```python
In [17]: import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
from sklearn.feature_selection import chi2
from sklearn.preprocessing import MinMaxScaler
import pandas as pd
```

```python
In [5]: import seaborn as sns
```

```python
In [6]: # Load the dataset
df = pd.read_csv('depression_data.csv')
df.head()
```

Out[6]:

	Name	Age	Marital Status	Education Level	Number of Children	Smoking Status	Physical Activity Level	Employment Status	Income	Alcohol Consumption	Dietary Habits	Sleep Patterns	History of Mental Illness	History of Substance Abuse	Family History of Depression	Chronic Medical Conditions
0	Christine Barker	31	Married	Bachelor's Degree	2	Non-smoker	Active	Unemployed	26265.67	Moderate	Moderate	Fair	Yes	No	Yes	Yes
1	Jacqueline Lewis	55	Married	High School	1	Non-smoker	Sedentary	Employed	42710.36	High	Unhealthy	Fair	Yes	No	No	Yes
2	Shannon Church	78	Widowed	Master's Degree	1	Non-smoker	Sedentary	Employed	125332.79	Low	Unhealthy	Good	No	No	Yes	No
3	Charles Jordan	58	Divorced	Master's Degree	3	Non-smoker	Moderate	Unemployed	9992.78	Moderate	Moderate	Poor	No	No	No	No
4	Michael Rich	18	Single	High School	0	Non-smoker	Sedentary	Unemployed	8595.08	Low	Moderate	Fair	Yes	No	Yes	Yes

```python
In [7]: # Information of the dataset
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 413768 entries, 0 to 413767
Data columns (total 16 columns):
 #   Column                        Non-Null Count   Dtype
---  ------                        --------------   -----
 0   Name                          413768 non-null  object
 1   Age                           413768 non-null  int64
 2   Marital Status                413768 non-null  object
 3   Education Level               413768 non-null  object
 4   Number of Children            413768 non-null  int64
 5   Smoking Status                413768 non-null  object
 6   Physical Activity Level       413768 non-null  object
 7   Employment Status             413768 non-null  object
 8   Income                        413768 non-null  float64
 9   Alcohol Consumption           413768 non-null  object
 10  Dietary Habits                413768 non-null  object
 11  Sleep Patterns                413768 non-null  object
 12  History of Mental Illness     413768 non-null  object
 13  History of Substance Abuse    413768 non-null  object
 14  Family History of Depression  413768 non-null  object
 15  Chronic Medical Conditions    413768 non-null  object
dtypes: float64(1), int64(2), object(13)
memory usage: 50.5+ MB
```

```python
In [8]: #Describes the stats of the dataset
df.describe()
```

Out[8]:

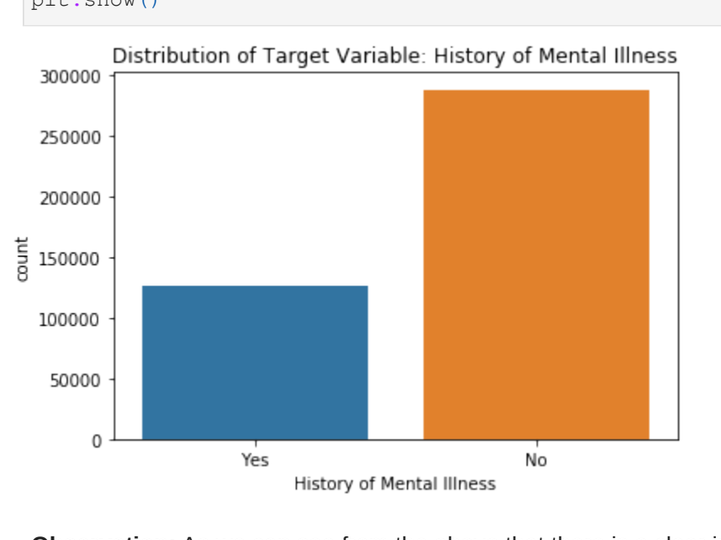
	Age	Number of Children	Income
count	413768.000000	413768.000000	413768.000000
mean	49.000713	1.298972	50661.707971
std	18.158759	1.237054	40624.100565
min	18.000000	0.000000	0.410000
25%	33.000000	0.000000	21001.030000
50%	49.000000	1.000000	37520.135000
75%	65.000000	2.000000	76616.300000
max	80.000000	4.000000	209995.220000

```python
In [9]: #Checks for missing values
df.isnull().sum()
```

```
Out[9]: Name                            0
Age                             0
Marital Status                  0
Education Level                 0
Number of Children              0
Smoking Status                  0
Physical Activity Level         0
Employment Status               0
Income                          0
Alcohol Consumption             0
Dietary Habits                  0
Sleep Patterns                  0
History of Mental Illness       0
History of Substance Abuse      0
Family History of Depression    0
Chronic Medical Conditions      0
dtype: int64
```

The following code shows the Distribution of Target Variable (History of mental Illness)

```python
In [11]: plt.figure(figsize=(6, 4))
sns.countplot(data=df, x='History of Mental Illness')
plt.title('Distribution of Target Variable: History of Mental Illness')
plt.show()
```

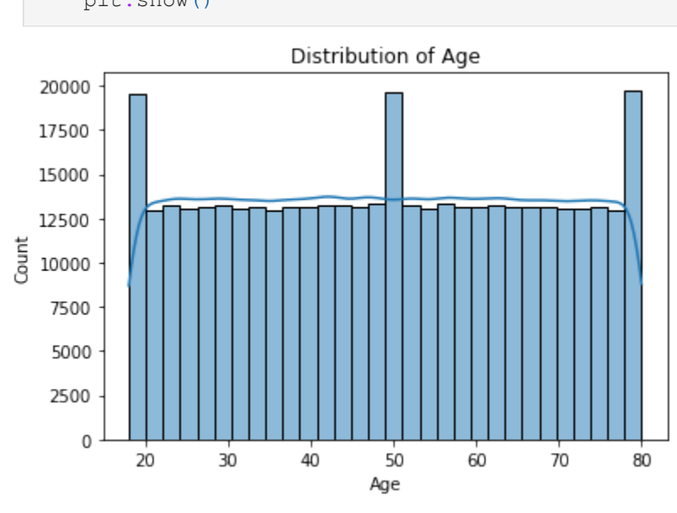


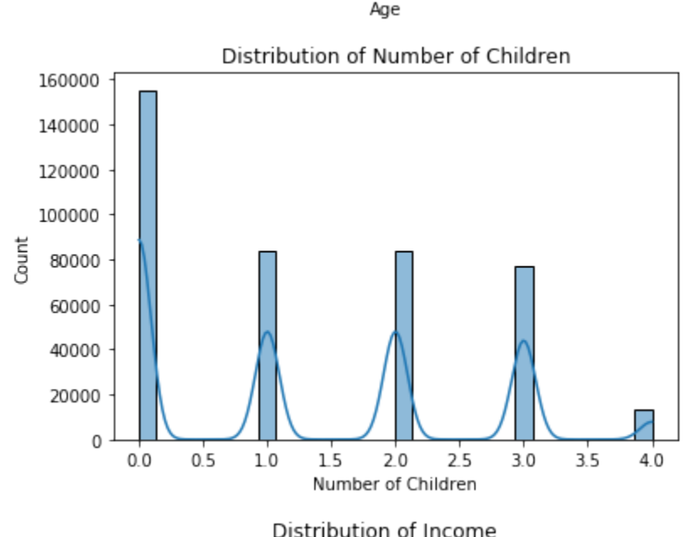
Observation: As we can see from the above that there is a clear imbalance with individuals with no history of mental Illness compared to do those who do. There is approximately 70% data of individuals with no history of mental Illness

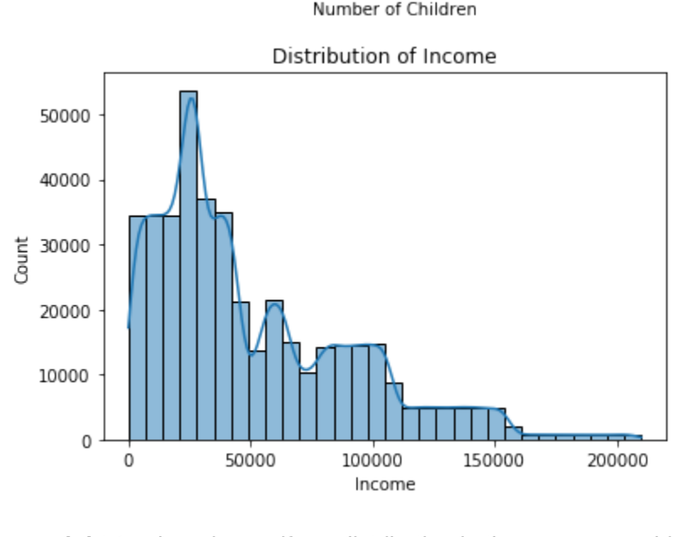
Implication: This suggest that we could use SMOTE resampling techniques for more accurate predictions

Distribution of Numerical values

```python
In [10]: numerical_cols = ['Age', 'Number of Children', 'Income']

for col in numerical_cols:
    plt.figure(figsize=(6, 4))
    sns.histplot(df[col], bins=30, kde=True)
    plt.title(f'Distribution of {col}')
    plt.show()
```







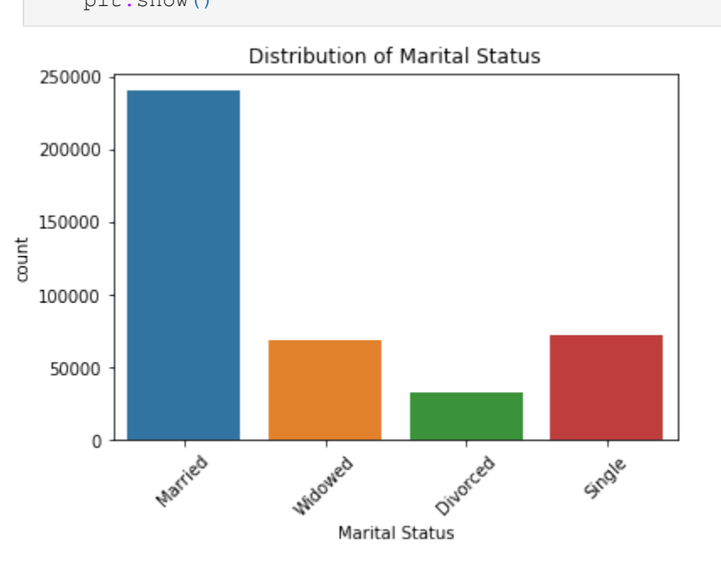
Insight-1: There is a uniform distribution in the age group which suggest that it is well balanced; The spikes indicate trends within specific age groups related to life events

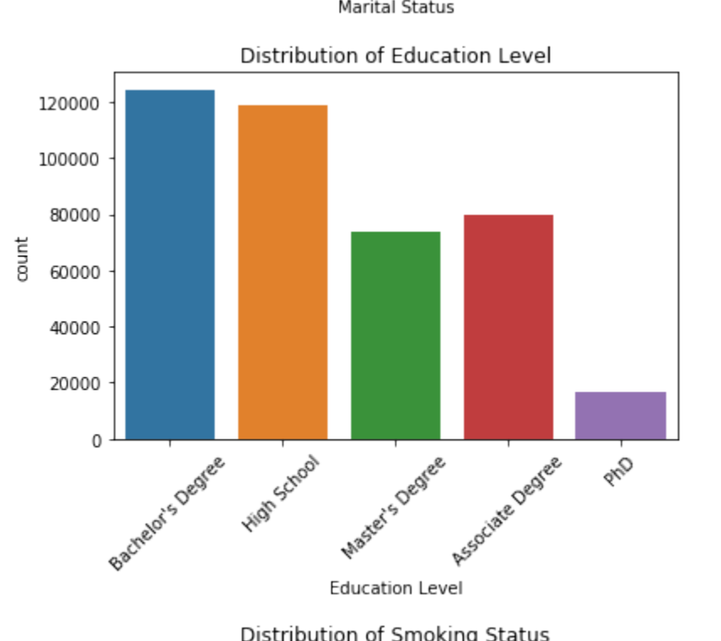
Insight-2: The right skewed distribution in income indicates that most individulas fall under low to moderate income brackets. Outliers such as individuals with very high income need to be accounted during model training using scaling or other outlier handling techniques

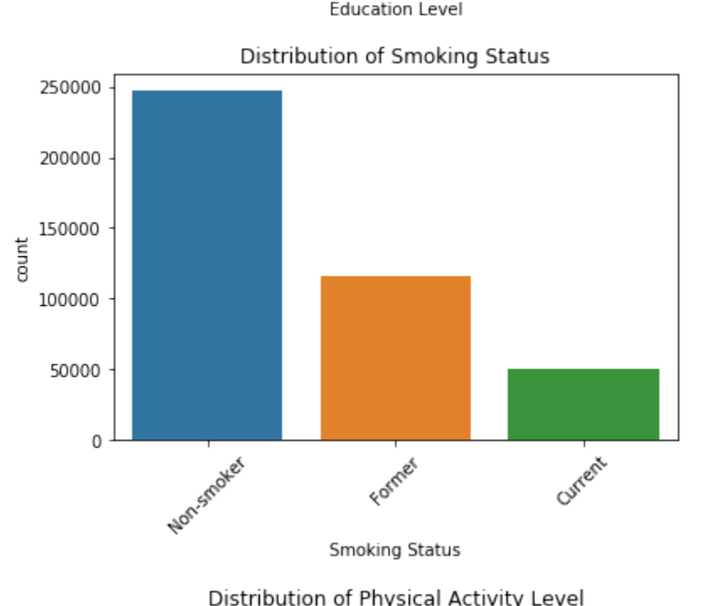
Insight-3: Most Individuals have either no or 1-2 children, also shows that individuals with large individuals are rare in this dataset

