Open in Colab

Sleep_health_and_lifestyle 데이터 준비

데이터 불러오기

- pandas를 이용해서 데이터 불러오기

In [13]:
```
import pandas as pd

sleep_data = pd.read_csv("/content/Sleep_health_and_lifestyle.csv")
sleep_data.head()
```

Out[13]:

	Gender	Age	Occupation	Sleep Duration	Quality of Sleep	Physical Activity Level	Stress Level	BMI Category	Blood Pressure	Heart Rate	Daily Steps	Sleep Disorder
0	Male	27	Software Engineer	6.1	6	42	6	Overweight	126/83	77	4200	NaN
1	Male	28	Doctor	6.2	6	60	8	Normal	125/80	75	10000	NaN
2	Male	28	Doctor	6.2	6	60	8	Normal	125/80	75	10000	NaN
3	Male	28	Sales Representative	5.9	4	30	8	Obese	140/90	85	3000	Sleep Apnea
4	Male	28	Sales Representative	5.9	4	30	8	Obese	140/90	85	3000	Sleep Apnea

- 데이터 칼럼 확인

In [14]:
```
sleep_data.columns
```

Out[14]:
```
Index(['Gender', 'Age', 'Occupation', 'Sleep Duration', 'Quality of Sleep',
       'Physical Activity Level', 'Stress Level', 'BMI Category',
       'Blood Pressure', 'Heart Rate', 'Daily Steps', 'Sleep Disorder'],
      dtype='object')
```

데이터 전처리

- '**Sleep Disorder**' 칼럼에 있는 결측 데이터를 확인해보았다.
  - NaN인 경우가 많았는데, 다음 칼럼 주제에 해당되지 않아서 적혀있지 않은 것 같다.
  - 'healthy'로 변경해주었다.

In [15]:
```
sleep_data[sleep_data["Sleep Disorder"].isna()]
```

Out[15]:

	Gender	Age	Occupation	Sleep Duration	Quality of Sleep	Physical Activity Level	Stress Level	BMI Category	Blood Pressure	Heart Rate	Daily Steps	Sleep Disorder
0	Male	27	Software Engineer	6.1	6	42	6	Overweight	126/83	77	4200	NaN
1	Male	28	Doctor	6.2	6	60	8	Normal	125/80	75	10000	NaN
2	Male	28	Doctor	6.2	6	60	8	Normal	125/80	75	10000	NaN
7	Male	29	Doctor	7.8	7	75	6	Normal	120/80	70	8000	NaN
8	Male	29	Doctor	7.8	7	75	6	Normal	120/80	70	8000	NaN
...	...	...	...	...	...	...	...	...	...	...	...	...
341	Female	56	Doctor	8.2	9	90	3	Normal Weight	118/75	65	10000	NaN
342	Female	56	Doctor	8.2	9	90	3	Normal Weight	118/75	65	10000	NaN
343	Female	57	Nurse	8.1	9	75	3	Overweight	140/95	68	7000	NaN
358	Female	59	Nurse	8.0	9	75	3	Overweight	140/95	68	7000	NaN
359	Female	59	Nurse	8.1	9	75	3	Overweight	140/95	68	7000	NaN

219 rows × 12 columns

In [16]:
```
sleep_data.fillna({"Sleep Disorder":"healthy"}, inplace= True)

sleep_data[sleep_data["Sleep Disorder"].isna()]
```

Out[16]:

	Gender	Age	Occupation	Sleep Duration	Quality of Sleep	Physical Activity Level	Stress Level	BMI Category	Blood Pressure	Heart Rate	Daily Steps	Sleep Disorder

---

데이터 가공

- 이 데이터들 중에서 이번 프로젝트에서는 수치형 데이터인 'Age', 'Sleep Duration', 'Quality of Sleep', 'Physical Activity Level', 'Stress Level', 'Heart Rate', 'Daily Steps' 만 이용한다.
  - 수치형 데이터가 아닌 나머지 칼럼은 제거한다.

In [17]:
```
sleep_data = sleep_data.drop(['Gender', 'Occupation', 'BMI Category', 'Blood Pressure', 'Sleep Disorder'], axis = 1)

sleep_data.columns
```

Out[17]:
```
Index(['Age', 'Sleep Duration', 'Quality of Sleep', 'Physical Activity Level',
       'Stress Level', 'Heart Rate', 'Daily Steps'],
      dtype='object')
```

데이터 저장

- 가공한 데이터를 csv파일로 저장한다.
- 이후 대시보드를 만드는데 활용한다.

In [18]:
```
from google.colab import files
```

In [19]:
```
sleep_data.to_csv("Sleep_health_and_lifestyle_dataset.csv", index=False)
```

In [20]:
```
files.download("Sleep_health_and_lifestyle_dataset.csv")
```