Data Wrangling and Exploration

By: Cameron Hartling, Jesse Jacobs, Timothy Perez, Jared Ponce

Data Wrangling

The data set we picked was published with all the data structured and enriched. Five features of the table have missing values, these features are awakening, caffeine consumption, alcohol consumption, smoking status, and exercise frequency. We didn't have many missing values. Awakenings and Exercise frequency have been filled with the average while Caffeine and Alcohol consumption have a rounded average. The rounded average is to stay the same as the other values that are whole numbers. Bedtime and Wakeup time were split into Bedtime, Bed date, Wakeup time, and Wakeup date. This allows for Data Enrichment where we can now see what Time of day they went to sleep.

```
Age                       0
Gender                    0
Bedtime                   0
Wakeup time               0
Sleep duration            0
Sleep efficiency          0
REM sleep percentage      0
Deep sleep percentage     0
Light sleep percentage    0
Awakenings               20
Caffeine consumption     25
Alcohol consumption      14
Smoking status            0
Exercise frequency        6
dtype: int64
```

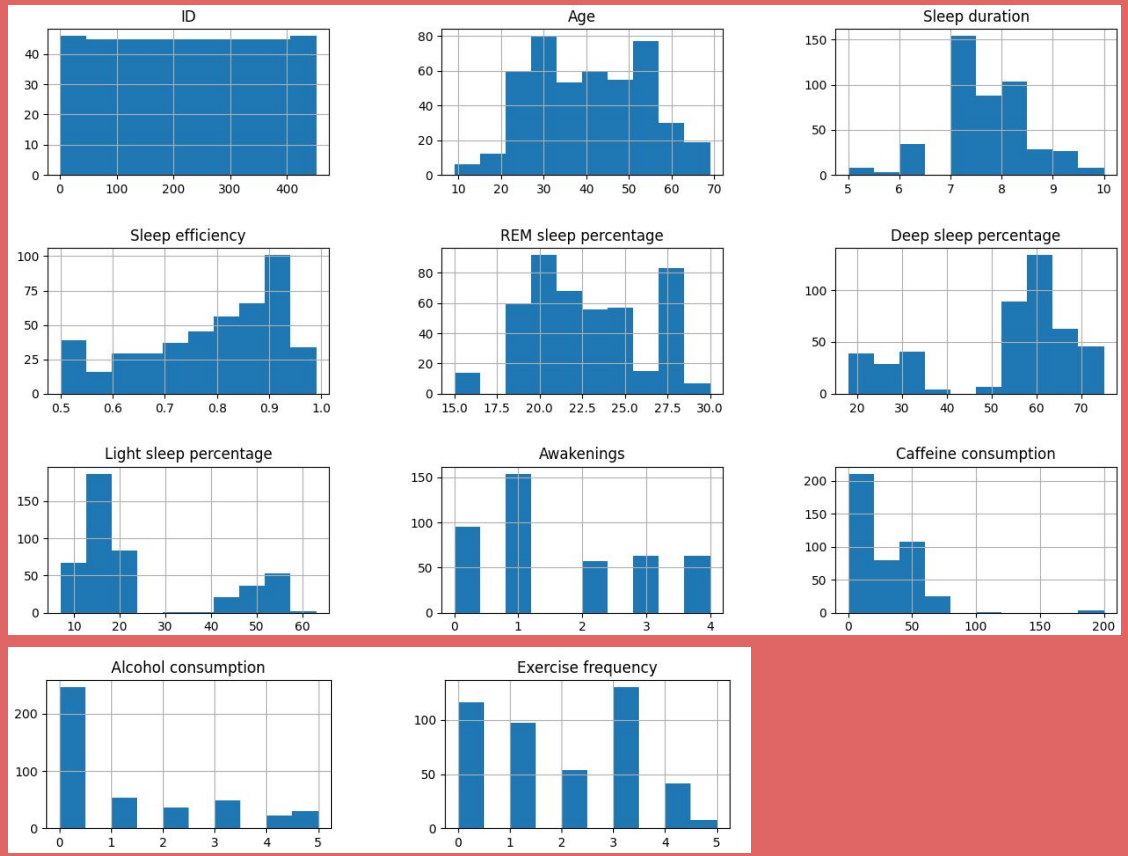
Number of null entries

```
Int64Index: 452 entries, 66 to 134
Data columns (total 17 columns):
 #   Column                  Non-Null Count  Dtype
---  ------                  --------------  -----
 0   Age                     452 non-null    int64
 1   Gender                  452 non-null    object
 2   Bedtime                 452 non-null    object
 3   Wakeup time             452 non-null    object
 4   Sleep duration          452 non-null    float64
 5   Sleep efficiency        452 non-null    float64
 6   REM sleep percentage    452 non-null    int64
 7   Deep sleep percentage   452 non-null    int64
 8   Light sleep percentage  452 non-null    int64
 9   Awakenings              452 non-null    float64
 10  Caffeine consumption    452 non-null    float64
 11  Alcohol consumption     452 non-null    float64
 12  Smoking status          452 non-null    object
 13  Exercise frequency      452 non-null    float64
 14  Bed date                452 non-null    object
 15  Wakeup date             452 non-null    object
 16  Time of day             452 non-null    object
dtypes: float64(6), int64(4), object(7)
```

