In [3]:
```
# Import necessary libraries
import pandas as pd
import numpy as np
```

In [13]:
```
# Load the dataset
file_path = "C:\\MISC\\alzheimers_prediction_dataset.csv"
df = pd.read_csv(file_path)
```

In [15]:
```
# Display first few rows of the dataset
print("First 5 rows of the dataset:")
display(df.head())
```

```
First 5 rows of the dataset:
```

	Country	Age	Gender	Education Level	BMI	Physical Activity Level	Smoking Status	Alcohol Consumption	Diabetes	Hy
0	Spain	90	Male	1	33.0	Medium	Never	Occasionally	No	
1	Argentina	72	Male	7	29.9	Medium	Former	Never	No	
2	South Africa	86	Female	19	22.9	High	Current	Occasionally	No	
3	China	53	Male	17	31.2	Low	Never	Regularly	Yes	
4	Sweden	58	Female	3	30.0	High	Former	Never	Yes	

5 rows × 25 columns

In [17]:
```
# Display column names
print("\nColumn Names in Dataset:")
print(df.columns)
```

```
Column Names in Dataset:
Index(['Country', 'Age', 'Gender', 'Education Level', 'BMI',
       'Physical Activity Level', 'Smoking Status', 'Alcohol Consumption',
       'Diabetes', 'Hypertension', 'Cholesterol Level',
       'Family History of Alzheimer's', 'Cognitive Test Score',
       'Depression Level', 'Sleep Quality', 'Dietary Habits',
       'Air Pollution Exposure', 'Employment Status', 'Marital Status',
       'Genetic Risk Factor (APOE-ε4 allele)', 'Social Engagement Level',
       'Income Level', 'Stress Levels', 'Urban vs Rural Living',
       'Alzheimer's Diagnosis'],
      dtype='object')
```

In [19]:
```
# Check for missing values
print("\nMissing Values Per Column:")
print(df.isnull().sum())
```

```
Missing Values Per Column:
Country                                0
Age                                    0
Gender                                 0
Education Level                        0
BMI                                    0
Physical Activity Level                0
Smoking Status                         0
Alcohol Consumption                    0
Diabetes                               0
Hypertension                           0
Cholesterol Level                      0
Family History of Alzheimer’s          0
Cognitive Test Score                   0
Depression Level                       0
Sleep Quality                          0
Dietary Habits                         0
Air Pollution Exposure                 0
Employment Status                      0
Marital Status                         0
Genetic Risk Factor (APOE-ε4 allele)   0
Social Engagement Level                0
Income Level                           0
Stress Levels                          0
Urban vs Rural Living                  0
Alzheimer’s Diagnosis                  0
dtype: int64
```

In [33]:
```python
# Standardize column names: Convert to lowercase, replace spaces, and remove "'"
df.columns = df.columns.str.strip().str.lower().str.replace(" ", "_").str.replace("

print(df.columns)  # Print modified column names to verify changes

# Now, update column selection with new names
selected_columns = ["education_level", "gender", "alzheimers_diagnosis"]  # Adjust
df_subset = df[selected_columns].copy()
```

```
Index(['country', 'age', 'gender', 'education_level', 'bmi',
       'physical_activity_level', 'smoking_status', 'alcohol_consumption',
       'diabetes', 'hypertension', 'cholesterol_level',
       'family_history_of_alzheimers', 'cognitive_test_score',
       'depression_level', 'sleep_quality', 'dietary_habits',
       'air_pollution_exposure', 'employment_status', 'marital_status',
       'genetic_risk_factor_(apoe-ε4_allele)', 'social_engagement_level',
       'income_level', 'stress_levels', 'urban_vs_rural_living',
       'alzheimers_diagnosis'],
      dtype='object')
```

In [41]:
```python
# Rename columns for consistency
df_subset.columns = ["education", "gender", "alzheimers_diagnosis"]
```

In [43]:
```python
print(df_subset.columns)
```

```
Index(['education', 'gender', 'alzheimers_diagnosis'], dtype='object')
```

In [45]:
```python
# Check unique values for categorical variables
print("\nUnique values in 'education' column:")
```

```
print(df_subset["education"].unique())

print("\nUnique values in 'gender' column:")
print(df_subset["gender"].unique())

print("\nUnique values in 'alzheimers_diagnosis' column:")
print(df_subset["alzheimers_diagnosis"].unique())
```

```
Unique values in 'education' column:
[ 1  7 19 17  3  2 18 11 15 10  6 13 12  4 16  5 14  0  8  9]

Unique values in 'gender' column:
['Male' 'Female']

Unique values in 'alzheimers_diagnosis' column:
['No' 'Yes']
```

In [47]:
```
# Handle missing values (Simple approach: drop rows with missing data)
df_cleaned = df_subset.dropna()
```

In [49]:
```
# Display summary statistics
print("\nSummary statistics of cleaned dataset:")
display(df_cleaned.describe(include="all"))
```

```
Summary statistics of cleaned dataset:
```

	education	gender	alzheimers_diagnosis
count	74283.000000	74283	74283
unique	NaN	2	2
top	NaN	Female	No
freq	NaN	37249	43570
mean	9.487514	NaN	NaN
std	5.757020	NaN	NaN
min	0.000000	NaN	NaN
25%	4.000000	NaN	NaN
50%	9.000000	NaN	NaN
75%	14.000000	NaN	NaN
max	19.000000	NaN	NaN

Save the cleaned dataset (optional)

df_cleaned.to_csv("cleaned_alzheimers_data.csv", index=False)

In [75]:
```
df_subset.to_csv(r"C:\\Users\\rdarn\Documents\\cleaned_alzheimers_data.csv", index=
```

In [ ]: