데이터처리프로그래밍(02) - 컴퓨터학과 20201001 이승연

1. 프로젝트 정의

- **데이터 분석의 목적**

'수면 시간이 많으면 수면의 질이 높아지는지'에 대한 궁금증을 기반으로 데이터를 탐색하고 분석 대상으로 선정하게 되었다. 수면 건강과 라이프스타일 데이터에 대한 분석 및 시각화를 통해 수면의 질과 다른 변수 간 관계를 이해하고, 수면의 질에 영향을 미치는 변수를 파악하는 것을 목적으로 한다.

- **데이터 분석 내용**

연령대, 수면의 질, 수면 시간, 스트레스 수준 등 각 변수 간의 관계를 탐색했다.

이를 위해 히스토그램, 파이 그래프, 바이올린 플롯, 막대 그래프, 산점도 등 다양한 시각화를 통해 살펴보았다.

또한 다중 선형 회귀 분석을 통해 수면의 질에 미치는 영향을 설명하는 모델을 도출했다.

- **기대효과**

선형 회귀 분석과 다양한 시각화를 통해 성별, 연령대, 스트레스 지수 등 변수 간의 관계를 파악하며, 이를 기반으로 수면의 질을 높이고 수면 건강을 개선하기 위한 향후 연구 방향을 도출할 수 있을 것으로 기대된다.

2. 데이터 소개

- **데이터 수집 일시**

해당 데이터는 2023년 12월 5일 공개 데이터 사이트 캐글에서 다운로드했다.

데이터 출처 링크:

https://www.kaggle.com/datasets/851c829b2a41e6dd0b5a60388cd4a2cfda2d54433450ed12141237416c8161bc

- **데이터 규모**

수면 건강 및 라이프스타일 데이터는 374개의 행과 13개의 열로 구성되어 있다.

- **데이터 수집 방법**

개인 식별자(Personal ID), 성별(Gender), 나이(Age), 직업(Occupation), 수면 시간(Sleep Duration), 수면의 질(Quality of Sleep), 신체 활동 수준(Physical Activity Level), 스트레스 수준(Stress Level), BMI 분류(BMI Category), 혈압(Blood Pressure), 심박수(Heart Rate), 일일 걸음 수(Daily Steps), 수면 장애(Sleep Disorder)로 열이 구성되어 있었다.

이 중에서 성별, 나이, 수면 시간, 수면의 질, 신체 활동 수준, 스트레스 수준, 수면 장애, 총 7개의 열을 가져왔다.

3. 데이터 전처리

- **데이터의 결측치 처리**

해당 데이터에는 null 값이 없어 별도의 처리를 하지 않았다.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 374 entries, 0 to 373
Data columns (total 7 columns):
 #   Column                  Non-Null Count  Dtype
---  ------                  --------------  -----
 0   Gender                  374 non-null    object
 1   Age                     374 non-null    int64
 2   Sleep Duration          374 non-null    float64
 3   Quality of Sleep        374 non-null    int64
 4   Physical Activity Level  374 non-null    int64
 5   Stress Level            374 non-null    int64
 6   Sleep Disorder          374 non-null    object
dtypes: float64(1), int64(4), object(2)
memory usage: 20.6+ KB
```

- **각 컬럼의 유일한 값 확인**

각 컬럼에 대해 unique() 함수를 적용하여, 아래와 같이 유일한 값들을 확인하였다.

```
Gender ['Male' 'Female']

Age [27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 48 49 50 51 52
 53 54 55 56 57 58 59]

Sleep Duration [6.1 6.2 5.9 6.3 7.8 6.  6.5 7.6 7.7 7.9 6.4 7.5 7.2 5.8 6.7 7.3 7.4 7.1
 6.6 6.9 8.  6.8 8.1 8.3 8.5 8.4 8.2]

Quality of Sleep [6 4 7 5 8 9]

Physical Activity Level [42 60 30 40 75 35 45 50 32 70 80 55 90 47 65 85]

Stress Level [6 8 7 4 3 5]

Sleep Disorder ['None' 'Sleep Apnea' 'Insomnia']
```

- **연령대 열 추가**

Age를 기반으로 setAgeGroup 함수를 작성하여 연령대를 나타내는 Ageg 열을 추가하였다.