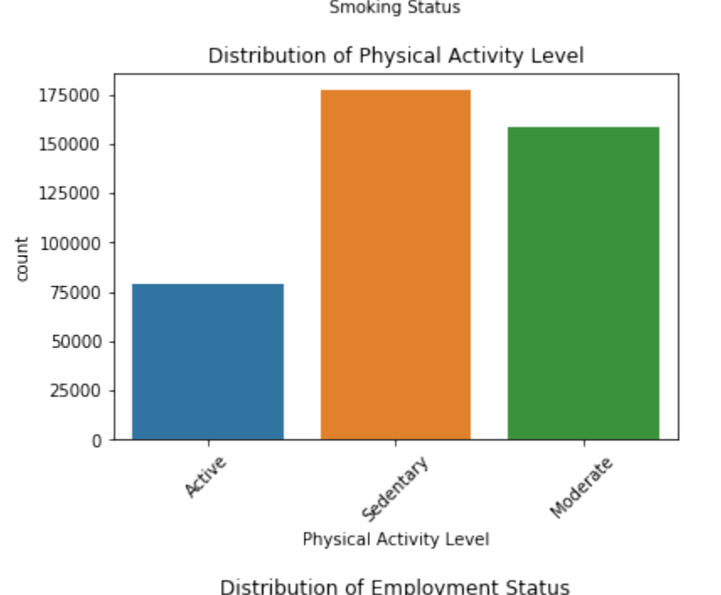
Distribution of Categorical values

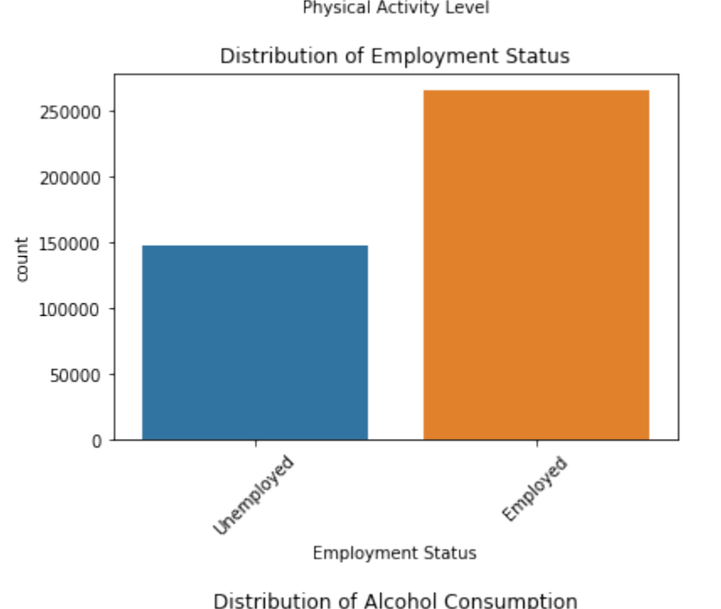
```python
In [12]: categorical_cols = ['Marital Status', 'Education Level', 'Smoking Status', 'Physical Activity Level',
                    'Employment Status', 'Alcohol Consumption', 'Dietary Habits', 'Sleep Patterns',
                    'History of Substance Abuse', 'Family History of Depression', 'Chronic Medical Conditions']

# Bar plots for categorical variables
for col in categorical_cols:
    plt.figure(figsize=(6, 4))
    sns.countplot(data=df, x=col)
    plt.title(f'Distribution of {col}')
    plt.xticks(rotation=45)
    plt.show()
```

