Data Mining

IBM-312

Group-Project

Autumn Semester, 2021-22

Sleeping Time Analysis Using Logistic Regression

Group No. : 21

Members:

1. Avi Gupta 19312006
2. Anurag Porwal 19312004
3. Harsh Indoria 19121012
4. Rakshit Joshi 19312025

Raw Data File: G21_Raw_Data_Combined.xlsx
Linear Regression File: G21_Multiple_Linear_Regression.csv
Data Processing & Logistic Regression File: G21_Logistic_Regression.ipynb

Introduction

In today's world, we find that sleep deprivation is a significant problem among people. We try in this project to find a relationship between factors that affect sleep and predict if a person gets a good amount of sleep when given certain information.

Background

According to several studies, an adult must get at least seven hours of sleep for good health. It helps our body to repair itself and function as it should, and is linked to better mental health and lower risk of many health conditions- including heart disease and diabetes. It's also been shown that not getting enough sleep is linked to cognitive decline and conditions such as Alzheimer's disease.

We combined it with various lifestyle decisions like the amount of time one spend on smartphones, number of meals in a day, beverages (tea and coffee), physical exercise routines, smoking or drinking habits, and age, gender, and physical illness to determine how they affect sleep time.

According to Harvard Study, nearly 50% of American adults polled said they used technology in bed at least once a week, and nearly 30% said they did so every day. Some 21% of adults even said if they woke up during the night, they would check their devices before going back to sleep.

Based on a study, exposure to all colors of light helps control your natural sleep-and-wake cycle or circadian rhythm. More so than any other color, blue light messes with one's body's ability to prepare for sleep because it blocks a hormone called melatonin that makes you sleepy.

The relationship between alcohol and sleep has been studied since the 1930s, yet many aspects of this relationship are still unknown. Research has shown sleepers who drink large amounts of alcohol before going to bed are often prone to delayed sleep onset, meaning they need more time to fall asleep. As liver enzymes metabolize the alcohol during their night and the blood alcohol level decreases, these individuals are also more likely to experience sleep disruptions and decreases in sleep quality.

The researchers examined 23 studies that evaluated sleep onset and quality in healthy adults who performed a single session of evening exercise compared with similar adults who did not. They found that not only did exercise not affect sleep, it seemed to help people fall asleep faster and spend more time in deep sleep. However, those who did high-intensity exercise — such as interval training — less than one hour before bedtime took longer to fall asleep and had poorer sleep quality.[Harvard Study]

The recommended sleeping direction per Vaastu shastra is that you lie down with your head pointed southward. This is because the human head is considered to have a polar-like attraction, and it needs to face southward to attract opposite poles while you sleep. Hence, the sleeping direction can affect the rhythm of sleep.

Raw Data File: G21_Raw_Data_Combined.xlsx
Linear Regression File: G21_Multiple_Linear_Regression.csv
Data Processing & Logistic Regression File: G21_Logistic_Regression.ipynb

Methodology

Data Collection

Our data consists of data collected from the internet, and we took a survey including our friends and family members. Surveyed data had 34 entries, and data collected from the internet has 46 entries.
*The survey is conducted using google forms.

Sample Data pre-processing

The data had many problems like the field of average screen time had entries like more than 5 hrs which we assumed to 6hrs for the sake of simplicity. We cleaned the data by replacing the text with numerical values and with standard values, using MinMaxScaler from sklearn.preprocessing.
We removed those entries where gender is not present, to avoid any error.

Results with Multi Linear Regression

Our initial guess was to use multilinear regression model. Our finding was that this model was poorly fitted with the sample data for the independent variables decided in the survey. R square(0.414) was less, significant F value of the model was too high (~0.6606).

