␣
  ↪population.
```



0.2 Conclusion: Age Distribution: Most individuals are concentrated in the middle-age group (30–50 years).

```
[ ]:
```