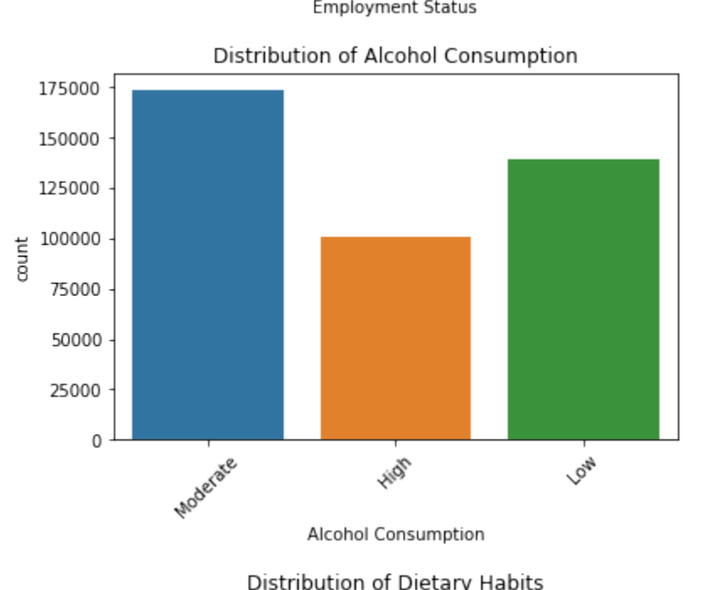


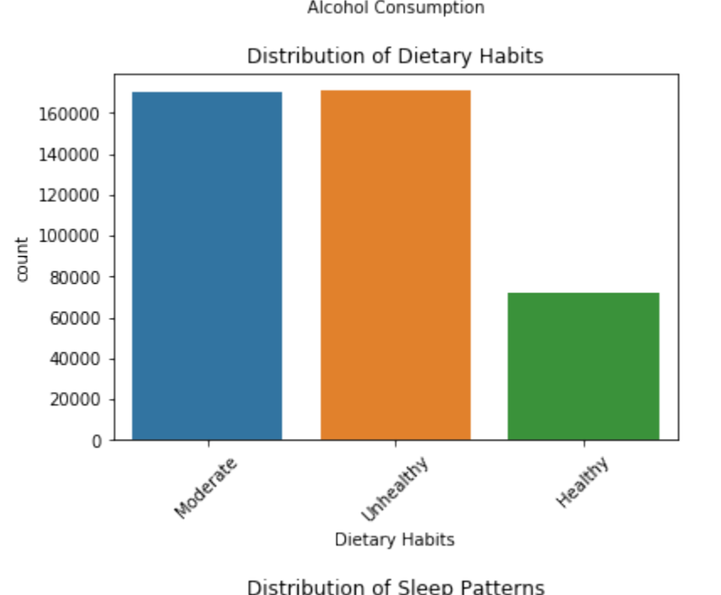


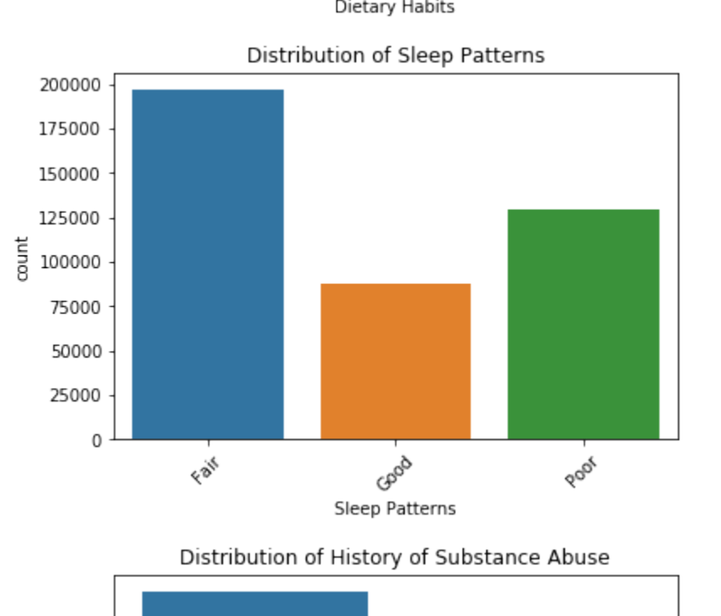


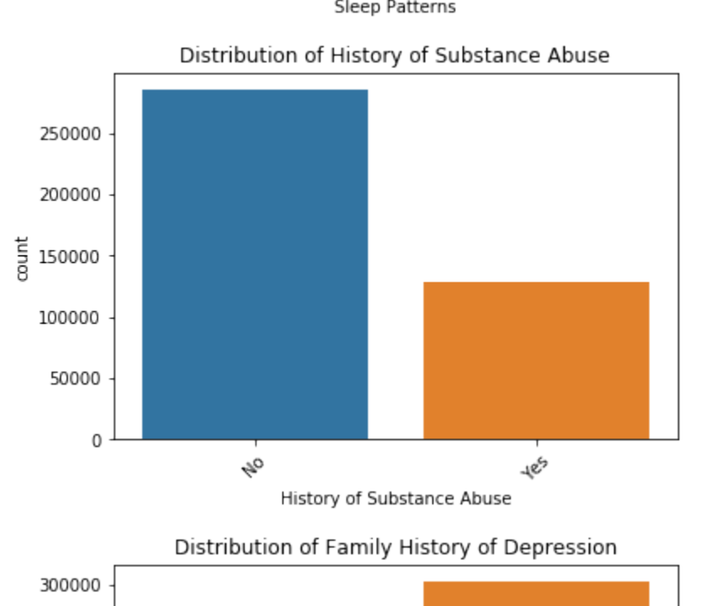


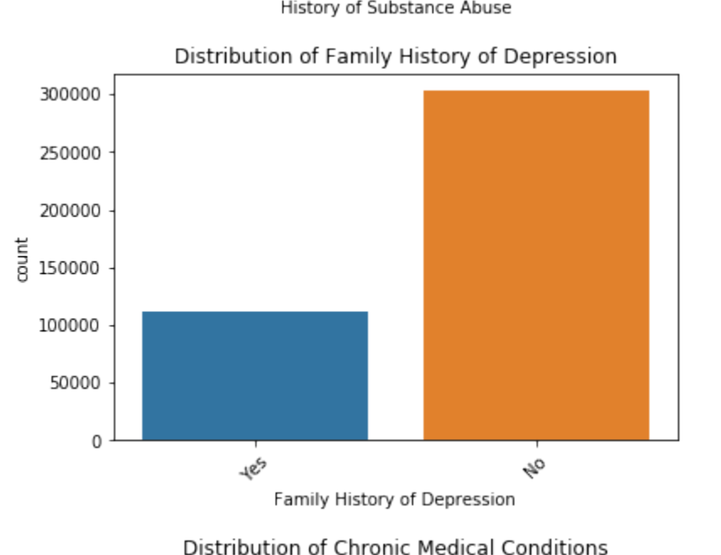


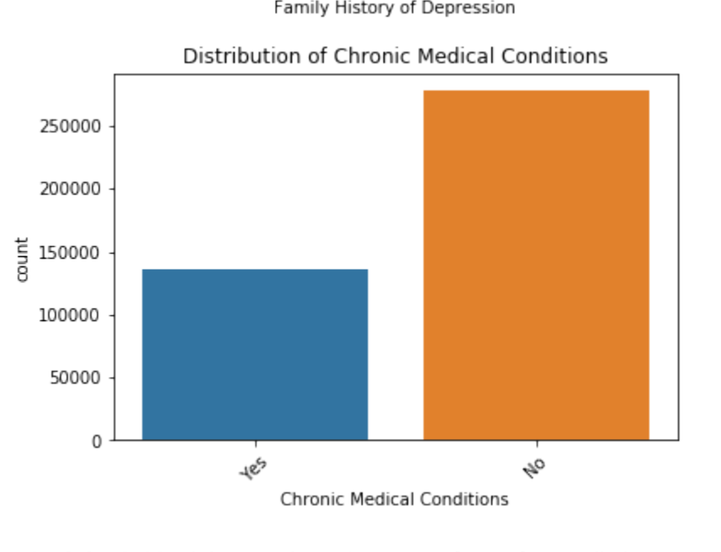












Insight-4: Martial status is an Important factor for predicting mental Illness and this dataset is dominated by Individuals who are married