Logistic Regression

Logistic regression is used to obtain odds ratio in the presence of more than one explanatory variable. The procedure is quite similar to multiple linear regression, with the exception that the response variable is binomial. The result is the impact of each variable on the odds ratio of the observed event of interest. The main advantage is to avoid confounding effects by analysing the association of all variables together.
In our case we wish to get the likelihood whether a person is having a good sleep or not based on other independent variables.

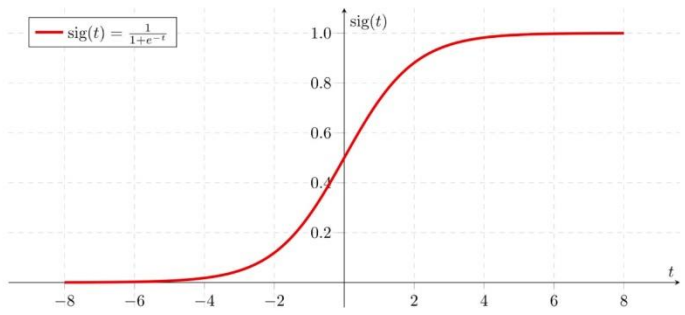
Model

Output = 0 or 1

Hypothesis => Z = WX + B

hΘ(x) = sigmoid (Z)

Sigmoid Function

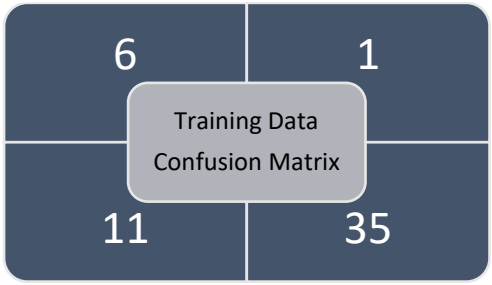


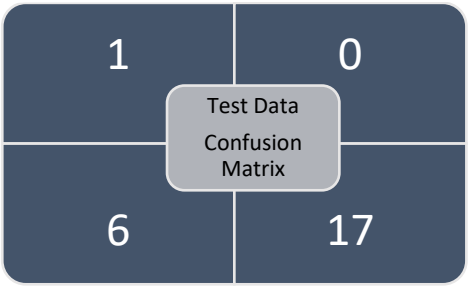
Raw Data File: G21_Raw_Data_Combined.xlsx
Linear Regression File: G21_Multiple_Linear_Regression.csv
Data Processing & Logistic Regression File: G21_Logistic_Regression.ipynb

Training with Logistic Regression Model
Taking the threshold good sleep time to be 7hrs we took a logistic regression of the data. The data was split into the train (70%) and test data (30%). Logistic Regression was imported from sklearn.linear_model library. This was fitted to the training data (X_train, Z_train). X_train, here, includes training data with parameters: Age, Gender, No of meals per day, Physical Illness, Screen Time, Blue light filter, Sleep direction, Exercise, Smoke/ Drink, Beverage. Z_train, here, includes training data with parameters: Sleep time (threshold of 7hrs).

Results

We imported accuracy_score from sklearn.metrics.
Accuracy score of training data: **77.35849%**
Accuracy score of test data: **75.0%**





Coefficients of the logistic regression model:

1.	Age:	-0.78287
2.	Gender:	0.50537
3.	Meals per day:	-0.20574
4.	Physical Illness:	-0.19331
5.	Screen Time:	0.33393
6.	Blue light filter:	0.5715
7.	Exercise:	-0.21316
8.	Smoke\Drink:	0.27227
9.	Tea:	-0.148307
10.	Coffee:	-0.89539
11.	West:	-0.117327
12.	East:	-0.09988
13.	South:	0.86594

Raw Data File: G21_Raw_Data_Combined.xlsx
Linear Regression File: G21_Multiple_Linear_Regression.csv
Data Processing & Logistic Regression File: G21_Logistic_Regression.ipynb

