Data Collection and Preprocessing Phase

Date	20 June 2024
Team ID	739813
Project Title	Optimizing Sleep Efficiency: Harnessing Machine Learning For Enhanced Restorative Rest
Maximum Marks	6 Marks

Data Exploration and Preprocessing Template

Identifies data sources, assesses quality issues like missing values and duplicates, and implements resolution plans to ensure accurate and reliable analysis.

Section	Description
Data Overview	Basic statistics, dimensions, and structure of the data. <class 'pandas.core.frame.DataFrame'> RangeIndex: 452 entries, 0 to 451 Data columns (total 15 columns): # Column Non-Null Count Dtype --- ------ -------------- ----- 0 ID 452 non-null int64 1 Age 452 non-null int64 2 Gender 452 non-null object 3 Bedtime 452 non-null object 4 Wakeup time 452 non-null object 5 Sleep duration 452 non-null float64 6 Sleep efficiency 452 non-null float64 7 REM sleep percentage 452 non-null int64

	8 Deep sleep percentage 452 non-null int64 9 Light sleep percentage 452 non-null int64 10 Awakenings 432 non-null float64 11 Caffeine consumption 427 non-null float64 12 Alcohol consumption 438 non-null float64 13 Smoking status 452 non-null object 14 Exercise frequency 446 non-null float64 dtypes: float64(6), int64(5), object(4) memory usage: 53.1+ KB
Univariate Analysis	Exploration of individual variables (mean, median, mode, etc.). ID Age Sleep duration Sleep efficiency REM sleep percentage Deep sleep percentageLight sleep percentage Awakenings Caffeine consumption Alcohol consumption Exercise frequency count 452.000000 452.000000 452.000000 452.000000 452.000000 452.000000 452.000000 452.000000 452.000000 452.000000 mean 226.500000 40.285398 7.465708 0.788916 22.615044 52.823009 24.561947 1.641204 23.653396 1.173516 1.791480 std 130.625419 13.172250 0.866625 0.135237 3.525963 15.654235 15.313665 1.326338 29.353745 1.596013 1.418603 min 1.000000 9.000000 5.000000 0.500000 15.000000 18.000000 7.000000 0.000000 0.000000 0.000000 0.000000 25% 113.750000 29.000000 7.000000 0.697500 20.000000 48.250000 15.000000 1.000000 0.000000 0.000000 0.000000 50% 226.500000 40.000000 7.500000 0.820000 22.000000 58.000000 18.000000 1.000000 23.653396 0.000000 2.000000 75% 339.250000 52.000000 8.000000 0.900000 25.000000 63.000000 32.500000 3.000000 50.000000 2.000000 3.000000 max 452.000000 69.000000 10.000000 0.990000 30.000000 75.000000 63.000000 4.000000 200.000000 5.000000 5.000000ID Age Sleep duration Sleep efficiency REM sleep percentage Deep sleep percentageLight sleep percentage

	Awakenings	Caffeine consumption	Alcohol consumption	Exercise frequency
count	452.000000	452.000000	452.000000	
	452.000000	452.000000	452.000000	
	452.000000	452.000000	452.000000	
	452.000000	452.000000		
mean	226.500000	40.285398	7.465708	0.788916
	22.615044	52.823009	24.561947	1.641204
	23.653396	1.173516	1.791480	
std	130.625419	13.172250	0.866625	0.135237
	3.525963	15.654235	15.313665	1.326338
	29.353745	1.596013	1.418603	
min	1.000000	9.000000	5.000000	0.500000
	15.000000	18.000000	7.000000	0.000000
	0.000000	0.000000	0.000000	
25%	113.750000	29.000000	7.000000	0.697500
	20.000000	48.250000	15.000000	1.000000
	0.000000	0.000000	0.000000	
50%	226.500000	40.000000	7.500000	0.820000
	22.000000	58.000000	18.000000	1.000000
	23.653396	0.000000	2.000000	
75%	339.250000	52.000000	8.000000	0.900000
	25.000000	63.000000	32.500000	3.000000
	50.000000	2.000000	3.000000	
max	452.000000	69.000000	10.000000	0.990000
	30.000000	75.000000	63.000000	4.000000
	200.000000	5.000000	5.000000	

Bivariate Analysis	Relationships between two variables (correlation, scatter plots).