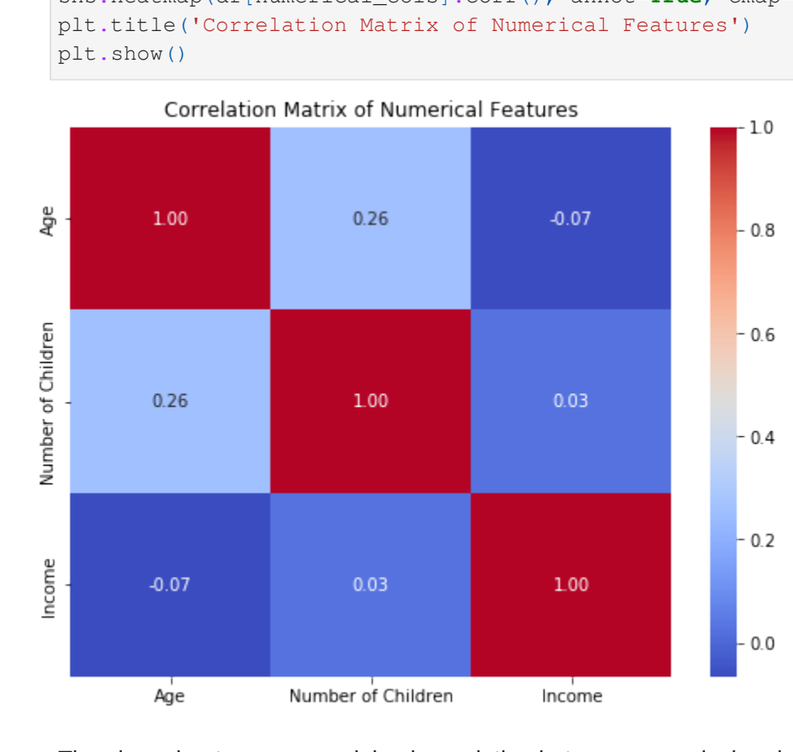
Insight-5: Education may play a vital role in predicting mental, and there is a significant portion with bachelors degree and fewer holding advanced degrees such as PHDs

Insight-6: Majority of the Individuals in this dataset are non-smokers and often smoking status is associated with mental Illness

Insight-7: Majority of Individuals in this dataset are employed and employment feature is a strong predictor of mental health and unemployement is often linked to stress and anxiety

Checking corelation of Numerical Features

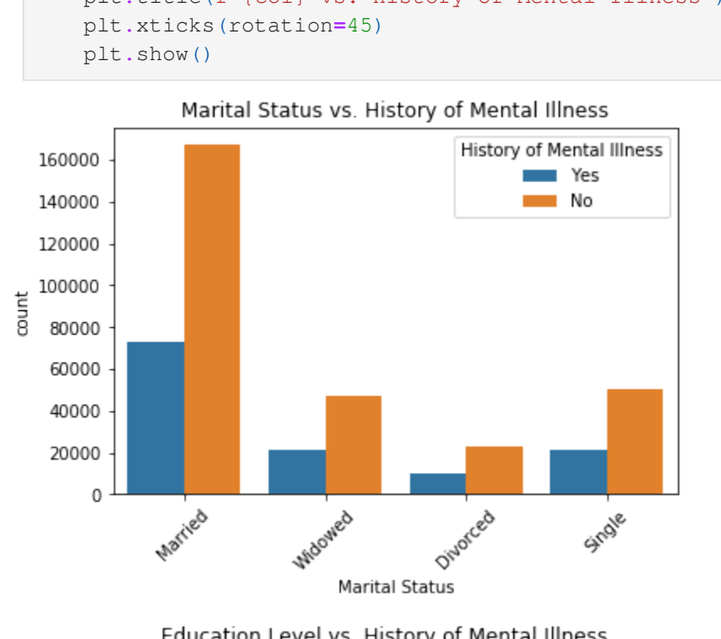
```python
In [14]: plt.figure(figsize=(8, 6))
sns.heatmap(df[numerical_cols].corr(), annot=True, cmap='coolwarm', fmt='.2f')
plt.title('Correlation Matrix of Numerical Features')
plt.show()
```

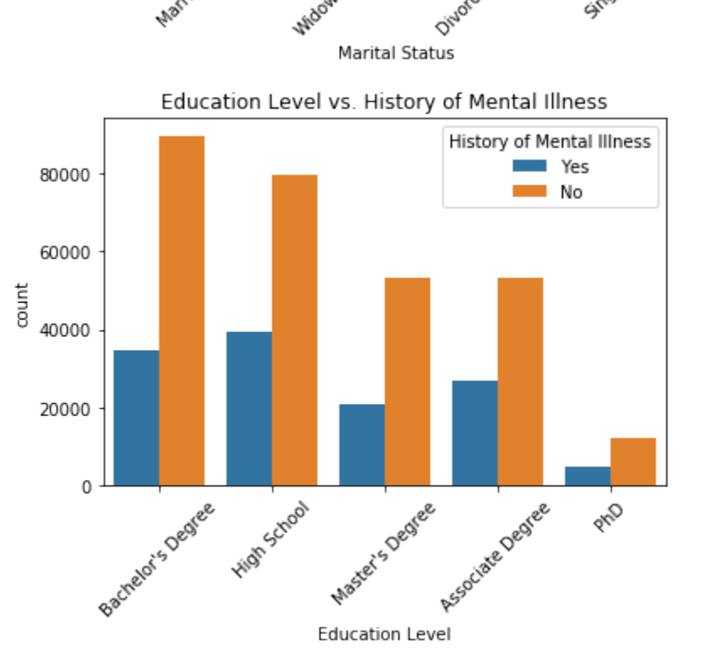


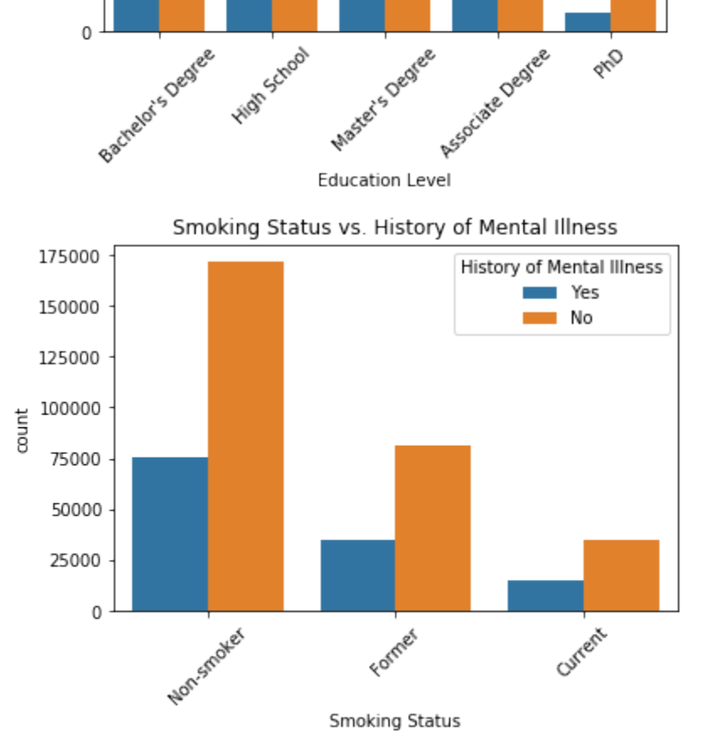
The above heatmap sows minimal correlation between numerical and categorical variables and there is no high correlation, multicollinearity should not be an issue