```
def setAgeGroup(x):
  if x >= 20 and x < 30:
    return '20 대'
  elif x >= 30 and x < 40:
    return '30 대'
  elif x >= 40 and x < 50:
    return '40 대'
  elif x >= 50 and x < 60:
    return '50 대'

lifestyle_df['Ageg'] = lifestyle_df['Age'].apply(setAgeGroup)
```

- **수면 장애 그룹화**

Sleep Apnea 열의 데이터에서 'None' 값을 '없음'으로, 'Sleep Apnea'(수면 무호흡증)과 'Insomnia'(불면증)를 '있음'으로 변경했다.

```
def setSleepDisorder(x):
  if x == 'Sleep Apnea' or x == 'Insomnia':
    return '있음'
  elif x == 'None':
    return '없음'
```

```
lifestyle_df['Sleep Disorder'] = lifestyle_df['Sleep
Disorder'].apply(setSleepDisorder)
```

- **컬럼명을 한글로 변경**

  시각화를 위해 컬럼명을 한글로 변경하였다.

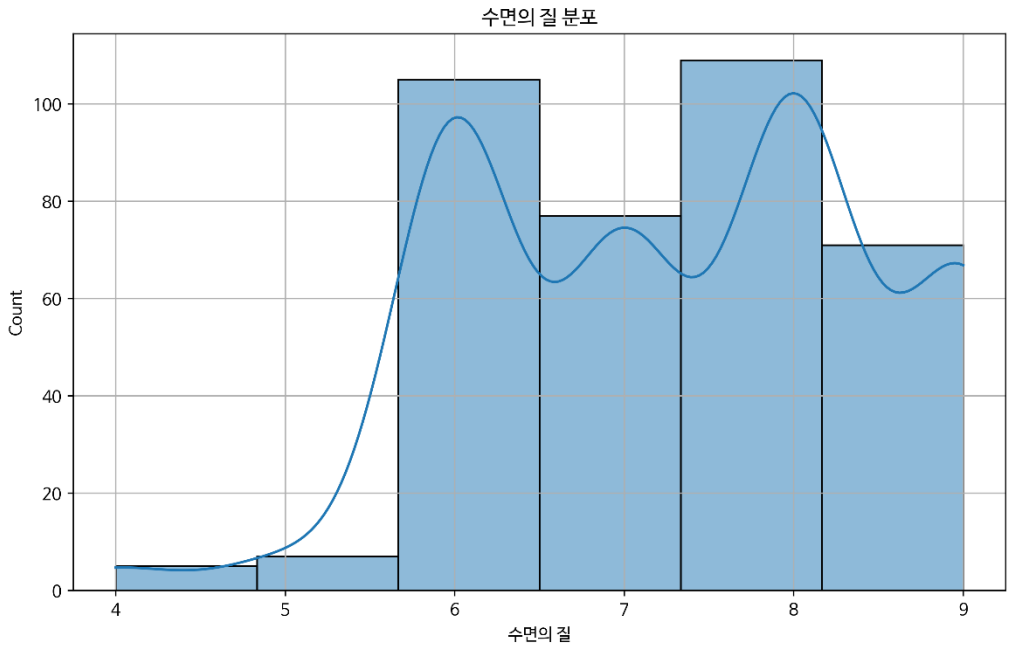
```
col_mapping = {'Gender':'성별', 'Age':'나이', 'Sleep Duration':'수면
시간', 'Quality of Sleep':'수면의 질', 'Physical Activity Level':'신체 활동
수준', 'Stress Level':'스트레스 수준', 'Sleep Disorder':'수면 장애',
'Ageg':'연령대'}
lifestyle_df = lifestyle_df.rename(columns = col_mapping)
```

4. 데이터 분석

- **탐색 및 시각화**

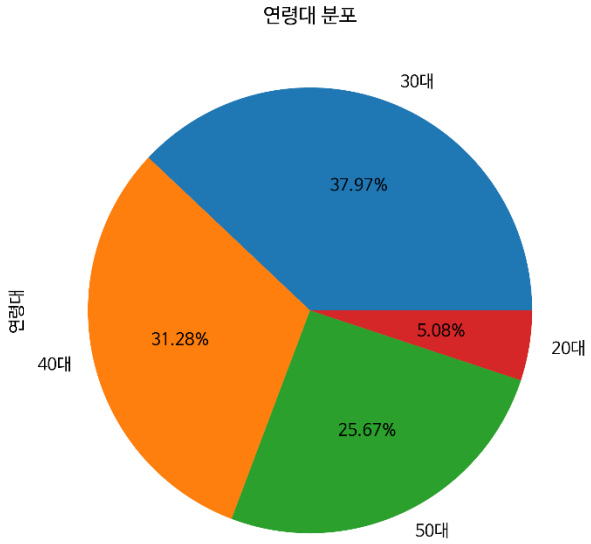
  1. 수면의 질 분포

  수면의 질 데이터를 기반으로 한 히스토그램으로, 전체 데이터에서 나타나는 수면의 질의 분포를 확인할 수 있다.



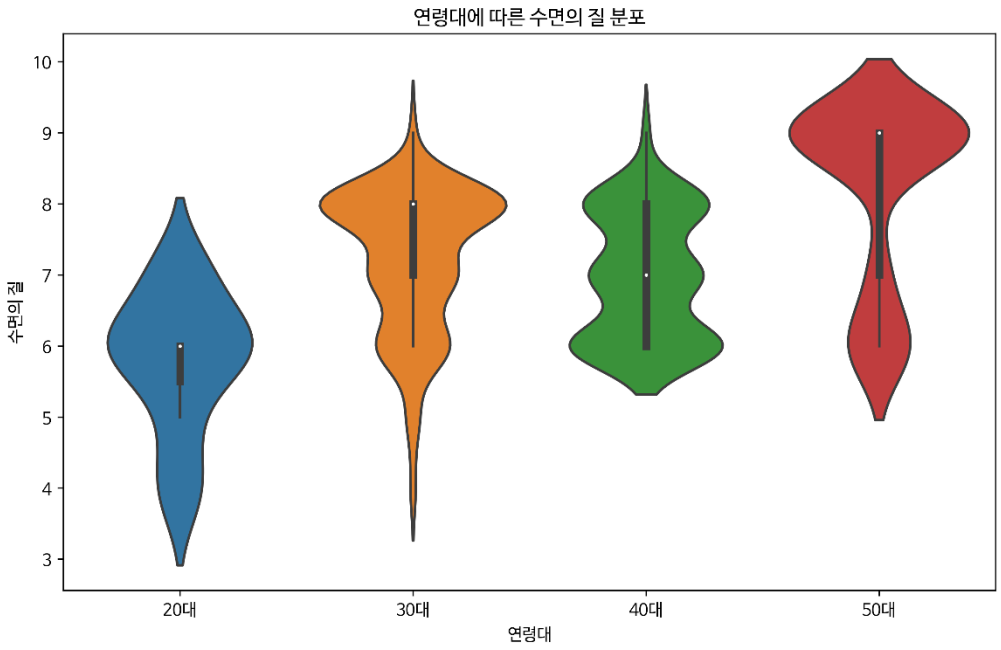
2. 연령대 분포

각 연령대의 비율을 파이 그래프로 시각화한 결과이다. 40대, 50대가 다른 연령대에 비해 상대적으로 많은 것을 확인할 수 있다.



