COGS 108 WI23 Final Project

Group 50: Leena Kang, Stephanie Park, Nicholas Azpeitia, Jorge Miguel Hernandez

Research Question

What (if any) effect do demographic factors such as gender and age have in determining an individual's sleep efficiency? What (if any) effect do living habits (smoking, drinking, caffeine consumption and exercise) have in determining an individual's sleep efficiency? Which of the two provides a more accurate prediction of an individual's sleep efficiency?

Background & Prior Work

- 4 stages of sleep: Stage 1, Stage 2, **Stage 3** & REM sleep
  - Stage 3 = deep sleep → good sleep efficiency
- Prior research shows impact of age, substance consumption & sleeping environment on sleep efficiency
  - Older age: harder to fall asleep, more likely to wake up
  - Alcohol/caffeine consumption: reduced sleep time

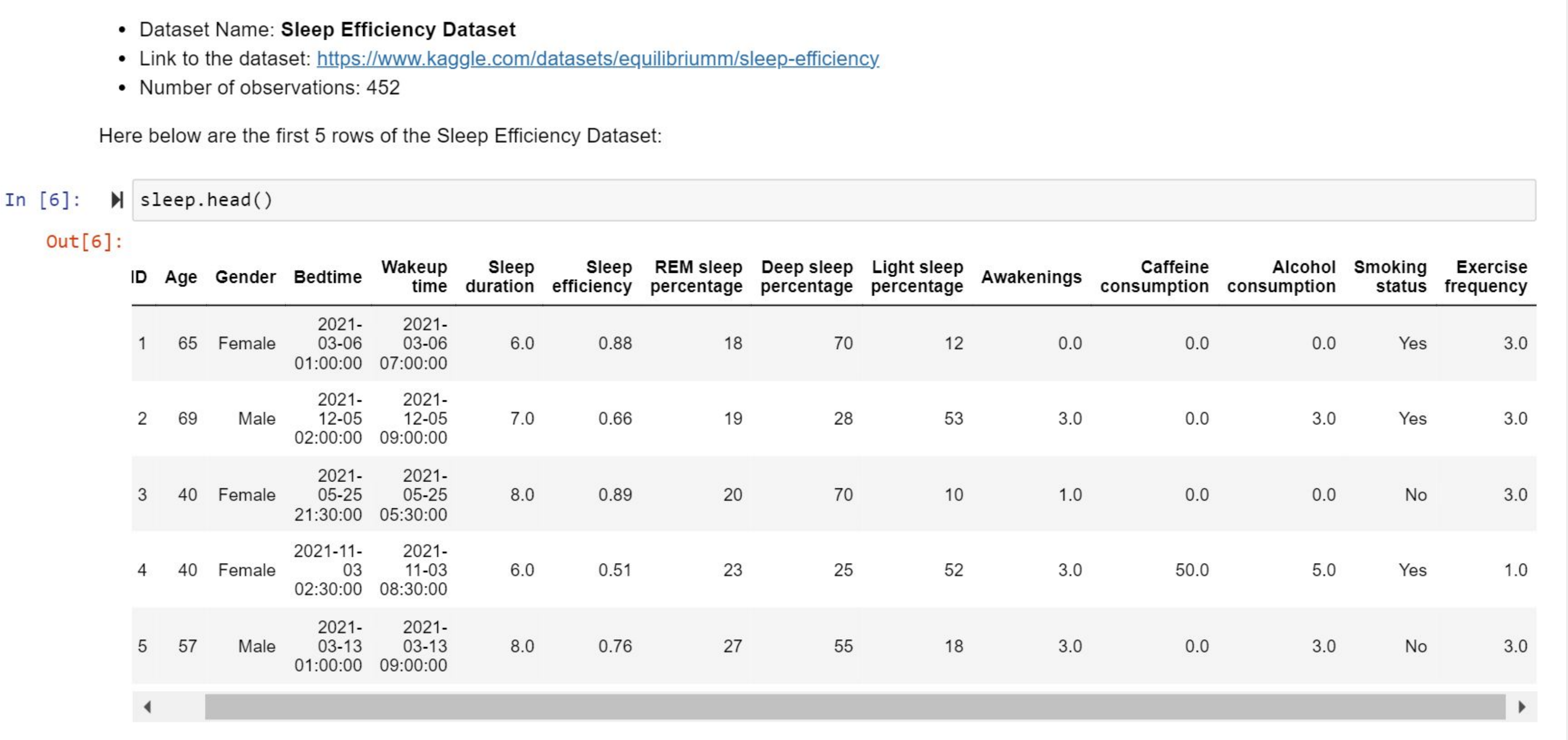
Hypothesis

Utilizing data on an individual's lifestyle habits & demographics in addition to observations on his/her sleep patterns, we can create a model that gives each individual a "sleep score" that measures one's quality of sleep.

Hypothesis

For demographic factors, we predict that older individuals will have a lower sleep score & that gender will not have a significant impact on one's sleep score. We predict that people with "negative" lifestyle habits such as substance use of caffeine or lack of exercise will have a lower sleep score. Out of the two categories, we predict that lifestyle habits will provide a more accurate prediction of an individual's sleep efficiency when compared to demographic factors.

Dataset



- Dataset Name: **Sleep Efficiency Dataset**
- Link to the dataset: https://www.kaggle.com/datasets/equilibriumm/sleep-efficiency
- Number of observations: 452

Here below are the first 5 rows of the Sleep Efficiency Dataset:

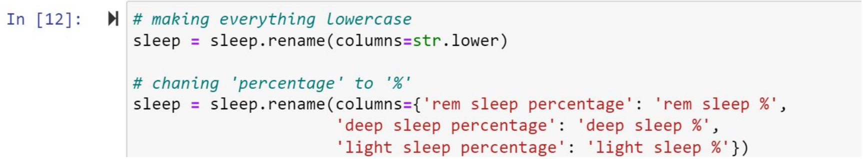
```
In [6]:  sleep.head()
```

```
Out[6]:
```

