```
import pandas as pd
import numpy as np
from matplotlib import pyplot as plt
import seaborn as sns
from scipy.stats import norm as norm
```

```
#mount to google drive
from google.colab import drive
drive.mount('/content/drive')
```

```
Mounted at /content/drive
```

Source of Data: Laksika Tharmalingam - Kaggle

https://www.kaggle.com/datasets/uom190346a/sleep-health-and-lifestyle-dataset

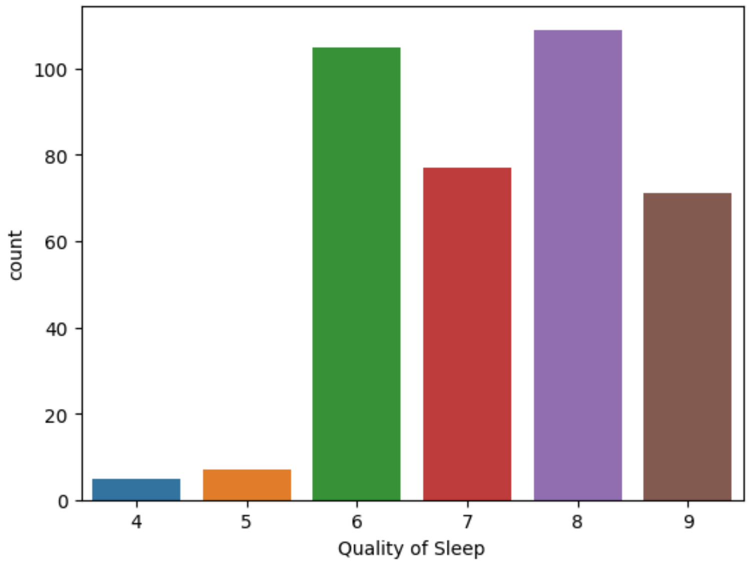
```
sleep_df = pd.read_csv('/content/archive.zip')
```

```
sleep_df.head()
```

	Person ID	Gender	Age	Occupation	Sleep Duration	Quality of Sleep	Physical Activity Level	Stress Level	BMI Category	Blood Pressure	Heart Rate	Daily Steps	Sleep Disorder
0	1	Male	27	Software Engineer	6.1	6	42	6	Overweight	126/83	77	4200	None
1	2	Male	28	Doctor	6.2	6	60	8	Normal	125/80	75	10000	None
2	3	Male	28	Doctor	6.2	6	60	8	Normal	125/80	75	10000	None
3	4	Male	28	Sales Representative	5.9	4	30	8	Obese	140/90	85	3000	Sleep Apnea
4	5	Male	28	Sales	5.9	4	30	8	Obese	140/90	85	3000	Sleep

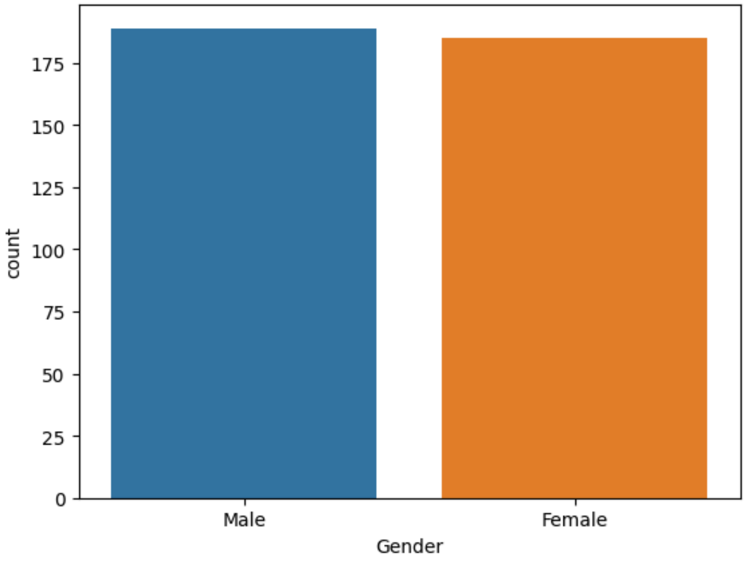
```
#Visualization One
sns.countplot(data= sleep_df, x= 'Quality of Sleep')
```

```
<Axes: xlabel='Quality of Sleep', ylabel='count'>
```