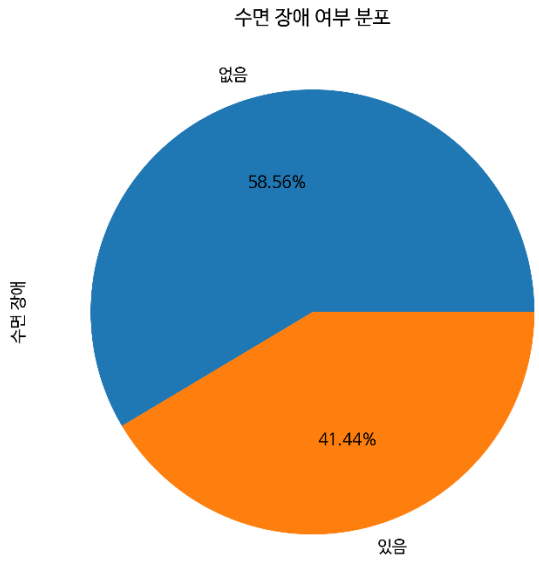
3. 연령대에 따른 수면의 질

각 연령대에 따른 수면의 질을 바이올린 플롯으로 시각화한 결과이다. 연령대가 높을수록 수면의 질이 높은 쪽에 많이 분포되어 있다.



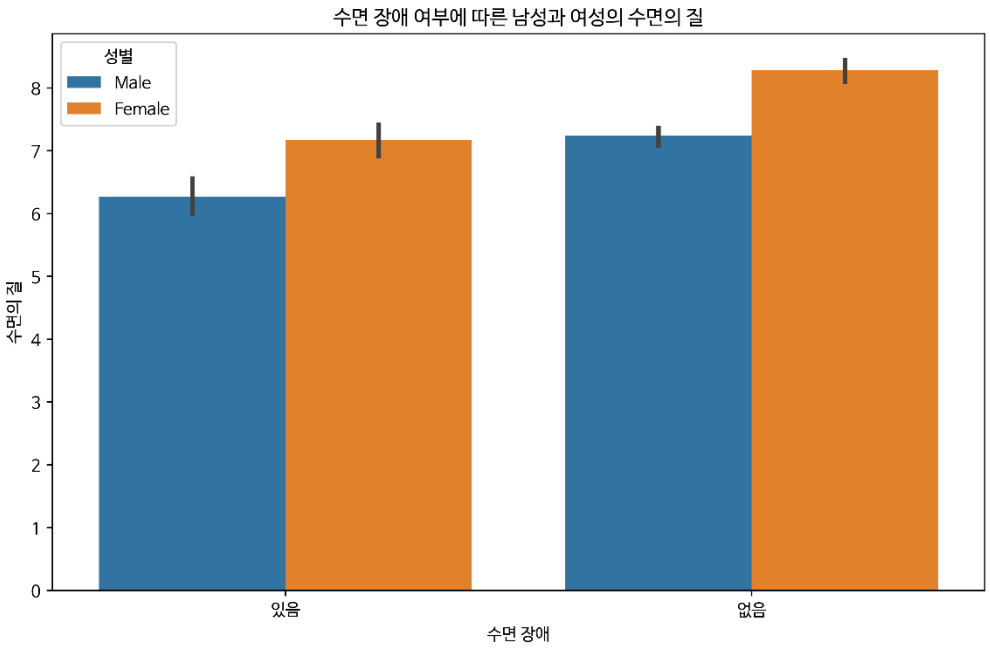
4. 수면 장애 여부 분포

수면 장애 비율을 파악하기 위해 파이 그래프로 시각화한 결과이다. 수면 장애가 없는 비율이 상대적으로 많은 비율을 차지하고 있다.



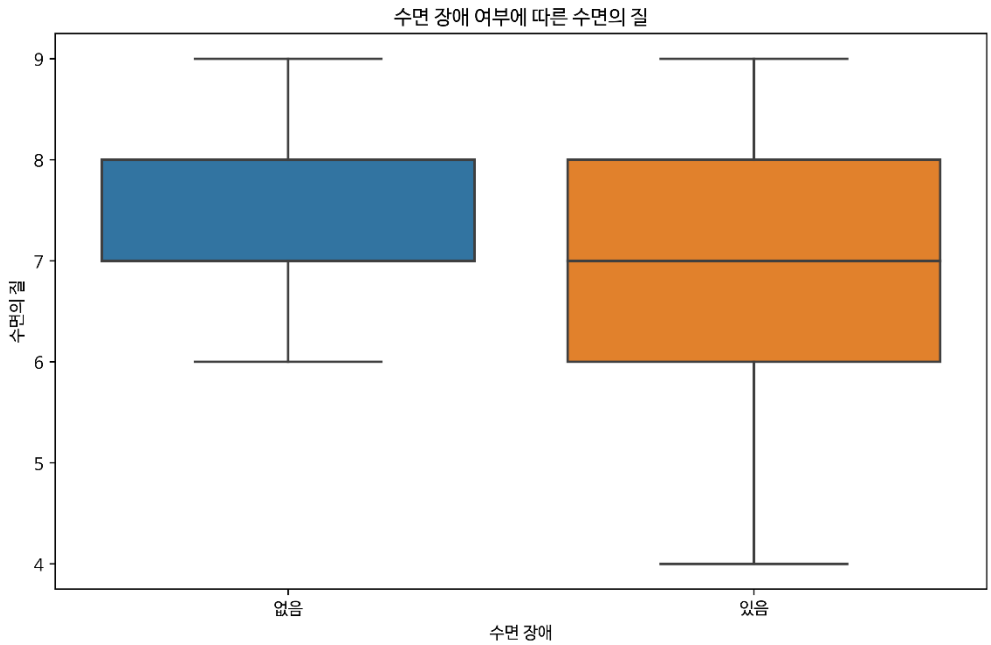
5. 수면 장애 여부에 따른 남성과 여성의 수면의 질

수면 장애 여부와 성별에 따라 수면의 질을 막대 그래프로 시각화한 결과이다. 남성과 여성 모두 수면 장애가 없는 경우 수면의 질이 높은 것을 확인할 수 있다.



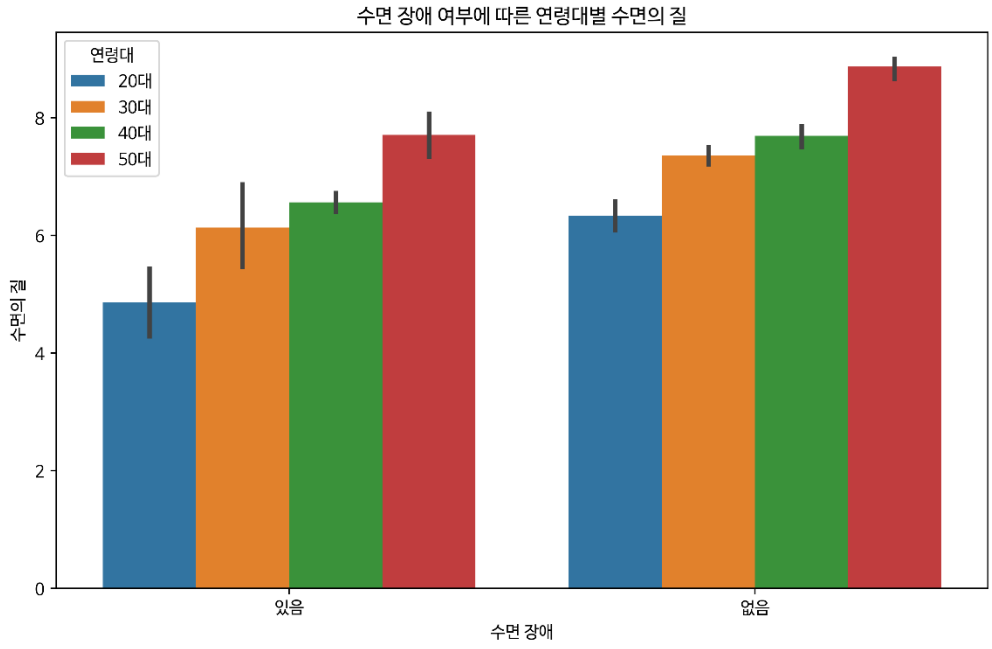
6. 수면 장애 여부에 따른 수면의 질

수면 장애 여부에 따른 수면의 질을 상자 수염 그래프로 시각화한 결과이다. 수면 장애가 있는 경우 수면의 질의 최솟값이 더 낮고 분포가 다양하게 나타난다. 이는 수면 장애가 수면의 질에 영향을 미칠 수 있다는 것을 시사한다.