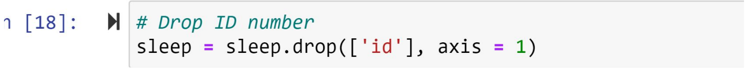
ID	Age	Gender	Bedtime	Wakeup time	Sleep duration	Sleep efficiency	REM sleep percentage	Deep sleep percentage	Light sleep percentage	Awakenings	Caffeine consumption	Alcohol consumption	Smoking status	Exercise frequency
1	65	Female	2021-03-06 01:00:00	2021-03-06 07:00:00	6.0	0.88	18	70	12	0.0	0.0	0.0	Yes	3.0
2	69	Male	2021-12-05 02:00:00	2021-12-05 09:00:00	7.0	0.66	19	28	53	3.0	0.0	3.0	Yes	3.0
3	40	Female	2021-05-25 21:30:00	2021-05-25 05:30:00	8.0	0.89	20	70	10	1.0	0.0	0.0	No	3.0
4	40	Female	2021-11-03 02:30:00	2021-11-03 08:30:00	6.0	0.51	23	25	52	3.0	50.0	5.0	Yes	1.0
5	57	Male	2021-03-13 01:00:00	2021-03-13 09:00:00	8.0	0.76	27	55	18	3.0	0.0	3.0	No	3.0

Data Cleaning

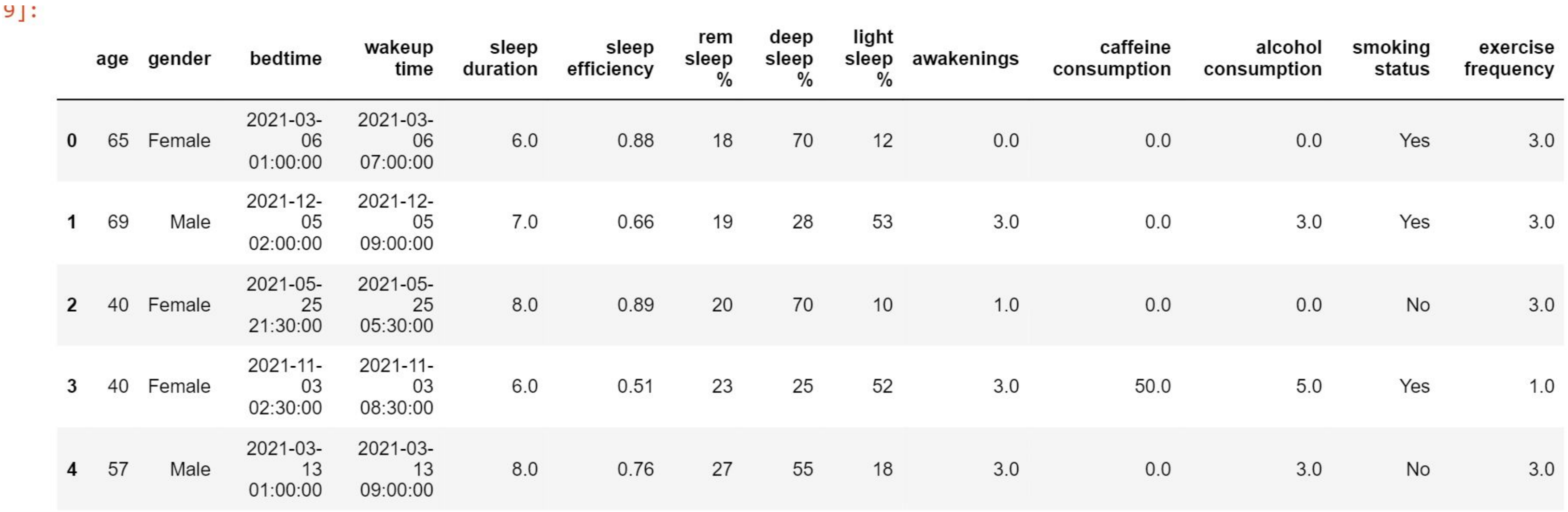
Set Up & Cleaning : Columns



```
In [12]:  # making everything lowercase
          sleep = sleep.rename(columns=str.lower)

          # chaning 'percentage' to '%'
          sleep = sleep.rename(columns={'rem sleep percentage': 'rem sleep %',
                                        'deep sleep percentage': 'deep sleep %',
                                        'light sleep percentage': 'light sleep %'})
```



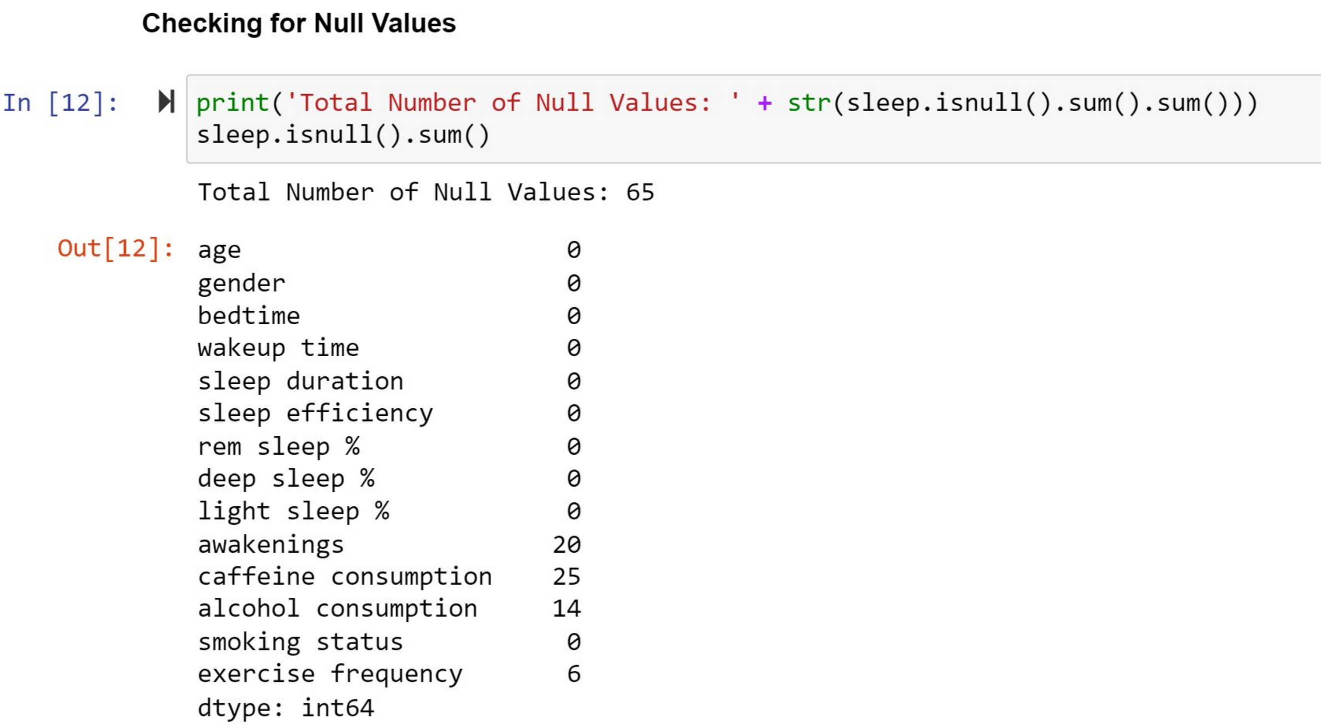
```
n [18]:  # Drop ID number
         sleep = sleep.drop(['id'], axis = 1)
```