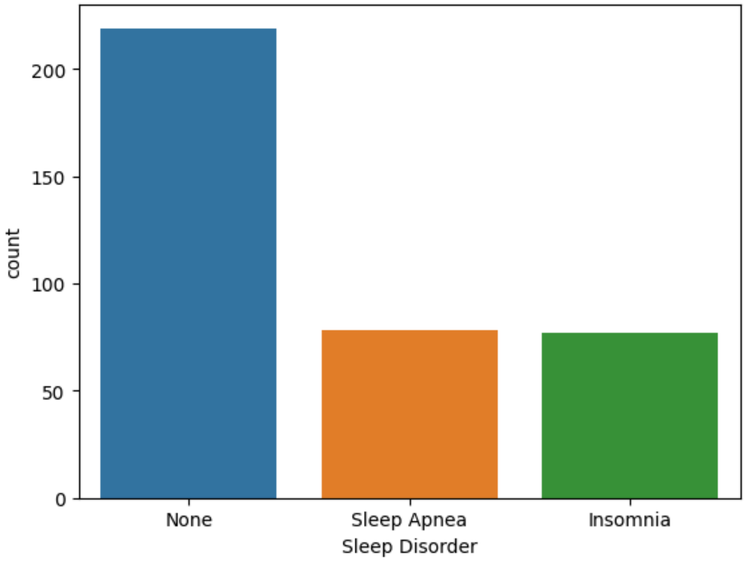
```
#Visualization Two
sns.countplot(data= sleep_df, x= 'Gender')
```

```
<Axes: xlabel='Gender', ylabel='count'>
```



```
#Visualization Three
sns.countplot(data= sleep_df, x='Sleep Disorder')
```

```
<Axes: xlabel='Sleep Disorder', ylabel='count'>
```



Hypothesis Test 1:

Is a docotor's average sleep duration statistically different from the other occupations' average sleep duration?

Null: A doctor's average sleep duration is not statistically different from other the other occupations

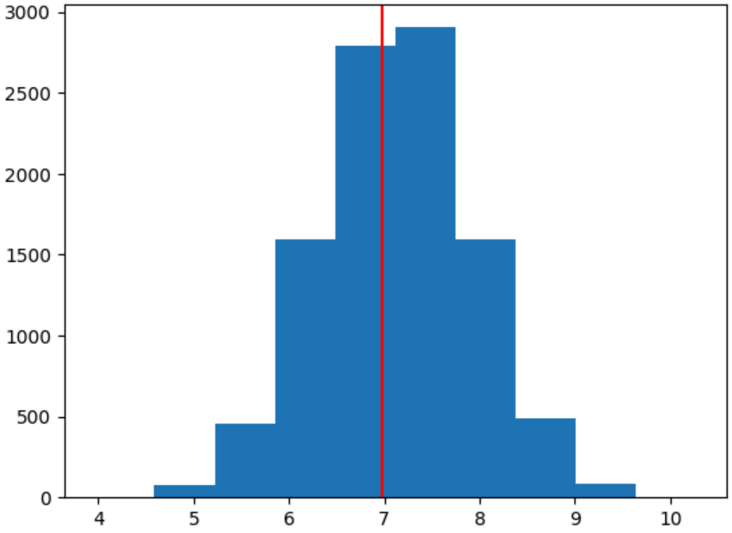
Alt: A doctor's average sleep duration is statistically different from the other occupations