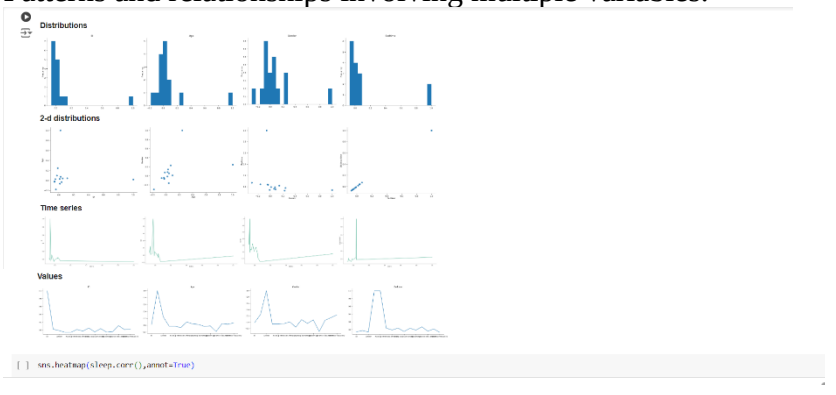
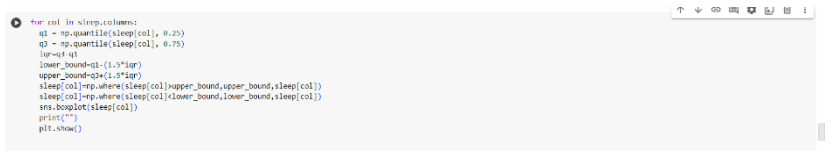
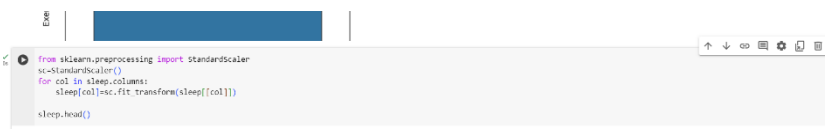
ID	Age	Gender	Bedtime	Wakeuptime	Sleepduration	Sleepefficiency	REMsleeppercentage	Deepsleeppercentage	Lightsleeppercentage	Awakenings
ID	1.000000	0.021310	-0.013124	-0.059307	-0.058310	0.014736	-0.025624	0.051636	-0.047507	0.037160
Age	0.021310	1.000000	0.244468	-0.024805	-0.025337	-0.065455	0.098357	0.042091	0.025800	-0.031354
Gender	-0.013124	0.244468	1.000000	-0.062276	-0.062457	-0.046941	0.010062	-0.161102	0.082634	-0.039541
Bedtime	-0.059307	-0.024805	-0.062276	1.000000	0.999929	0.038189	-0.009028	0.038519	-0.049915	0.034731
Wakeuptime	-0.058310	-0.025337	-0.062457	0.999929	1.000000	0.040773	-0.009607	0.037928	-0.050076	0.034976

1023.45833

1028.20833

	Sleep duration	0.014736	-0.065455	-0.046941	0.038189	0.040773	1.000000	-0.025352	-0.018780	-0.035441	0.037799	0.000343	-0.017130	0.05
	Sleep efficiency	-0.025624	0.098357	0.010062	-0.009028	-0.009607	-0.025352	1.000000	0.062362	0.788332	-0.819075	-0.554498	0.053539	0.08
	REM sleep percentage	0.051636	0.042091	-0.161102	0.038519	0.037928	-0.018780	0.062362	1.000000	-0.198674	-0.015556	-0.024992	0.073566	0.05
	Deep sleep percentage	-0.047507	0.025800	0.082634	-0.049915	-0.050076	-0.035441	0.788332	-0.198674	1.000000	-0.974266	-0.298433	-0.013825	0.05
	Light sleep percentage	0.037160	-0.031354	-0.039541	0.034731	0.034976	0.037799	-0.819075	-0.015556	-0.974266	1.000000	0.313352	-0.005775	0.07
	Awakenings	-0.055429	-0.017468	0.074205	-0.012240	-0.011645	0.003434	-0.554498	-0.024992	-0.078433	0.313352	1.000000	-0.099927	0.09
	Caffeine con	-0.040	-0.180	-0.299	0.069783	0.069779	-0.017	0.053539	0.073566	-0.013825	-0.005775	-0.099927	1.000000	0.01

sumption	342	135	386			130								3
Alcohol consumption	0.113164	0.046474	0.058287	-0.040375	-0.040204	-0.051693	-0.383671	-0.051709	-0.357714	0.374904	0.198491	-0.113368		1 0 0
Smoking status	0.010804	0.031237	0.143020	0.017204	0.017658	0.007307	-0.290026	0.032184	-0.226446	0.223500	-0.020781	0.026496		0 7 6
Exercise frequency	0.024416	0.071892	0.230169	-0.071991	-0.073468	-0.070873	0.257974	0.031165	0.182975	-0.191706	-0.212485	-0.084257		0 0 8

Multivariate Analysis	Patterns and relationships involving multiple variables. 
Outliers and Anomalies	
Data Preprocessing Code Screenshots	
Loading Data	`sleep=pd.read_csv("/content/Sleep_Efficiency.csv")`
Handling Missing Data	
Data Transformation	
Feature Engineering	Code for creating new features or modifying existing ones.
Save Processed Data	Code to save the cleaned and processed data for future use. sleep = data