Insights and Analysis

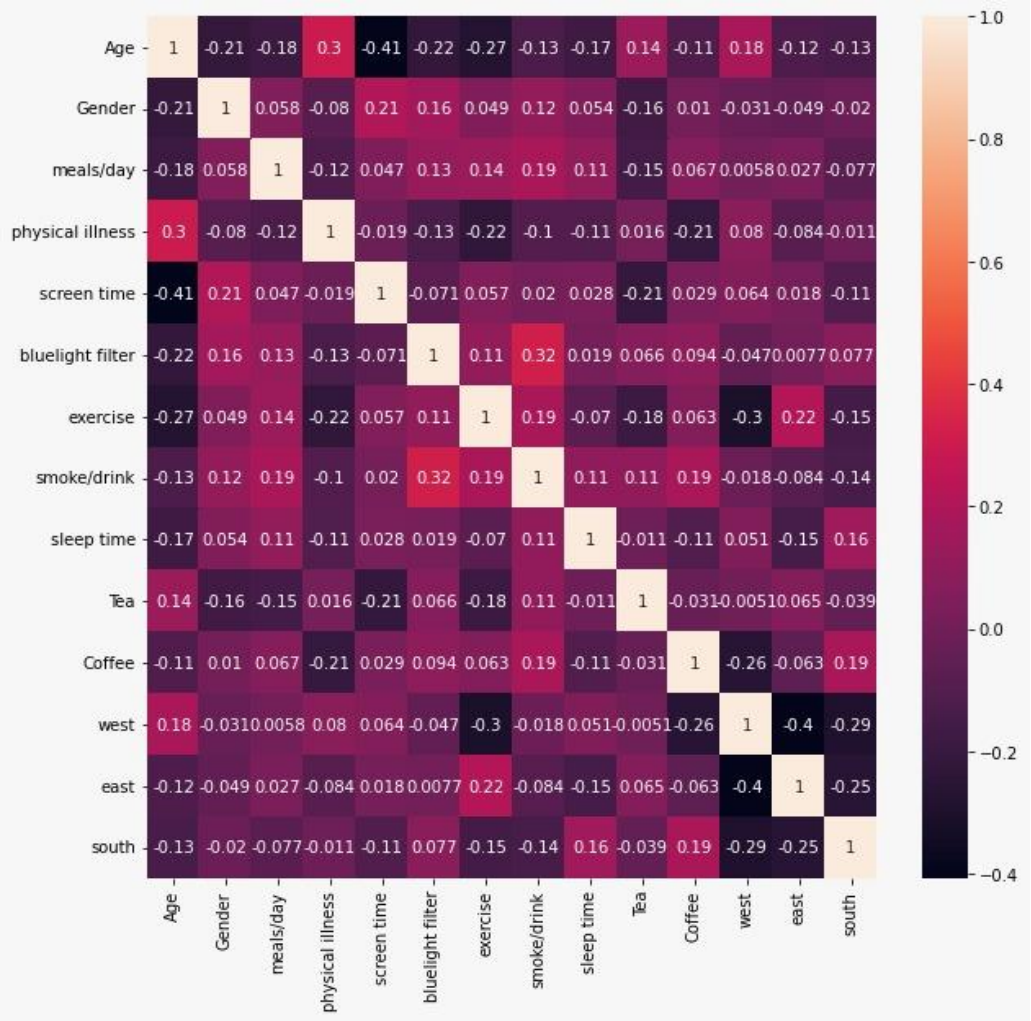
- Acc. to this survey, we find that increase in the age, meals per day, physical illness, exercise, taking tea or coffee, sleeping in west or east direction decrease the odds of having a good sleep time of a person.
- Acc. to this link, people find harder to sleep with growing age which supports our finding that sleep time decreases with age. https://medlineplus.gov/ency/article/004018.htm
- On other hand, being male implies higher odds of good sleep time, and higher screen time, applying blue light filter, smoking/drinking, sleeping in south direction increases the odds of having a good sleep time.
- Smoking/Drinking increases odds of quality sleep is counter-intuitive.
- Screen time tends to increase the odds which seems counter-intuitive acc. to this article: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5839336/ which states that screen time reduces sleep duration in young people.
- Meals per day has negative coefficient which is counter-intuitive as it is known that eating more leads to more sleepiness.
- Chances are there can be some co-relation between the independent variables which results in the counter-intuition (which we have highlighted in below in the 'data visualization' section)
- The independent variables have a low VIF score, less than 5.
- After taking regression of independent variables separately we find that the sign of coefficients remains same. Hence, we can say that the variables are less likely or not correlated.
- People consuming coffee and tea reduces the odds of having good sleep. In fact, the coefficient for coffee is greater than the tea's coefficient which means coffee is better in reducing sleep.
- Acc. to the data, direction for good sleep is in the order: South> North> East>West.