9]:

	age	gender	bedtime	wakeup time	sleep duration	sleep efficiency	rem sleep %	deep sleep %	light sleep %	awakenings	caffeine consumption	alcohol consumption	smoking status	exercise frequency
0	65	Female	2021-03-06 01:00:00	2021-03-06 07:00:00	6.0	0.88	18	70	12	0.0	0.0	0.0	Yes	3.0
1	69	Male	2021-12-05 02:00:00	2021-12-05 09:00:00	7.0	0.66	19	28	53	3.0	0.0	3.0	Yes	3.0
2	40	Female	2021-05-25 21:30:00	2021-05-25 05:30:00	8.0	0.89	20	70	10	1.0	0.0	0.0	No	3.0
3	40	Female	2021-11-03 02:30:00	2021-11-03 08:30:00	6.0	0.51	23	25	52	3.0	50.0	5.0	Yes	1.0
4	57	Male	2021-03-13 01:00:00	2021-03-13 09:00:00	8.0	0.76	27	55	18	3.0	0.0	3.0	No	3.0

Set Up & Cleaning : Dropping Null Values

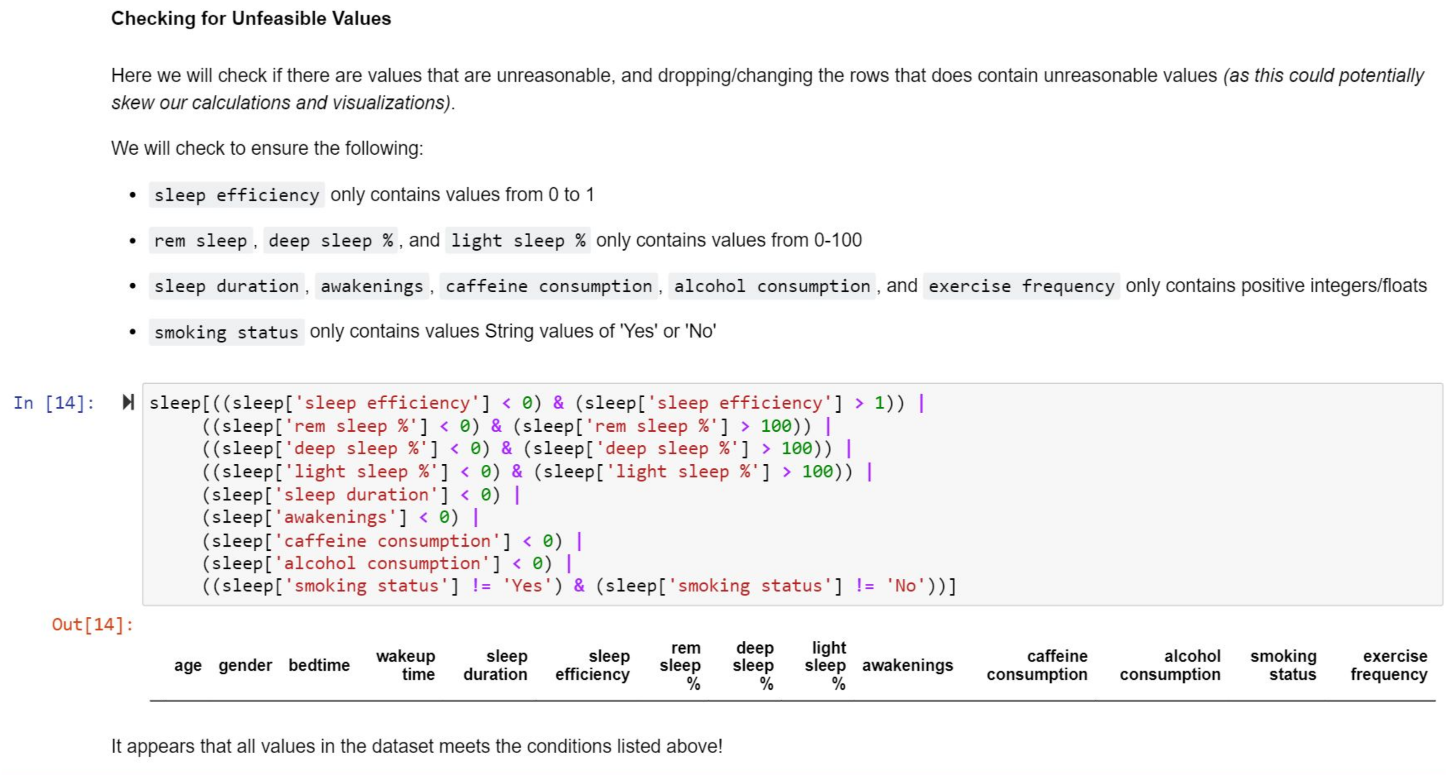


Checking for Null Values

```
In [12]:  print('Total Number of Null Values: ' + str(sleep.isnull().sum().sum()))
          sleep.isnull().sum()
```

```
          Total Number of Null Values: 65

Out[12]:  age                       0
          gender                    0
          bedtime                   0
          wakeup time               0
          sleep duration            0
          sleep efficiency          0
          rem sleep %               0
          deep sleep %              0
          light sleep %             0
          awakenings               20
          caffeine consumption     25
          alcohol consumption      14
          smoking status            0
          exercise frequency        6
          dtype: int64
```

Set Up & Cleaning : Unfeasible Data



Checking for Unfeasible Values

Here we will check if there are values that are unreasonable, and dropping/changing the rows that does contain unreasonable values *(as this could potentially skew our calculations and visualizations)*.