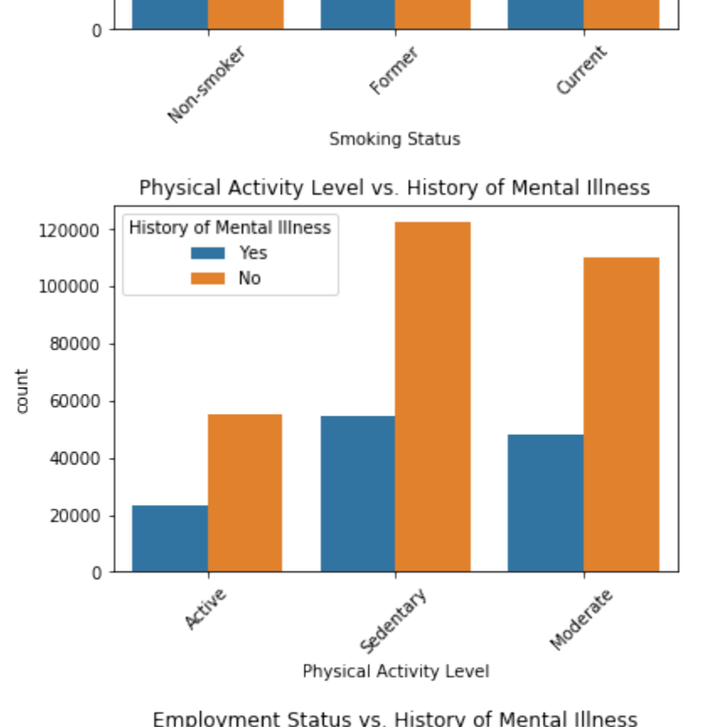
Analysing relationships between categorical variables and target variables

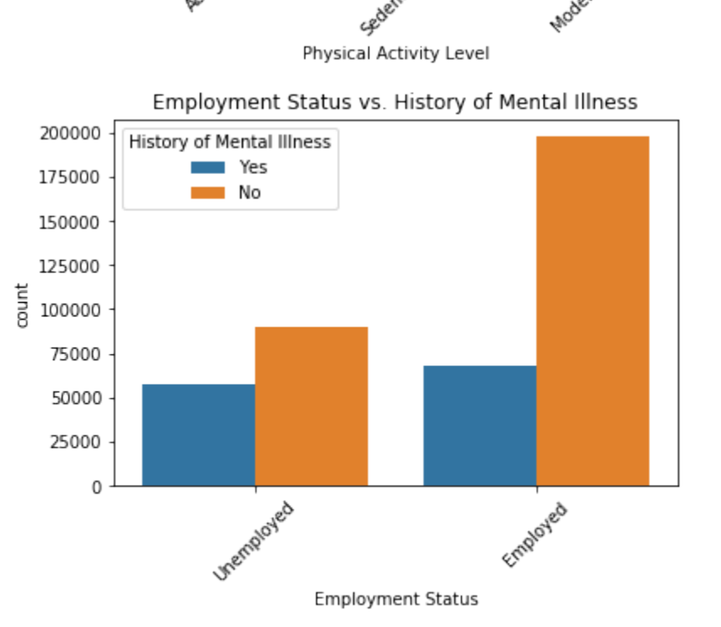
```python
In [15]: for col in categorical_cols:
    plt.figure(figsize=(6, 4))
    sns.countplot(data=df, x=col, hue='History of Mental Illness')
    plt.title(f'{col} vs. History of Mental Illness')
    plt.xticks(rotation=45)
    plt.show()
```