Significance Level: If a doctor's sleep duration is 1.5 hours different from other occupations than doctor's average sleep duration is significantly higher than the other occupations.

```
#observed value
doc_df= sleep_df[sleep_df['Occupation']== 'Doctor']
doc_sleep_durat= doc_df['Sleep Duration'].values
avg_doc_sleep_durat= doc_sleep_durat.mean() #Test statistic

#mean sleep duration for other occupations
mean_sleep_durat= sleep_df['Sleep Duration'].mean()

#standard deviation of sleep duration for other occupations
std_sleep_durat= sleep_df['Sleep Duration'].std()
```

```
#generate 10,000 random variables
random_variables= norm.rvs(loc=mean_sleep_durat, scale=std_sleep_durat, size=10000)


#Histogram
all_sleep_durat=sleep_df['Sleep Duration']
plt.hist(random_variables)
plt.axvline(x = avg_doc_sleep_durat, color='red')
```

```
<matplotlib.lines.Line2D at 0x7fc53f389cc0>
```



```
#p-value here
p_value= len(np.where(random_variables>=avg_doc_sleep_durat)[0])/10000
p_value
```

```
0.5778
```

Conclusion:

I can accept the null hypothesis because the observed value is within the blue. Meaning, a doctor's average sleep duration is not statistically different from other occupations.

Hypothesis Test 2:

Does quality of sleep effect average sleep duration? To do this, I will split quality up into "good" and "bad" and then compare the average sleep durations for each category.

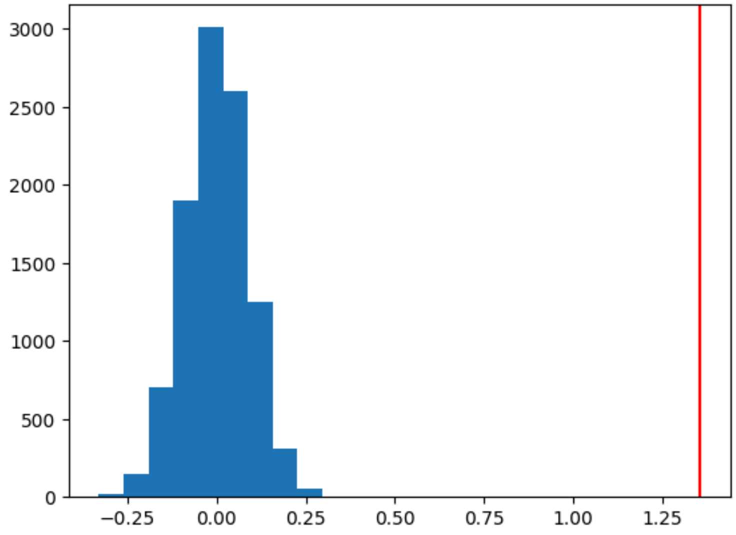
Null: The quality of sleep does not effect average sleep duration.

Alt: The quality of sleep does effect average sleep duration.

```
#spliting up sleep quality
good_quality= sleep_df[sleep_df['Quality of Sleep']>=7]
bad_quality= sleep_df[sleep_df['Quality of Sleep']<7]


#creating test statistic
good_sleep_durat= good_quality['Sleep Duration'].values
avg_good_sleep_durat= good_sleep_durat.mean()

bad_sleep_durat= bad_quality['Sleep Duration'].values
avg_bad_sleep_durat= bad_sleep_durat.mean()

test_statistic= avg_good_sleep_durat - avg_bad_sleep_durat


good_bad_quality= sleep_df[(sleep_df['Quality of Sleep']>=7)| (sleep_df['Quality of Sleep']< 7)]
sleep_durat_all_quality = good_bad_quality['Sleep Duration'].values
all_quality = good_bad_quality['Quality of Sleep'].values
```

```
def simulate_test_stat(sleep_durat_all_quality, all_quality):
  shuffle_quality= all_quality
  np.random.shuffle(shuffle_quality)

  idx_good= np.where(shuffle_quality>=7)
  avg_good_sleep_durat= np.average(sleep_durat_all_quality[idx_good])

  idx_bad= np.where(shuffle_quality < 7)
  avg_bad_sleep_durat= np.average(sleep_durat_all_quality[idx_bad])

  sim_test_stat= avg_good_sleep_durat - avg_bad_sleep_durat
  return sim_test_stat


#simulate 10,000 test stats
sim_test_stats = []
numsim= 10000
for i in range(numsim):
 stat = simulate_test_stat(sleep_durat_all_quality, all_quality)
 sim_test_stats.append(stat)


#create a histogram
plt.hist(sim_test_stats)
plt.axvline(x=test_statistic, c= 'r')
```

```
<matplotlib.lines.Line2D at 0x7c44f9681960>
```