Raw Data File: G21_Raw_Data_Combined.xlsx
Linear Regression File: G21_Multiple_Linear_Regression.csv
Data Processing & Logistic Regression File: G21_Logistic_Regression.ipynb

Biases

- Our exercise data didn't include separately for morning and evening exercise but we learnt that evening exercise tends to decrease the quality of sleep while morning exercise tends to increase as per studies.
- Convenience Bias: our own survey was limited to our friends and family members and we strongly feel that this has biased the data from people who consume alcohol or smoke that has a very crucial role in determining the quality of sleep.
- Response Bias: There might be a chance that people who participated in the survey were intentionally or non-intentionally provide false feedback for the asked questions.
- Our data lacked female candidate participation which could have been better for a more correct prediction.
- Data revolves around urban middle-class citizens. This changes for different class of society like Rural areas.

Data Visualisation

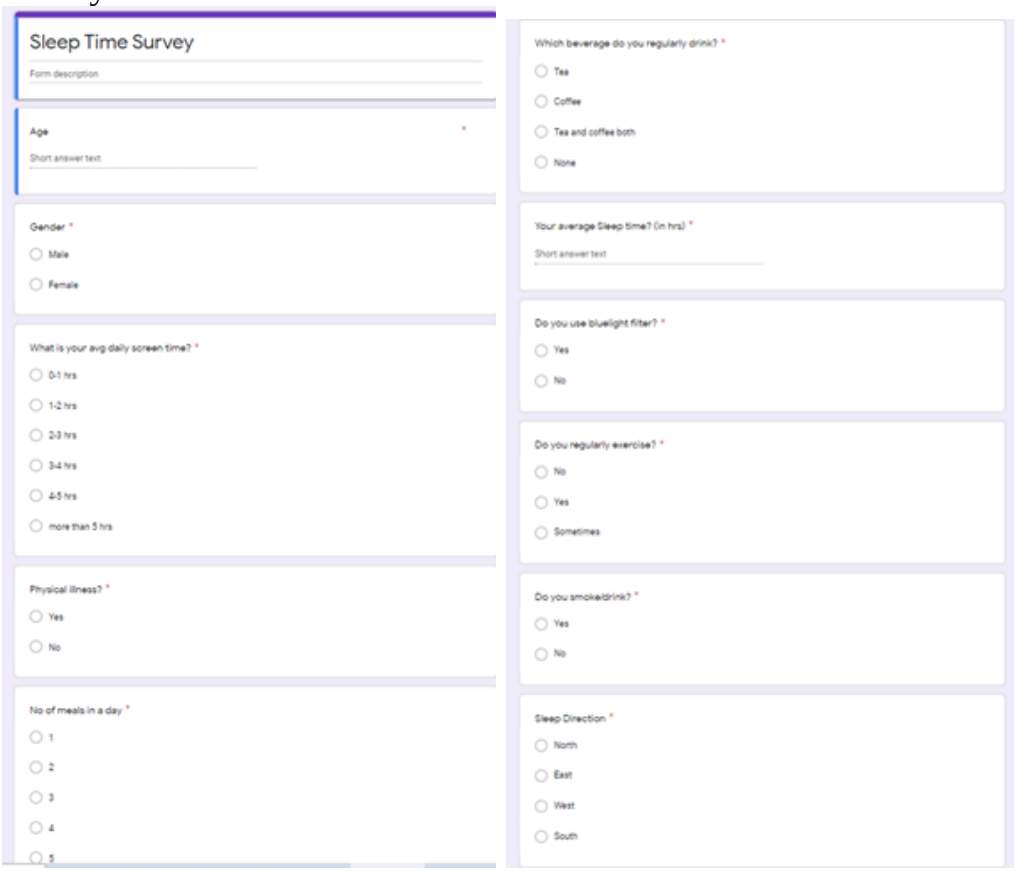


Above heat map of pairwise correlation, we can see that Age is working as a latent variable due to which there are some counter intuitive correlations of sleep time with screen time. According to studies, and intuition, increase in screen time leads to reduction in sleep time, however, our logistic regression model shows a positive correlation between these. This is because of the strong negative correlation between age, and screen time, and strong negative correlation between age and sleep time. Therefore, increase in screen time means the subject is of lower age and so he/she sleeps more. Hence, age acts as a latent variable here.