We will check to ensure the following:

- `sleep efficiency` only contains values from 0 to 1
- `rem sleep`, `deep sleep %`, and `light sleep %` only contains values from 0-100
- `sleep duration`, `awakenings`, `caffeine consumption`, `alcohol consumption`, and `exercise frequency` only contains positive integers/floats
- `smoking status` only contains values String values of 'Yes' or 'No'

In [14]:

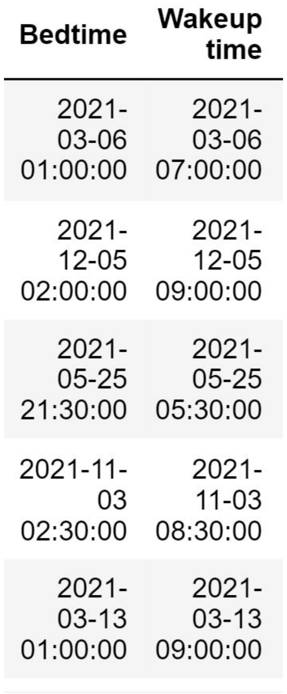
```
sleep[((sleep['sleep efficiency'] < 0) & (sleep['sleep efficiency'] > 1)) |
     ((sleep['rem sleep %'] < 0) & (sleep['rem sleep %'] > 100)) |
     ((sleep['deep sleep %'] < 0) & (sleep['deep sleep %'] > 100)) |
     ((sleep['light sleep %'] < 0) & (sleep['light sleep %'] > 100)) |
     (sleep['sleep duration'] < 0) |
     (sleep['awakenings'] < 0) |
     (sleep['caffeine consumption'] < 0) |
     (sleep['alcohol consumption'] < 0) |
     ((sleep['smoking status'] != 'Yes') & (sleep['smoking status'] != 'No'))]
```

Out[14]:

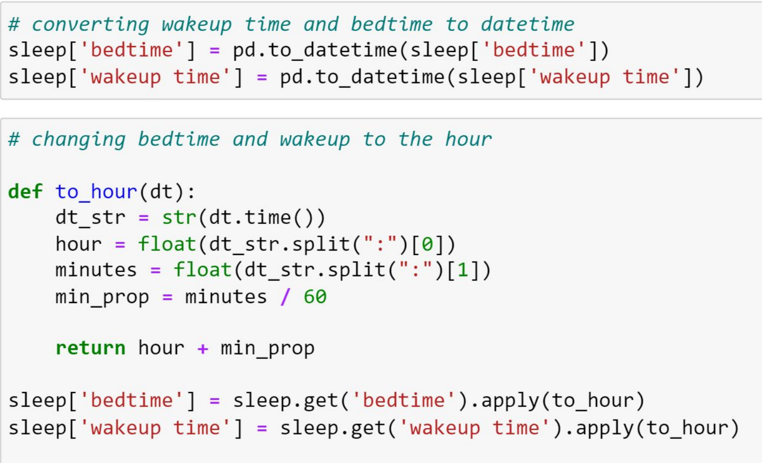
age	gender	bedtime	wakeup time	sleep duration	sleep efficiency	rem sleep %	deep sleep %	light sleep %	awakenings	caffeine consumption	alcohol consumption	smoking status	exercise frequency

It appears that all values in the dataset meets the conditions listed above!

Set Up & Cleaning : bedtime and wakeup time

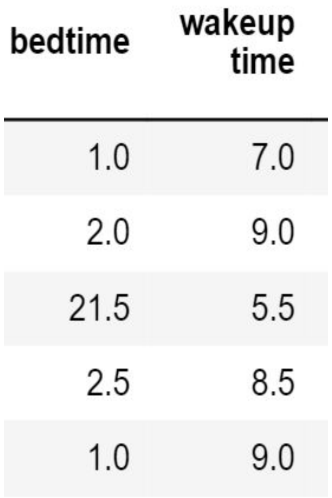


Bedtime	Wakeup time
2021-03-06 01:00:00	2021-03-06 07:00:00
2021-12-05 02:00:00	2021-12-05 09:00:00
2021-05-25 21:30:00	2021-05-25 05:30:00
2021-11-03 02:30:00	2021-11-03 08:30:00
2021-03-13 01:00:00	2021-03-13 09:00:00



```python
# converting wakeup time and bedtime to datetime
sleep['bedtime'] = pd.to_datetime(sleep['bedtime'])
sleep['wakeup time'] = pd.to_datetime(sleep['wakeup time'])
```

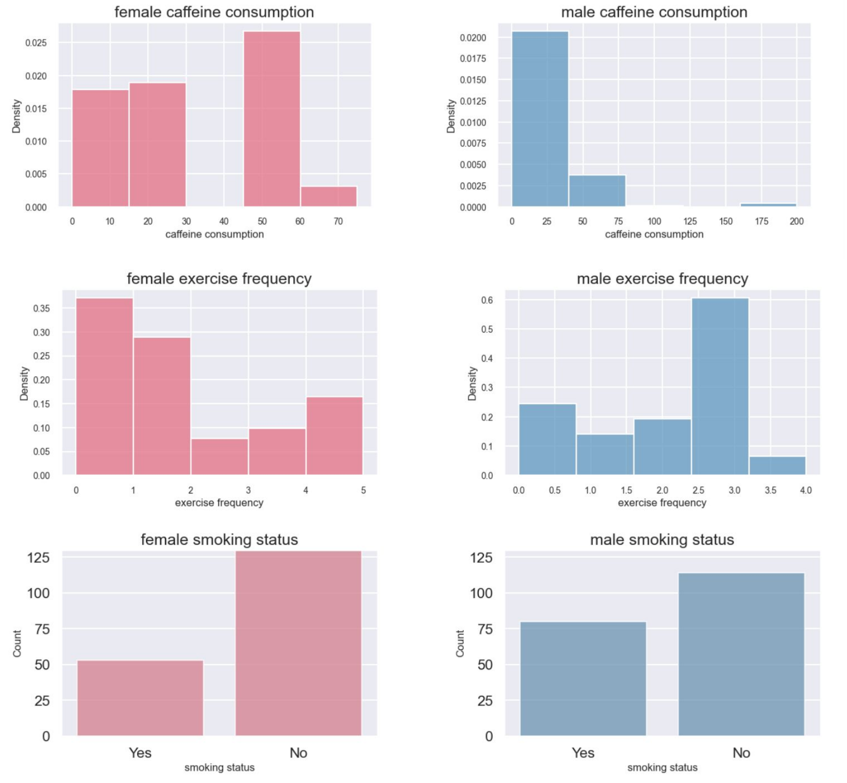
```python
# changing bedtime and wakeup to the hour

def to_hour(dt):
    dt_str = str(dt.time())
    hour = float(dt_str.split(":")[0])
    minutes = float(dt_str.split(":")[1])
    min_prop = minutes / 60

    return hour + min_prop

sleep['bedtime'] = sleep.get('bedtime').apply(to_hour)
sleep['wakeup time'] = sleep.get('wakeup time').apply(to_hour)
```



