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


mh_df = pd.read_csv('/content/Mental Health Dataset/mental_health_dataset.csv')
mh_df.head(5)
```

	age	gender	employment_status	work_environment	mental_health_history	seeks_treatment	stress_level	sleep_hours	physical_activity
0	56	Male	Employed	On-site	Yes	Yes	6	6.2	
1	46	Female	Student	On-site	No	Yes	10	9.0	
2	32	Female	Employed	On-site	Yes	No	7	7.7	
3	60	Non-binary	Self-employed	On-site	No	No	4	4.5	
4	25	Female	Self-employed	On-site	Yes	Yes	3	5.4	

Frequency Table and Bar Charts

```
# Categorical variables
categorical_vars = [
    'gender', 'employment_status', 'work_environment',
    'mental_health_history', 'seeks_treatment', 'mental_health_risk'
]

# Frequency and Percentage Table + Bar Chart
for var in categorical_vars:
    print(f"\n--- {var.upper()} ---")

    freq_table = mh_df[var].value_counts().reset_index()
    freq_table.columns = [var, 'Frequency']
    freq_table['Percentage (%)'] = (freq_table['Frequency'] / freq_table['Frequency'].sum()) * 100

    print(freq_table)

    # Bar chart plot
    plt.figure(figsize=(8, 5))
    sns.countplot(data=mh_df, x=var, palette='pastel', order=mh_df[var].value_counts().index)
    plt.title(f'Distribution of {var}')
    plt.ylabel('Count')
    plt.xticks(rotation=45)
    plt.tight_layout()
    plt.savefig(f'{var}_bar_chart.png')
    plt.show()
```

Show hidden output

Observation

Frequency distribution of Categorical variables

> The descriptive analysis revealed a nearly equal distribution of male and female participants, indicating a balanced gender representation within the dataset. In terms of employment status, the majority of respondents were employed, suggesting that the sample predominantly reflects individuals engaged in the workforce.
>
> For work environment, most participants reported working on-site, followed by those working remotely, while hybrid workers constituted the smallest group. Regarding mental health history, a larger proportion of individuals reported having no prior mental health issues.
>
> When asked whether they currently seek mental health treatment, the majority responded no, indicating a low rate of active help-seeking behavior among participants. Finally, with respect to mental health risk, most respondents fell into the medium risk category, followed by those in the high and then low risk groups.

Histogram

```
# Continuous variables
continuous_vars = [
    'age', 'stress_level', 'sleep_hours', 'physical_activity_days',
    'depression_score', 'anxiety_score', 'social_support_score', 'productivity_score'
]

# Plot all histograms at once
plt.figure(figsize=(16, 12))
for i, var in enumerate(continuous_vars, 1):
    plt.subplot(3, 3, i)
    sns.histplot(mh_df[var], kde=True, color='skyblue', bins=20)
    plt.title(f'Distribution of {var}')
    plt.xlabel(var)
    plt.ylabel('Frequency')
    plt.tight_layout()

plt.suptitle("Histograms of Continuous Variables", y=1.02)
plt.tight_layout()
plt.savefig('histograms.png')
plt.show()
```

Show hidden output

Observation from Histogram

Sleep Hours

> The histogram analysis showed that among the continuous variables, only sleep hours appeared to follow a normal distribution, with a central peak around 7.75 hours, suggesting that the average participant gets an optimal amount of sleep.

Productivity score

> In contrast, the distribution of productivity scores was left-skewed (negatively skewed), indicating that most participants reported higher productivity, while a smaller portion reported lower productivity levels. This suggests that low productivity scores were relatively less common in the sample.

ALL other variables

> All other variables **(age, stress levels, physical activity days, depression score, anxiety score and social support score)**, show a "Platykurtic Distribution", indicating a low kurtosis.

Interpretation

> The values are more evenly distributed across the range.
>
> There is less concentration of scores around the mean.
>
> No strong central tendency — no single value dominates.

Cross Tabulation (Contingency Table

```
# 1. Work Environment vs Seeks Treatment
ct1 = pd.crosstab(mh_df['work_environment'], mh_df['seeks_treatment'], margins=True, normalize='index') * 100
print("\nWork Environment vs Seeks Treatment (Row %):")
print(ct1.round(2))

# 2. Work Environment vs Mental Health Risk
ct2 = pd.crosstab(mh_df['work_environment'], mh_df['mental_health_risk'], margins=True, normalize='index') * 100
print("\nWork Environment vs Mental Health Risk (Row %):")
print(ct2.round(2))

# 3. Employment Status vs Seeks Treatment
```

```
ct3 = pd.crosstab(mh_df['employment_status'], mh_df['seeks_treatment'], margins=True, normalize='index') * 100
print("\nEmployment Status vs Seeks Treatment (Row %):")
print(ct3.round(2))

# 4. Employment Status vs Mental Health Risk
ct4 = pd.crosstab(mh_df['employment_status'], mh_df['mental_health_risk'], margins=True, normalize='index') * 100
print("\nEmployment Status vs Mental Health Risk (Row %):")
print(ct4.round(2))
```

Show hidden output

Observation from contingency table

The contingency table analysis showed a remarkably consistent pattern across different categories of work environment and employment status:

work_environment vs seeks_treatment

> For the relationship between work environment and treatment-seeking behavior, approximately 60% of individuals across all categories (hybrid, on-site, and remote workers) reported not seeking mental health treatment.

work_environment vs mental_health_risk

> Similarly, in the cross-tabulation of work environment and mental health risk, around 59% of respondents across all work settings were classified under the medium mental health risk category.

employment_status vs seeks_treatment

> With respect to employment status and treatment-seeking behavior, a uniform trend was also observed: about 60% of individuals in each employment group (self-employed, employed, students, and unemployed) indicated that they did not seek treatment.

employment_status vs mental_health_risk

> Likewise, the distribution of mental health risk across employment status revealed that approximately 60% of individuals in each category were categorized as having medium risk

Interpretation

> These patterns suggest a lack of variability in both treatment-seeking behavior and perceived mental health risk across different work-related categories.

Scatter Plot (Pairplot)

```
# List of continuous variables
continuous_vars = [
    'age', 'stress_level', 'sleep_hours', 'physical_activity_days',
    'depression_score', 'anxiety_score', 'social_support_score', 'productivity_score'
]

# Pairplot (scatter plots in a grid + histograms)
sns.pairplot(mh_df[continuous_vars], corner=True, plot_kws={'alpha': 0.6, 's': 40})
plt.suptitle("Scatter Plot Matrix of Continuous Variables", y=1.02)
plt.tight_layout()
plt.savefig('scatter_plot_matrix.png')
plt.show()
```

Show hidden output

Observation of Scatter Plot

depression_score vs productivity_score

> There appears to be a negative correlation between the variables.

> As depression score increases, productivity tends to decrease.
>
> The relationship appears linear and strong.

Start coding or generate with AI.