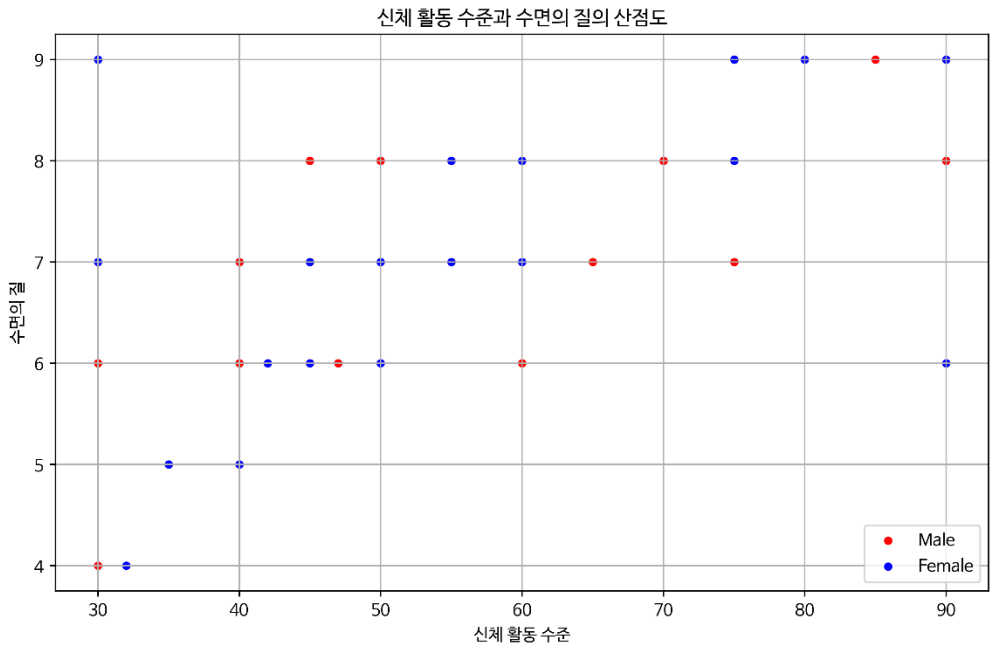
7. 수면 장애 여부에 따른 연령대별 수면의 질

수면 장애가 없는 경우 수면의 질이 전체적으로 높다. 또한 수면 장애 여부와 관계없이 연령대가 높을수록 수면의 질이 높게 나타났다.



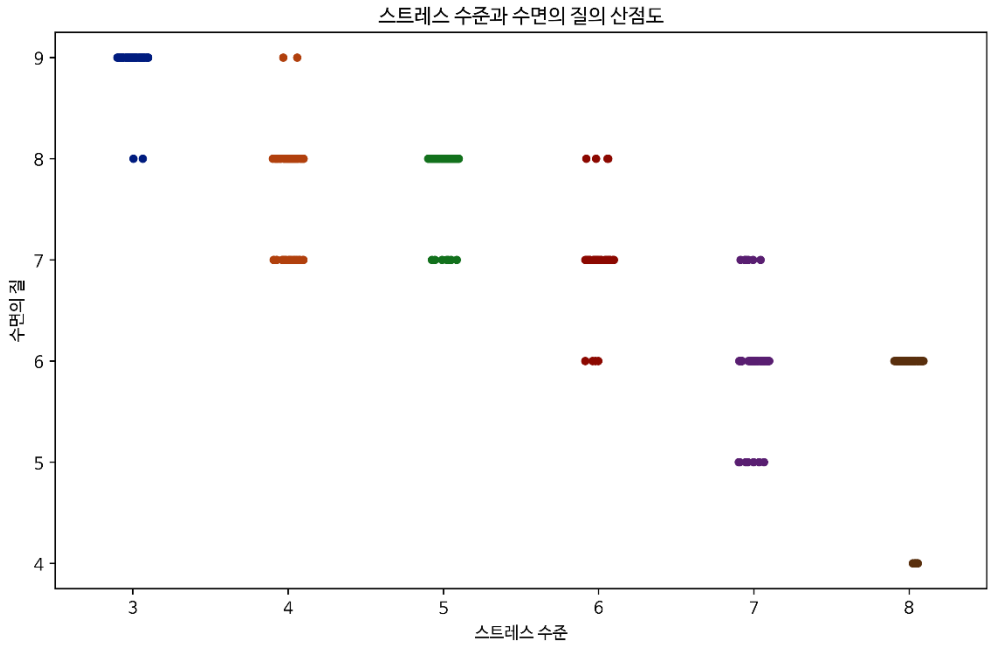
8. 신체 활동 수준과 수면의 질

산점도를 통해 신체 활동 수준과 수면의 질 관계를 보면 성별 간 큰 차이가 나타나지 않는 것을 알 수 있다.



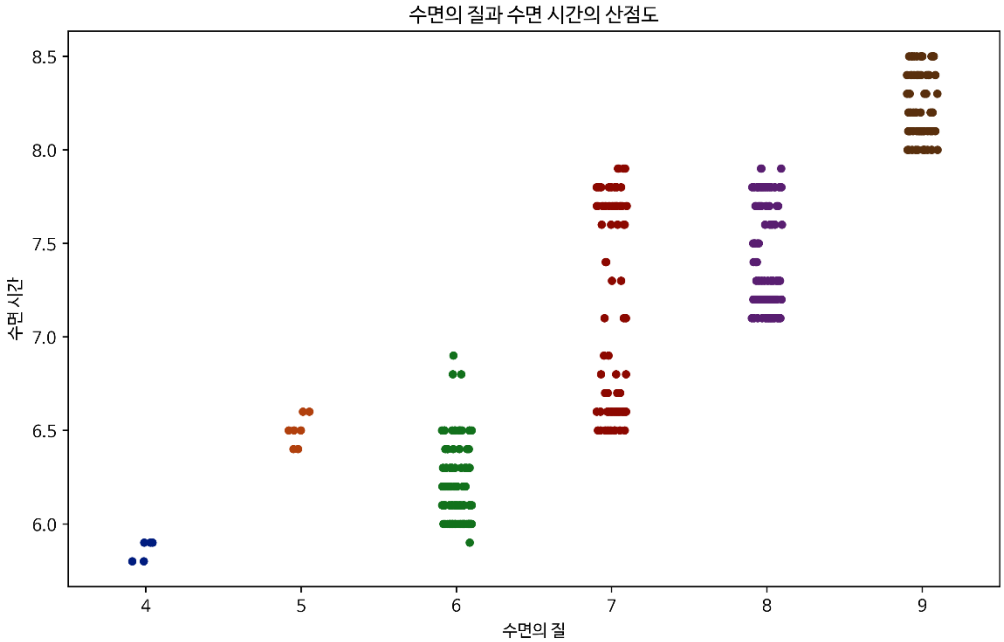
9. 스트레스 수준과 수면의 질

스트레스 수준과 수면의 질 간의 산점도를 통해 스트레스 수준이 높을수록 수면의 질이 감소하는 것을 확인할 수 있다. 이는 스트레스 수준이 수면의 질에 영향을 미칠 수 있다는 것을 시사한다.



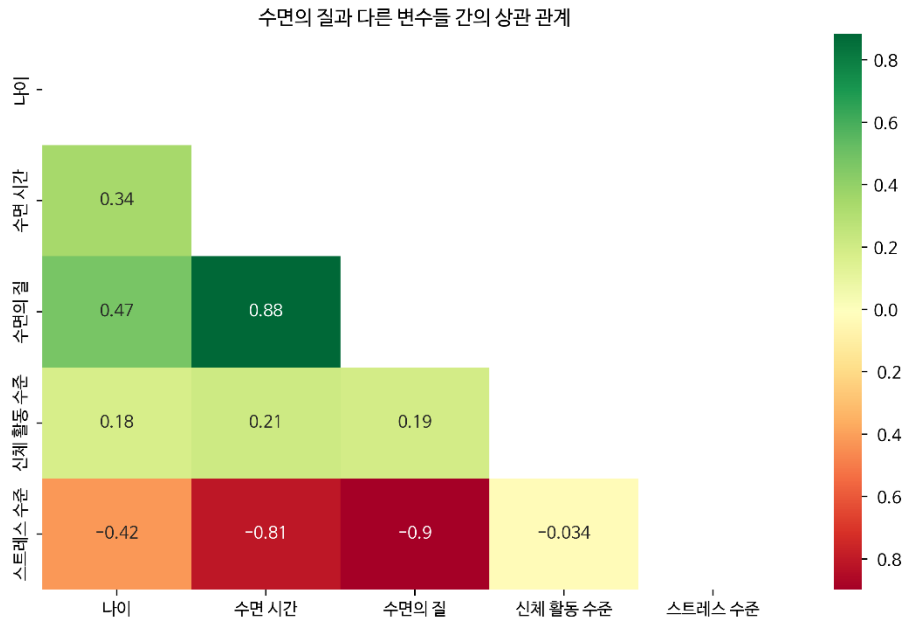
10. 수면의 질과 수면 시간

수면의 질과 수면 시간의 산점도를 보면 수면 시간이 많을수록 수면의 질이 높게 나타난다.



11. 변수 간의 상관관계 분석