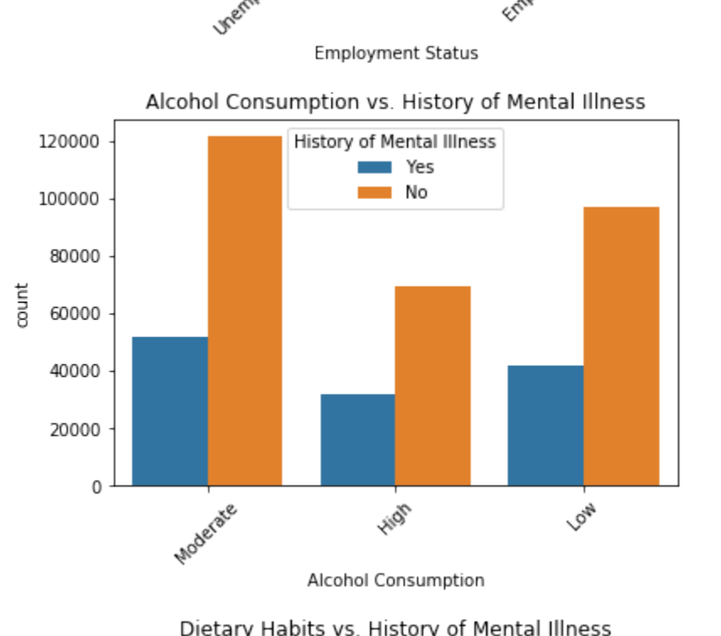


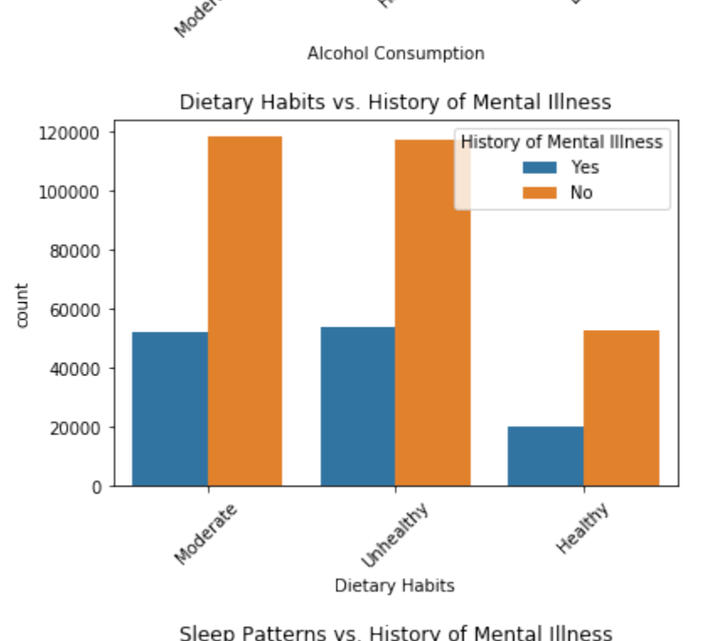


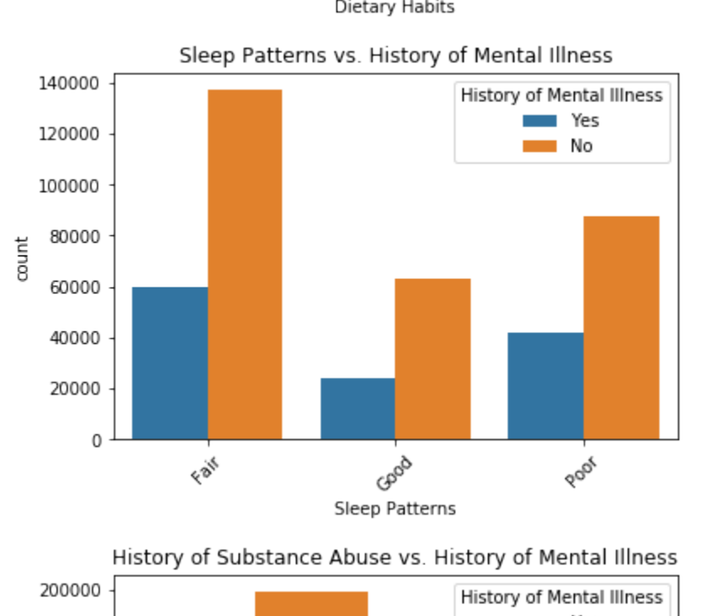


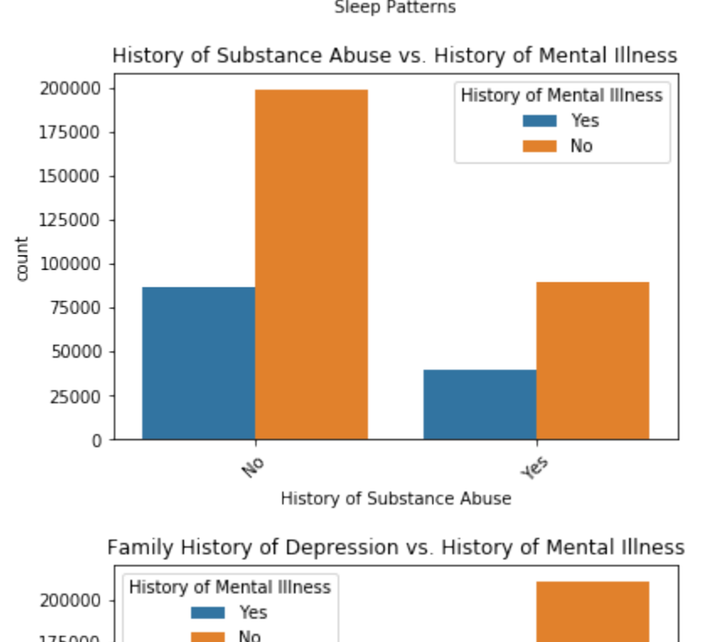


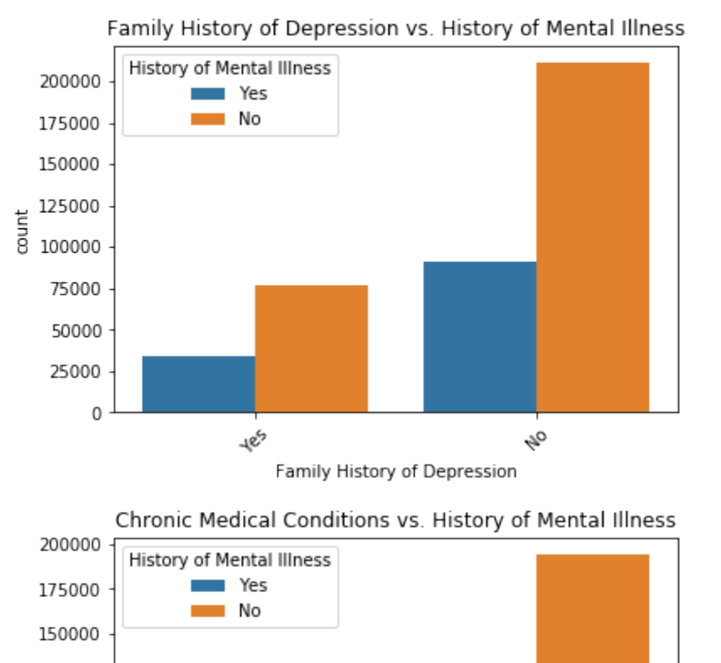


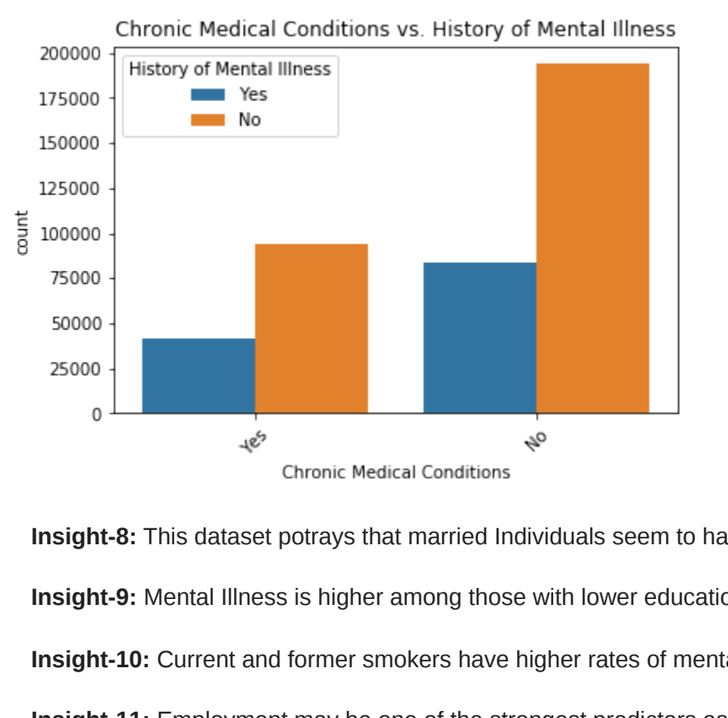












Insight-8: This dataset potrays that married Individuals seem to have lower rates of Mental health issues while single and widowed individuals show higher proportions of mental Illness

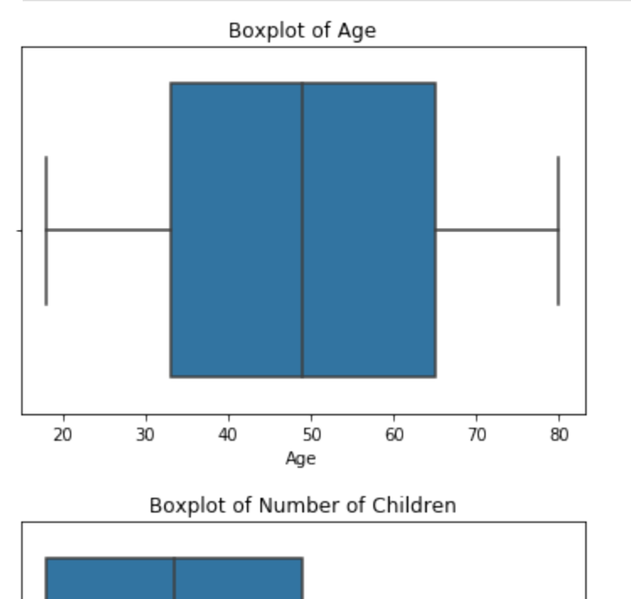
Insight-9: Mental Illness is higher among those with lower educational backgrounds while it is lower with advanced degrees