Results after the changes

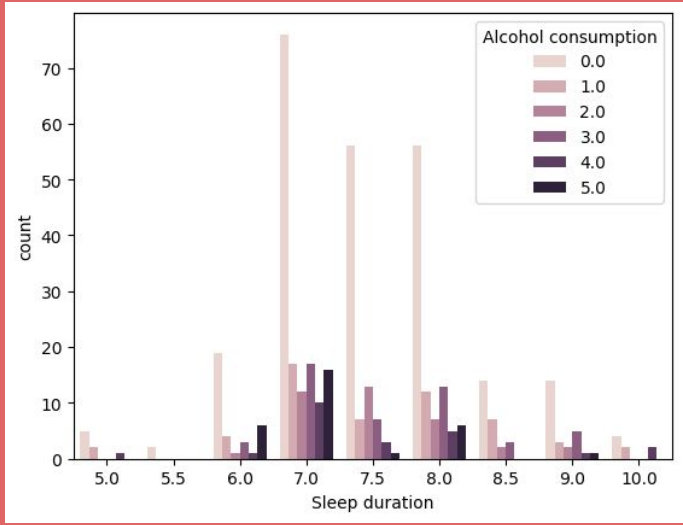
Exploration Overview

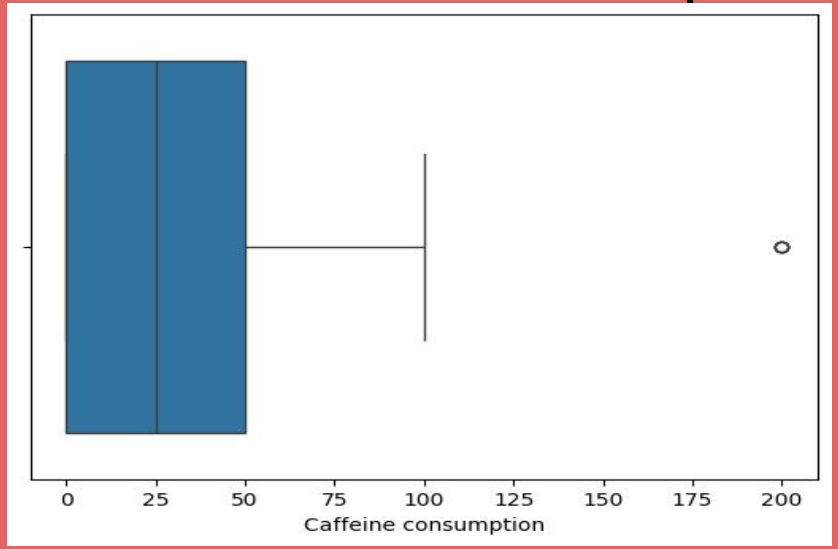
This Data set is comprised of eleven features that include continuous and categorical data. This general overview of all the features was made using the .hist() function.



Exploration Alcohol Consumption & Caffeine Consumption

We found that the data for sleep duration based on alcohol consumption followed a normal curve regardless of the alcohol consumed before bed.





We discovered an outlier for caffeine consumption for this data set at 200mg using the boxplot function.

Exploration Smoking Status & Exercise Frequency

Exercise frequency	ID	Age	Sleep duration	Sleep efficiency	REM sleep percentage	Deep sleep percentage	Light sleep percentage	Awakenings	Caffeine consumption	Alcohol consumption
0.0	222.845455	41.790909	7.563636	0.744818	21.872727	48.309091	29.818182	1.963636	16.363636	1.209091
1.0	235.038462	35.974359	7.391026	0.775128	23.923077	53.102564	22.974359	1.666667	41.025641	0.961538
2.0	223.533333	35.955556	7.211111	0.810222	21.666667	56.866667	21.466667	1.711111	32.222222	1.177778
3.0	220.477876	45.707965	7.539823	0.799204	23.247788	53.265487	23.486726	1.566372	13.053097	1.221239
4.0	229.200000	41.028571	7.228571	0.890286	22.114286	59.942857	17.942857	0.771429	22.142857	1.114286
5.0	278.285714	31.428571	7.571429	0.844286	21.714286	51.857143	26.428571	0.142857	14.285714	1.000000

In the above graph various features' means were explored with respect to exercise frequency. There are certain relationships that can be explored between a feature and how often one exercises.

Smoking status	ID	Age	Sleep duration	Sleep efficiency	REM sleep percentage	Deep sleep percentage	Light sleep percentage	Awakenings	Caffeine consumption	Alcohol consumption	Exercise frequency
No	227.462745	40.666667	7.449020	0.817922	22.584314	55.215686	22.200000	1.623529	22.156863	1.054902	1.784314
Yes	223.954887	41.142857	7.454887	0.734286	22.864662	48.233083	28.902256	1.609023	23.684211	1.323308	1.706767

In the above graph certain features' means were explored with respect to smoking status. There are certain relationships that can be explored between a feature and whether an individual smokes or not.