히트맵을 통해 변수 간의 상관관계를 시각화한 결과이다. 녹색은 양의 상관관계, 빨간색은 음의 상관관계를 나타내며 진할수록 강한 상관관계를 나타낸다. 히트맵 분석 결과, 수면의 질은 수면 시간과 강한 양의 상관관계를 보이며 스트레스 수준은 수면 시간, 수면의 질과 강한 음의 상관관계를 보인다.



- 회귀 분석

다중 선형 회귀 분석 결과, R-squared(결정 계수) 값은 0.892로, 설명력이 89.2% 정도이다. P-value 값을 보면 0.05보다 작기 때문에 이 모델의 계수들은 유의미한 것으로 나타난다.

따라서 나이, 수면 시간, 신체 활동 수준, 스트레스 수준이 수면의 질에 유의미한 영향을 미치며, 이 모델은 전반적으로 데이터를 잘 설명하고 있다고 볼 수 있다.

OLS Regression Results

Dep. Variable:	수면의질	R-squared:	0.892
Model:	OLS	Adj. R-squared:	0.891
Method:	Least Squares	F-statistic:	762.7
Date:	Mon, 18 Dec 2023	Prob (F-statistic):	6.44e-177
Time:	09:22:26	Log-Likelihood:	-181.05
No. Observations:	374	AIC:	372.1
Df Residuals:	369	BIC:	391.7
Df Model:	4		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	3.9571	0.439	9.023	0.000	3.095	4.819
나이	0.0137	0.003	5.180	0.000	0.009	0.019
수면시간	0.6207	0.046	13.372	0.000	0.529	0.712
신체활동수준	0.0040	0.001	3.796	0.000	0.002	0.006
스트레스수준	-0.3505	0.021	-16.466	0.000	-0.392	-0.309

Omnibus:	35.280	Durbin-Watson:	1.084
Prob(Omnibus):	0.000	Jarque-Bera (JB):	54.069
Skew:	-0.628	Prob(JB):	1.82e-12
Kurtosis:	4.375	Cond. No.	1.63e+03

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 1.63e+03. This might indicate that there are