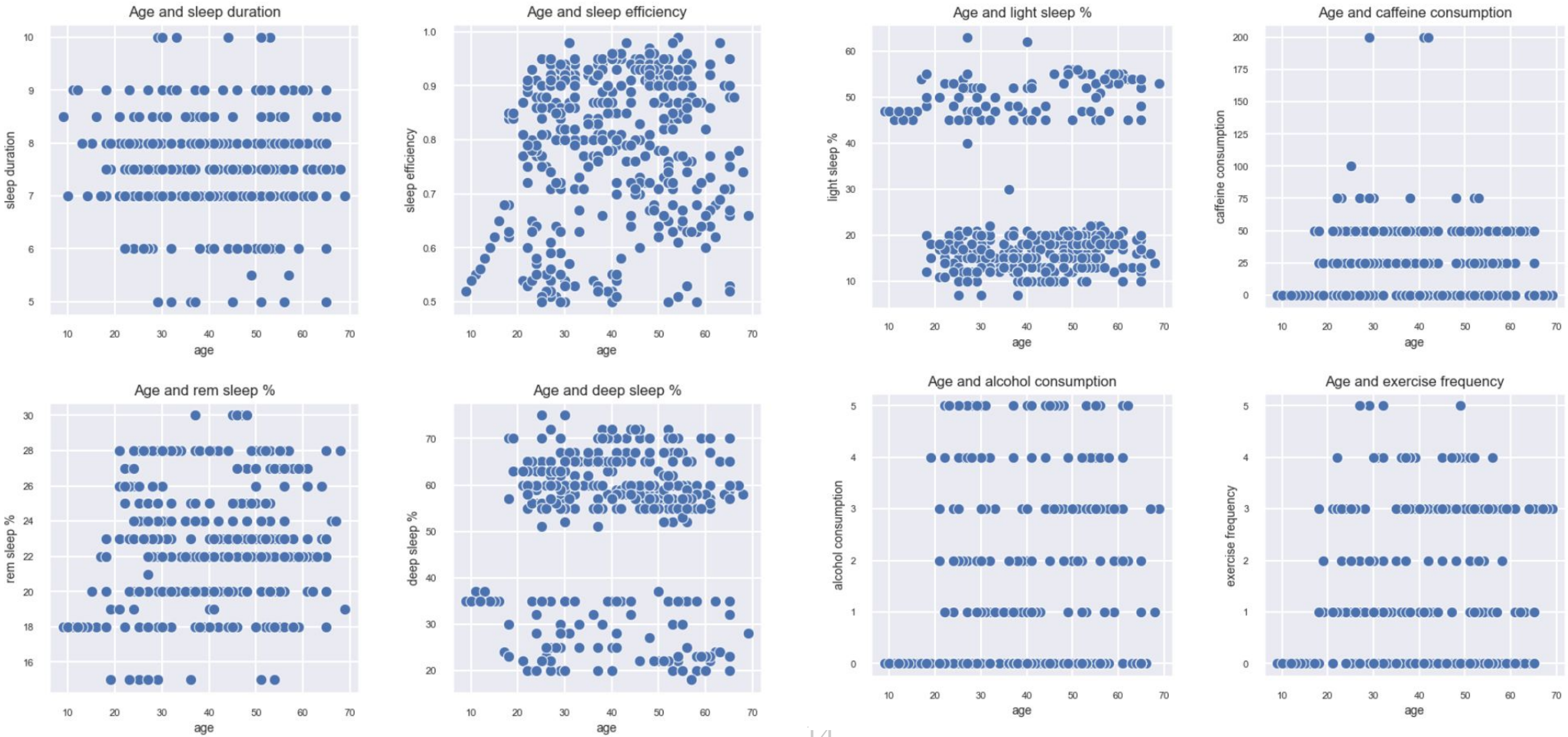
bedtime	wakeup time
1.0	7.0
2.0	9.0
21.5	5.5
2.5	8.5
1.0	9.0