Conclusion:

Since the observed value does not fall within the values simulated under the null, we reject the null and accept the alternative. Meaning, the quality of sleep does effect the average sleep duration.

Hypothesis Test 3:

Are the average daily steps for doctors statistically different from the average daily steps for sales representatives? To do this, I will bootstrap a 95% confidence interval.

Null: The average daily steps of a doctor is not statistically different from the average daily steps of sales representatives.

Alt: The average daily steps of a doctor is statistically different from the average daily steps of sales representatives.

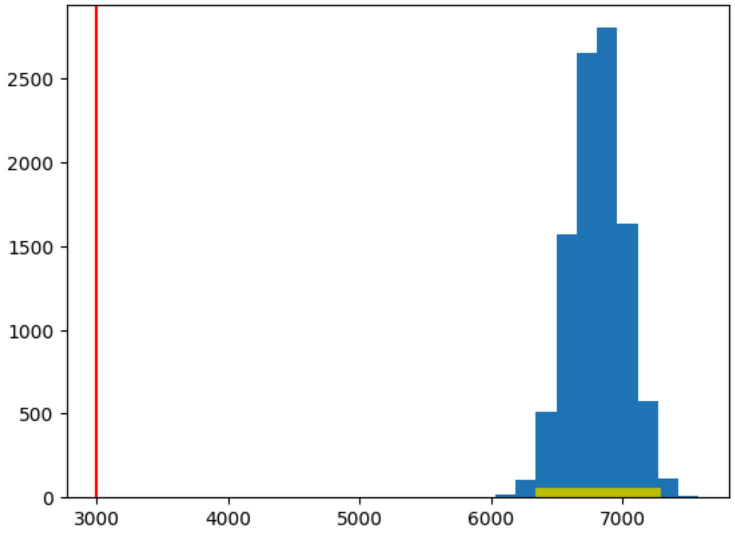
```
#Finding the point estimate of the mean of daily steps of doctors
doctor_df= sleep_df[sleep_df['Occupation']=='Doctor']
doctor_steps= doctor_df['Daily Steps']
avg_doctor_steps= doctor_df['Daily Steps'].mean()


def one_bootstrap_mean(sample_df):
  bootstrapped_mean = sample_df.sample(n=len(sample_df), replace = True)['Daily Steps'].mean()
  return bootstrapped_mean
```

```
num_sim=10000
bootstrap_means= []
for i in range(num_sim):
  bootstrap_means.append(one_bootstrap_mean(doctor_df))

left_interval_endpoint = np.percentile(bootstrap_means, 2.5)
print(left_interval_endpoint)

right_interval_endpoint = np.percentile(bootstrap_means, 97.5)
print(right_interval_endpoint)
```

```
6407.042253521126
7223.978873239436
```

```
#create histogram
interval= np.array([left_interval_endpoint, right_interval_endpoint])

sales_df=sleep_df[sleep_df['Occupation']=='Sales Representative']
avg_sales_steps= sales_df['Daily Steps'].mean()
print(avg_sales_steps)

plt.hist(bootstrap_means)
plt.plot(interval, [0, 0], linewidth = 10, c = 'y') #confidence interval
plt.axvline(x=avg_sales_steps, c= 'r')
```

```
3000.0
<matplotlib.lines.Line2D at 0x7c44f91fe350>
```



Conclusion:

Because the red line does not fall within the confidence region, we can conclude that the average daily steps of doctors is statistically different from the average daily steps of sales representatives.