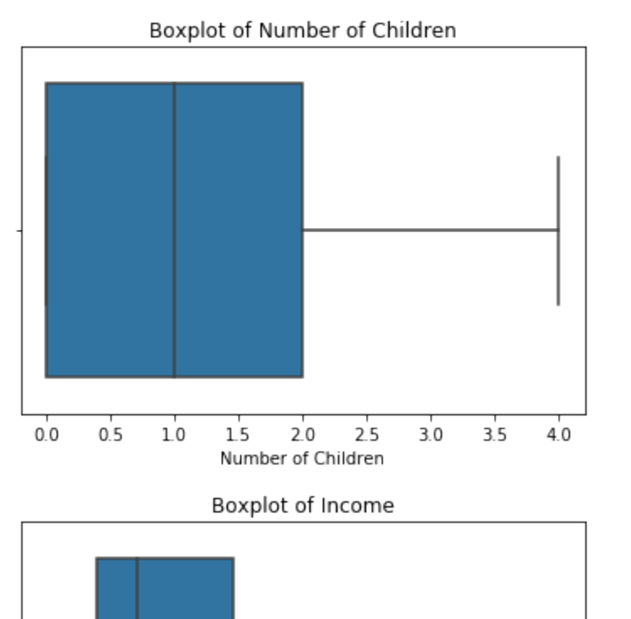
Insight-10: Current and former smokers have higher rates of mental illness compared to non-smokers.

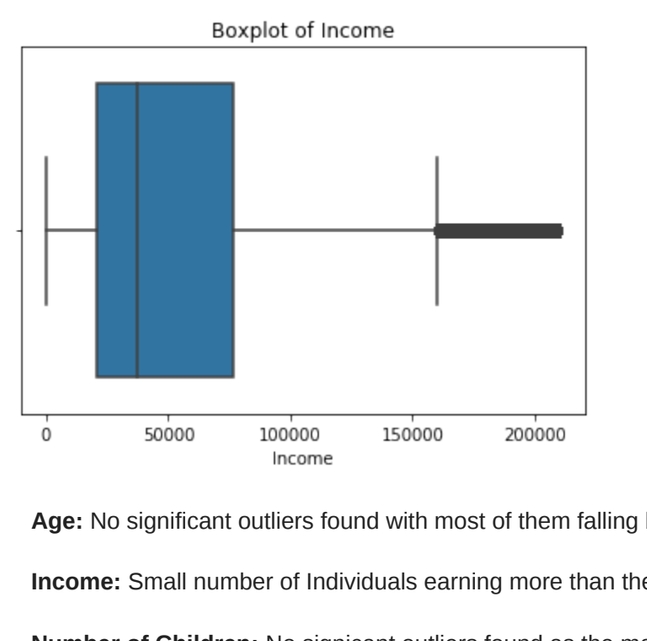
Insight-11: Employment may be one of the strongest predictors as no employmnent often leads to major stress problems

Checking for Outliers in the Numerical variables

```python
In [19]: for col in numerical_cols:
    plt.figure(figsize=(6, 4))
    sns.boxplot(data=df, x=col)
    plt.title(f'Boxplot of {col}')
    plt.show()
```







Age: No significant outliers found with most of them falling between 20 and 80

Income: Small number of Individuals earning more than the majority.These outliers may need to be removed

Number of Children: No signicant outliers found as the maximum is capped at 4

Running Chi-Square Test to understand Feature Importance

```python
In [18]: from sklearn.feature_selection import chi2
from sklearn.preprocessing import MinMaxScaler
import pandas as pd

# List of categorical columns
categorical_cols = ['Marital Status', 'Education Level', 'Smoking Status', 'Physical Activity Level',
                    'Employment Status', 'Alcohol Consumption', 'Dietary Habits', 'Sleep Patterns',
                    'History of Substance Abuse', 'Family History of Depression', 'Chronic Medical Conditions']

# Encoding the target variable
df['History of Mental Illness'] = df['History of Mental Illness'].map({'Yes': 1, 'No': 0})

# One-hot encode the categorical columns, excluding the target column
df_encoded = pd.get_dummies(df, columns=categorical_cols, drop_first=False)

# Features (X) and target (y)
X = df_encoded.drop(columns=['Name', 'History of Mental Illness'])  # Dropping irrelevant 'Name' column
y = df_encoded['History of Mental Illness']

# Normalizing data for Chi-Square test (Chi-Square test requires non-negative values)
X_minmax = MinMaxScaler().fit_transform(X)

# Perform Chi-Square test
chi_scores = chi2(X_minmax, y)

# Display p-values and feature importance
p_values = pd.Series(chi_scores[1], index=X.columns)
important_features = p_values[p_values < 0.05]  # Features significant at p < 0.05
print("Important Features based on Chi-Square test:\n", important_features)
```

```
Important Features based on Chi-Square test:
 Age                                 4.827889e-11
Income                             3.229536e-263
Marital Status_Divorced             5.765434e-05
Marital Status_Single               3.872224e-05
Marital Status_Widowed              2.375366e-08
Education Level_Associate Degree    8.923698e-73
Education Level_Bachelor's Degree   2.545691e-93
Education Level_High School         2.585653e-89
Education Level_Master's Degree     7.878033e-46
Education Level_PhD                 7.261697e-13
Smoking Status_Current              1.103899e-02
Physical Activity Level_Active      7.036633e-05
Physical Activity Level_Sedentary   2.488983e-04
Employment Status_Employed          0.000000e+00
Employment Status_Unemployed        0.000000e+00
Alcohol Consumption_High            6.473697e-14
Alcohol Consumption_Moderate        4.240635e-05
Dietary Habits_Healthy              5.917721e-62
Dietary Habits_Unhealthy            6.081420e-20
Sleep Patterns_Good                 2.675515e-76
Sleep Patterns_Poor                 3.034061e-56
Family History of Depression_Yes    7.185919e-03
Chronic Medical Conditions_No       2.585736e-02
Chronic Medical Conditions_Yes      1.467724e-03
dtype: float64
```

Key Insights

- The following Chi-Square test reveals that Socio-demographic variables(age,martial status,education), lifestyle choices (physical activity,smoking and alcohol consumption) and health related factors(sleep patterns,family history of depression, chronic medical conditions) are associated with mental health

```python
In [ ]:
```