Raw Data File: G21_Raw_Data_Combined.xlsx
Linear Regression File: G21_Multiple_Linear_Regression.csv
Data Processing & Logistic Regression File: G21_Logistic_Regression.ipynb

Snapshots

Survey Form



Data Snapshot: Excel file

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
1		Age	Gender	meals/day	physical illr	screen tim	bluelight fil	exercise	smoke/dri	Tea	Coffee	west	east	south	sleep time	sleep direction	beverage
2	0	20	1	4	0	6	0	1	0	1	0	1	0	0	8	West	Tea
3	1	21	1	3	0	6	0	1	0	1	1	0	0	0	8	North	Tea and Coffee both
4	2	20	1	3	0	6	0	1	0	0	0	0	0	0	7	North	None
5	3	19	1	4	0	4.5	0	1	0	0	0	0	1	0	8	East	None
6	4	45	0	3	0	3.5	0	1	0	0	0	0	0	0	6	North	None
7	5	20	1	5	1	6	0	1	0	0	0	1	0	0	5.9	West	None
8	6	17	1	5	0	6	0	0	0	0	0	1	0	0	8.5	West	None
9	7	20	0	3	0	4.5	1	1	0	0	0	1	0	0	7	West	None
10	8	20	1	4	0	3.5	1	1	1	0	1	0	0	0	8	North	Coffee
11	9	21	1	4	0	6	1	1	0	0	0	0	1	0	8	East	None
12	10	20	0	4	0	2.5	0	1	0	0	0	0	1	0	8.5	East	None
13	11	19	1	3	0	6	1	1	0	0	1	0	0	1	8.5	South	Coffee
14	12	20	1	4	0	6	1	1	1	1	1	0	0	0	8	North	Tea and Coffee both
15	13	21	1	3	0	4.5	0	0	0	1	0	1	0	0	8	West	Tea
16	14	21	1	3	0	6	1	1	1	1	0	1	0	0	7	West	Tea
17	15	19	1	4	0	1.5	1	1	1	0	0	0	0	0	8	North	None
18	16	19	1	5	0	6	0	1	0	0	0	1	0	0	8	West	None
19	17	26	1	2	0	6	0	1	0	0	0	0	1	0	6	East	None
20	18	21	1	3	0	6	1	0	0	1	0	0	1	0	6.5	East	Tea
21	19	47	0	3	1	1.5	0	0	0	1	0	0	0	0	6	North	Tea
22	20	58	0	2	1	2.5	0	0	0	1	0	1	0	0	7.5	West	Tea
23	21	18	1	3	0	4.5	1	1	0	0	0	0	0	0	6.5	North	None
24	22	20	1	3	0	6	0	1	0	0	0	0	0	0	6	North	None
25	23	21	1	2	0	3.5	1	0	0	1	1	0	0	1	10	South	Tea and Coffee both
26	24	17	0	2	0	4.5	0	0	0	1	0	0	0	1	8	South	Tea
27	25	25	1	3	0	1.5	1	1	0	1	0	0	1	0	7	East	Tea
28	26	30	1	2	0	6	0	1	0	1	1	0	0	0	6	North	Tea and Coffee both