Data Visualization and Analysis

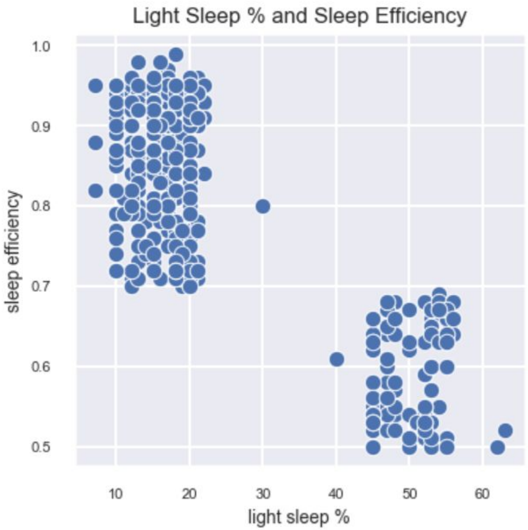
Data Analysis & Visualization : Gender



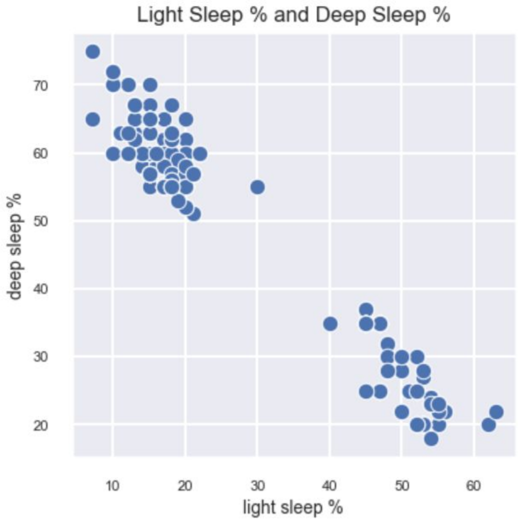
Data Analysis & Visualization : Age



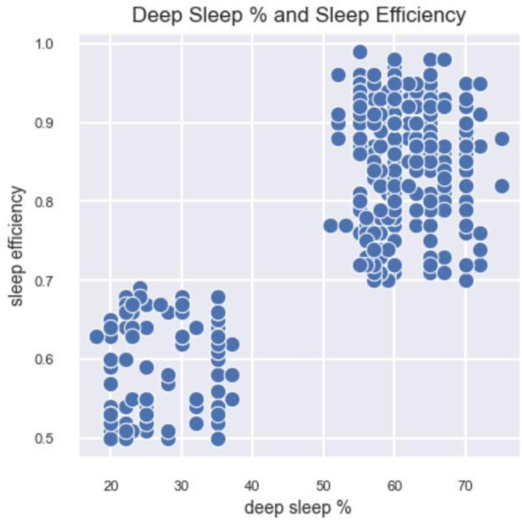
Data Analysis & Visualization : Sleep


Light Sleep % and Sleep Efficiency

Correlation: -0.82


Light Sleep % and Deep Sleep %

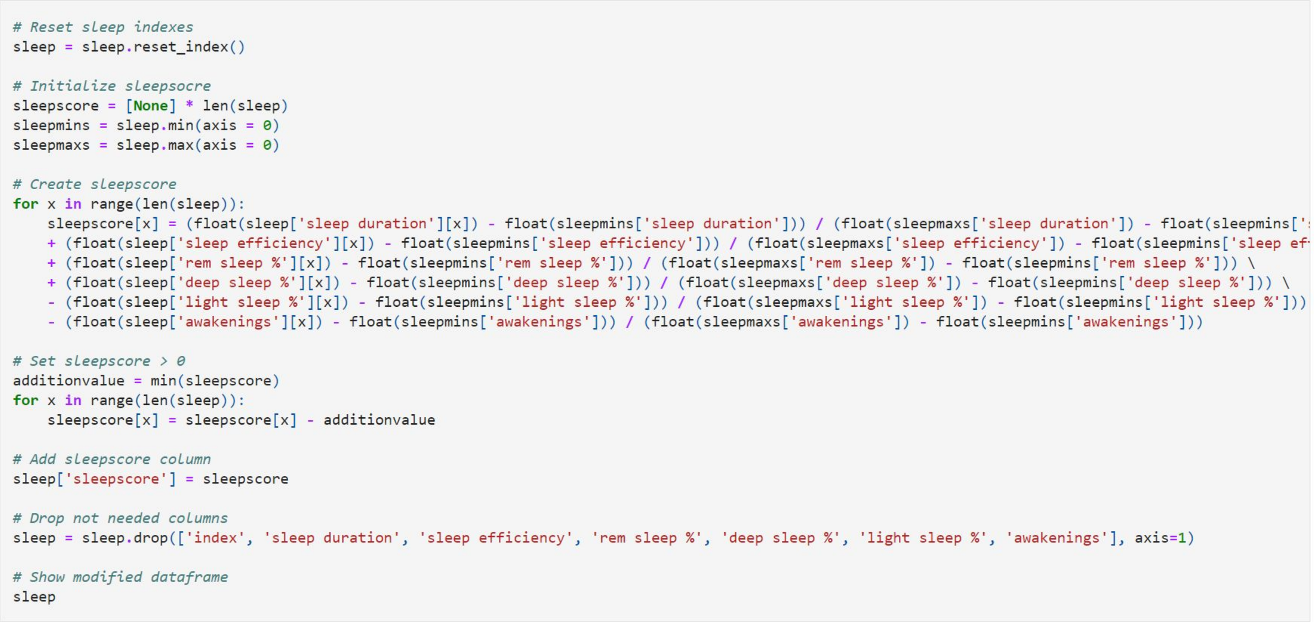
Correlation: -0.98


Deep Sleep % and Sleep Efficiency

Correlation: 0.79

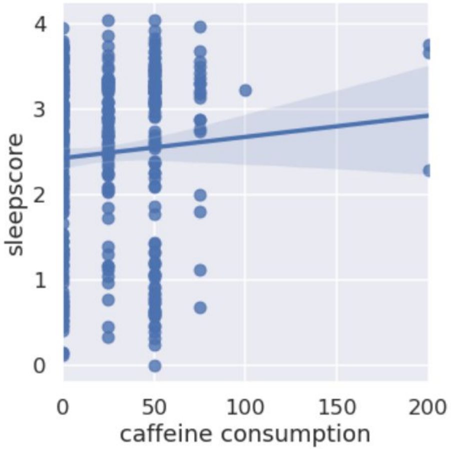
Data Analysis & Visualization

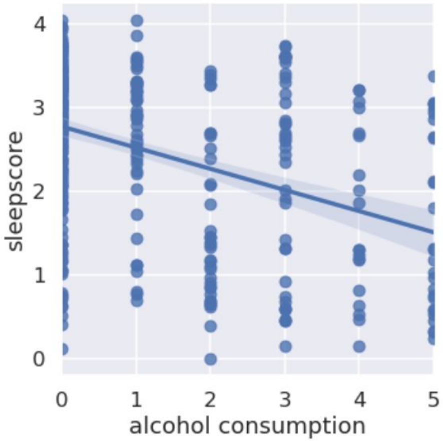
Creating the Sleep Score

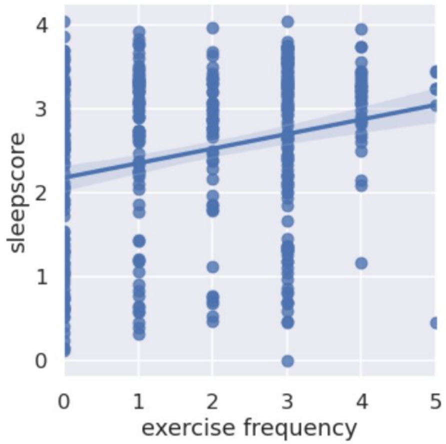


```python
# Reset sleep indexes
sleep = sleep.reset_index()

# Initialize sleepsocre
sleepscore = [None] * len(sleep)
sleepmins = sleep.min(axis = 0)
sleepmaxs = sleep.max(axis = 0)

# Create sleepscore
for x in range(len(sleep)):
    sleepscore[x] = (float(sleep['sleep duration'][x]) - float(sleepmins['sleep duration'])) / (float(sleepmaxs['sleep duration']) - float(sleepmins['
    + (float(sleep['sleep efficiency'][x]) - float(sleepmins['sleep efficiency'])) / (float(sleepmaxs['sleep efficiency']) - float(sleepmins['sleep ef
    + (float(sleep['rem sleep %'][x]) - float(sleepmins['rem sleep %'])) / (float(sleepmaxs['rem sleep %']) - float(sleepmins['rem sleep %'])) \
    + (float(sleep['deep sleep %'][x]) - float(sleepmins['deep sleep %'])) / (float(sleepmaxs['deep sleep %']) - float(sleepmins['deep sleep %'])) \
    - (float(sleep['light sleep %'][x]) - float(sleepmins['light sleep %'])) / (float(sleepmaxs['light sleep %']) - float(sleepmins['light sleep %']))
    - (float(sleep['awakenings'][x]) - float(sleepmins['awakenings'])) / (float(sleepmaxs['awakenings']) - float(sleepmins['awakenings']))

# Set sleepscore > 0
additionvalue = min(sleepscore)
for x in range(len(sleep)):
    sleepscore[x] = sleepscore[x] - additionvalue

# Add sleepscore column
sleep['sleepscore'] = sleepscore