strong multicollinearity or other numerical problems.

5. 결론

- 프로젝트를 통해 얻은 결론

'수면 시간이 많으면 수면의 질이 높아지는지'에 대한 궁금증에서 시작하여 데이터를 선택하고, 분석하고 시각화를 해 보았다. 다중 선형 회귀 분석 결과 수면 시간, 나이, 신체 활동 수준, 스트레스 수준이 수면의 질에 미치는 영향을 설명하는 모델이 유의미하게 도출되었다. 또한 히트맵을 통한 변수 간 상관 분석 결과, 수면 시간과 수면의 질은 높은 양의 상관관계(0.88)를 보였다. 따라서 수면 시간이 많을수록 수면의 질이 높다는 결론을 도출할 수 있었다.

이외에도 스트레스 수준, 신체 활동 능력, 연령대 등 변수들이 수면의 질과 어떤 관계를 가지는지 분석하기 위해 시각화를 진행했다. 산점도를 통해 스트레스 수준이 낮으면 수면의 질이 낮고, 신체 활동 수준이 높으면 수면의 질이 높게 나타났다. 이 또한 히트맵에서 수면의 질과 상관관계가 높게 나타났으므로 수면의 질에 영향을 미치는 변수인 것을 확인할 수 있었다.

이런 결과를 토대로 개인의 수면 시간, 스트레스 수준 등 라이프스타일에 주의를 기울인다면 수면의 질 향상에 도움이 될 것으로 기대된다.