Sheet1 | Combined_Sleep_Analysis

Raw Data File: G21_Raw_Data_Combined.xlsx
Linear Regression File: G21_Multiple_Linear_Regression.csv
Data Processing & Logistic Regression File: G21_Logistic_Regression.ipynb

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.375808
R Square	0.141231
Adjusted R	-0.03597
Standard E	1.094708
Observatio	77

ANOVA

	df	*SS*	*MS*	*F*	*ignificance F*
Regression	13	12.41631	0.955101	0.796989	0.660645
Residual	63	75.49834	1.198386		
Total	76	87.91465			

	Coefficients	*andard Erro*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*ower 95.0%*	*pper 95.0%*
Intercept	7.753935	1.077826	7.194053	9.17E-10	5.600072	9.907798	5.600072	9.907798
Age	-0.0231	0.019937	-1.15844	0.251056	-0.06294	0.016745	-0.06294	0.016745
Gender	0.019304	0.287269	0.067198	0.946637	-0.55476	0.593365	-0.55476	0.593365
meals/day	0.116034	0.148653	0.780572	0.437977	-0.18102	0.413093	-0.18102	0.413093
physical illr	-0.43136	0.474682	-0.90874	0.366952	-1.37994	0.517215	-1.37994	0.517215
screen time	-0.01667	0.082894	-0.20114	0.841237	-0.18232	0.148976	-0.18232	0.148976
bluelight fil	-0.16745	0.277545	-0.60332	0.54846	-0.72208	0.387181	-0.72208	0.387181
exercise	-0.27206	0.337418	-0.80631	0.423096	-0.94634	0.402212	-0.94634	0.402212
smoke/drir	0.573728	0.497793	1.152543	0.253452	-0.42103	1.568489	-0.42103	1.568489
Tea	0.000908	0.273245	0.003323	0.997359	-0.54513	0.546944	-0.54513	0.546944
Coffee	-0.48989	0.300268	-1.63151	0.107772	-1.08993	0.110148	-1.08993	0.110148
west	0.016719	0.361734	0.046219	0.963282	-0.70615	0.739588	-0.70615	0.739588
east	-0.26943	0.357639	-0.75335	0.454048	-0.98411	0.445259	-0.98411	0.445259
south	0.47673	0.43581	1.093893	0.278167	-0.39417	1.347626	-0.39417	1.347626

Sheet1 Combined_Sleep_Analysis

Multiple regression analysis

References

1. https://www.sleepfoundation.org/nutrition/alcohol-and-sleep
2. https://www.health.harvard.edu/staying-healthy/does-exercising-at-night-affect-sleep
3. Survey Link: https://docs.google.com/spreadsheets/d/1C5PDqf-TC8pw1XjbEuCZSe1Y4XI4uf6WDWorANS-SSY/edit#gid=29110817
4. Data from internet: https://www.kaggle.com/krupa1999/sleep-pattern
5. https://www.mayoclinic.org/healthy-lifestyle/adult-health/expert-answers/how-many-hours-of-sleep-are-enough/faq-20057898
6. https://towardsdatascience.com/logistic-regression-detailed-overview-46c4da4303bc

Raw Data File: G21_Raw_Data_Combined.xlsx
Linear Regression File: G21_Multiple_Linear_Regression.csv
Data Processing & Logistic Regression File: G21_Logistic_Regression.ipynb