# Drop not needed columns
sleep = sleep.drop(['index', 'sleep duration', 'sleep efficiency', 'rem sleep %', 'deep sleep %', 'light sleep %', 'awakenings'], axis=1)

# Show modified dataframe
sleep
```

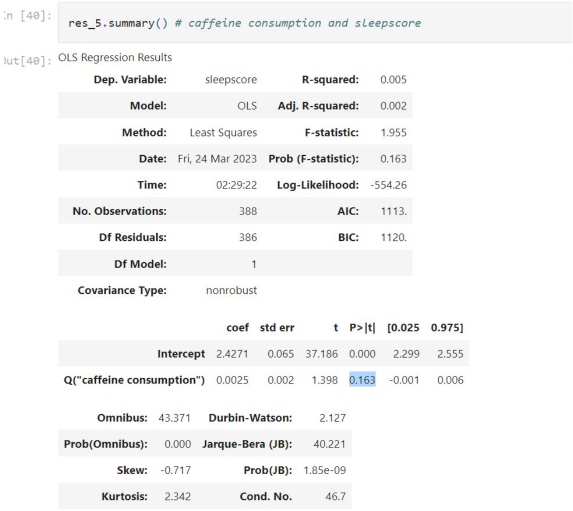
Data Analysis & Visualization







Data Analysis & Visualization



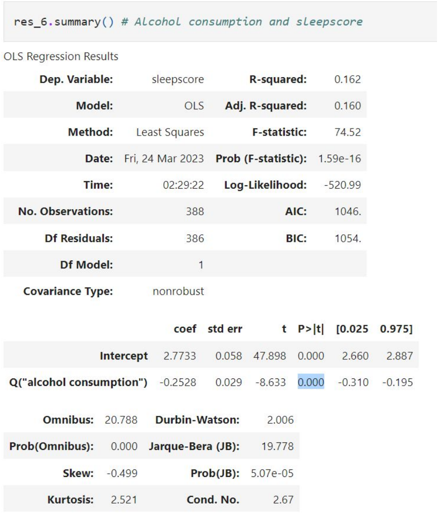
```
n [40]: res_5.summary() # caffeine consumption and sleepscore
```

ut[40]: OLS Regression Results

Dep. Variable:	sleepscore	**R-squared:**	0.005
Model:	OLS	**Adj. R-squared:**	0.002
Method:	Least Squares	**F-statistic:**	1.955
Date:	Fri, 24 Mar 2023	**Prob (F-statistic):**	0.163
Time:	02:29:22	**Log-Likelihood:**	-554.26
No. Observations:	388	**AIC:**	1113.
Df Residuals:	386	**BIC:**	1120.
Df Model:	1		
Covariance Type:	nonrobust		

	coef	**std err**	**t**	**P>|t|**	**[0.025**	**0.975]**
Intercept	2.4271	0.065	37.186	0.000	2.299	2.555
Q("caffeine consumption")	0.0025	0.002	1.398	0.163	-0.001	0.006

Omnibus:	43.371	**Durbin-Watson:**	2.127
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	40.221
Skew:	-0.717	**Prob(JB):**	1.85e-09
Kurtosis:	2.342	**Cond. No.**	46.7



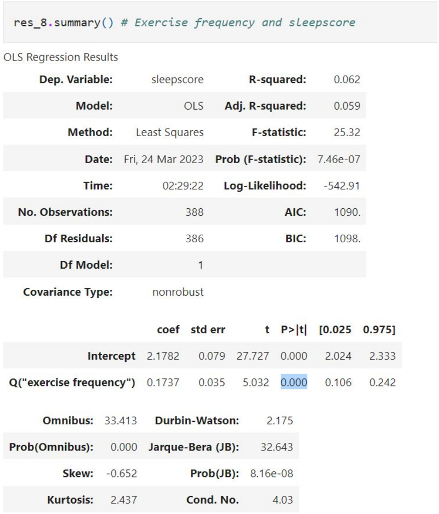
```
res_6.summary() # Alcohol consumption and sleepscore
```

OLS Regression Results

Dep. Variable:	sleepscore	**R-squared:**	0.162
Model:	OLS	**Adj. R-squared:**	0.160
Method:	Least Squares	**F-statistic:**	74.52
Date:	Fri, 24 Mar 2023	**Prob (F-statistic):**	1.59e-16
Time:	02:29:22	**Log-Likelihood:**	-520.99
No. Observations:	388	**AIC:**	1046.
Df Residuals:	386	**BIC:**	1054.
Df Model:	1		
Covariance Type:	nonrobust		

	coef	**std err**	**t**	**P>|t|**	**[0.025**	**0.975]**
Intercept	2.7733	0.058	47.898	0.000	2.660	2.887
Q("alcohol consumption")	-0.2528	0.029	-8.633	0.000	-0.310	-0.195

Omnibus:	20.788	**Durbin-Watson:**	2.006
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	19.778
Skew:	-0.499	**Prob(JB):**	5.07e-05
Kurtosis:	2.521	**Cond. No.**	2.67



```
res_8.summary() # Exercise frequency and sleepscore
```

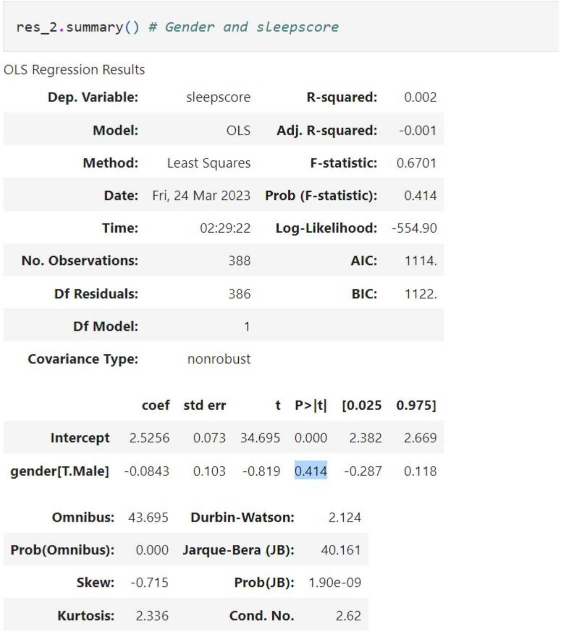
OLS Regression Results

Dep. Variable:	sleepscore	**R-squared:**	0.062
Model:	OLS	**Adj. R-squared:**	0.059
Method:	Least Squares	**F-statistic:**	25.32
Date:	Fri, 24 Mar 2023	**Prob (F-statistic):**	7.46e-07
Time:	02:29:22	**Log-Likelihood:**	-542.91
No. Observations:	388	**AIC:**	1090.
Df Residuals:	386	**BIC:**	1098.
Df Model:	1		
Covariance Type:	nonrobust		

	coef	**std err**	**t**	**P>|t|**	**[0.025**	**0.975]**
Intercept	2.1782	0.079	27.727	0.000	2.024	2.333
Q("exercise frequency")	0.1737	0.035	5.032	0.000	0.106	0.242

Omnibus:	33.413	**Durbin-Watson:**	2.175
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	32.643
Skew:	-0.652	**Prob(JB):**	8.16e-08
Kurtosis:	2.437	**Cond. No.**	4.03

Data Analysis & Visualization



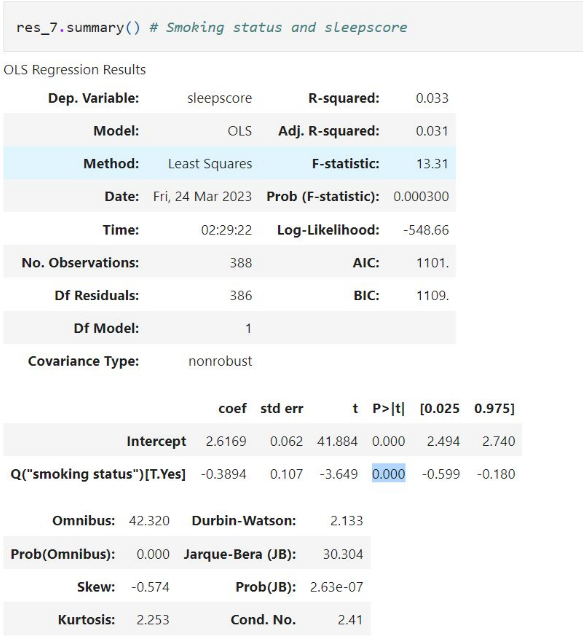
```
res_2.summary() # Gender and sleepscore
```

OLS Regression Results

Dep. Variable:	sleepscore	**R-squared:**	0.002
Model:	OLS	**Adj. R-squared:**	-0.001
Method:	Least Squares	**F-statistic:**	0.6701
Date:	Fri, 24 Mar 2023	**Prob (F-statistic):**	0.414
Time:	02:29:22	**Log-Likelihood:**	-554.90
No. Observations:	388	**AIC:**	1114.
Df Residuals:	386	**BIC:**	1122.
Df Model:	1		
Covariance Type:	nonrobust		

	coef	**std err**	**t**	**P>|t|**	**[0.025**	**0.975]**
Intercept	2.5256	0.073	34.695	0.000	2.382	2.669
gender[T.Male]	-0.0843	0.103	-0.819	0.414	-0.287	0.118

Omnibus:	43.695	**Durbin-Watson:**	2.124
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	40.161
Skew:	-0.715	**Prob(JB):**	1.90e-09
Kurtosis:	2.336	**Cond. No.**	2.62



```
res_7.summary() # Smoking status and sleepscore
```

OLS Regression Results

Dep. Variable:	sleepscore	**R-squared:**	0.033
Model:	OLS	**Adj. R-squared:**	0.031
Method:	Least Squares	**F-statistic:**	13.31
Date:	Fri, 24 Mar 2023	**Prob (F-statistic):**	0.000300
Time:	02:29:22	**Log-Likelihood:**	-548.66
No. Observations:	388	**AIC:**	1101.
Df Residuals:	386	**BIC:**	1109.
Df Model:	1		
Covariance Type:	nonrobust		

	coef	**std err**	**t**	**P>|t|**	**[0.025**	**0.975]**
Intercept	2.6169	0.062	41.884	0.000	2.494	2.740
Q("smoking status")[T.Yes]	-0.3894	0.107	-3.649	0.000	-0.599	-0.180

Omnibus:	42.320	**Durbin-Watson:**	2.133
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	30.304
Skew:	-0.574	**Prob(JB):**	2.63e-07
Kurtosis:	2.253	**Cond. No.**	2.41

Results & Conclusion

It can be said that more exercise and an earlier bedtime may have a positive correlation with one's quality of sleep. It can also be said that smoking and consuming more alcohol may have a negative correlation with one's quality of sleep.

Results & Conclusion

Looking back at our hypothesis, we had varying correctness. Overall, we were correct in that lifestyle habits provided a more accurate prediction of an individual's sleep score when compared to demographic factors. While substance use of caffeine did not have a strong correlation with sleep score, bedtime, exercise, smoking status, and alcohol all showed strong correlations either negative or positive.

Ethics & Privacy

- Publicly available data
    - Personal information (tobacco/alcohol consumption) BUT subjects indicated with ID numbers for privacy
- Potential bias: data collection region
- Future considerations: more detailed data on other factors that could be responsible for sleep efficiency
    - Ex: underlying health conditions, sleep environment