Unlocking the Secrets of Quality Sleep - Data Report

Author: Ammar Al-Baadani

Date: 9/15/2023

Class: ITSS 4V95 - Seminar Series In Information Systems

Table of Contents:

Business Problem Overview

In todays fast-paced environment many individuals neglect the importance of sleep quality. Quality sleep directly influences productivity, cognitive performance, emotional well-being, and overall health. Recognizing this, businesses understand that well-rested employees are not only more efficient but also happier and more engaged.

This project is driven by the critical need to uncover the factors that influence sleep quality and to derive actionable insights that empower individuals to enhance their sleep experiences.

Data Overview and Pre-processing

Data Overview *Provide information related to data source and update frequency. Use knowledge learned to classify and define data and give insights into the overall quality of data.*

Example:

● The data contains 13 columns and information about 374 synthesized data related to individuals with different sleep quality, occupations, and more.

● The data used was uploaded on kaggle and is 4 months old https://www.kaggle.com/datasets/uom190346a/sleep-health-and-lifestyle-dataset (https://www.kaggle.com/datasets/uom190346a/sleep-health-and-lifestyle-dataset).

● The characteristics include ('Person_ID', 'Gender', 'Age', 'Occupation', 'Sleep Duration', 'Quality of Sleep', 'Physical Activity Level', 'Stress Level', 'BMI Category','Blood Pressure', 'Heart Rate', 'Daily Steps', 'Sleep Disorder')

● No columns have missing values, but I will exclude Blood Pressure and Heart Rate.

Data Pre-processing

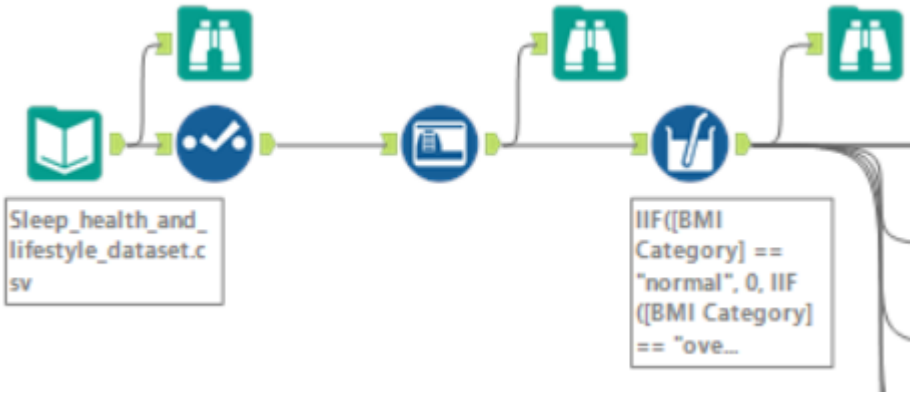
The tool of choice for this project was Alteryx. I started by selecting only 10 columns out of the 13, these include (Gender', 'Age', 'Occupation', 'Sleep Duration', 'Quality of Sleep', 'Physical Activity Level', 'Stress Level', 'BMI Category','Daily Steps', 'Sleep Disorder') and renamed 'Physical Activity Level' and 'Stress Level' to 'Activness_lvl' and 'Stress_lvl' respectivly. I then change the data type of all the column to the correct to the appropriate type using (Auto Filed).

Since There are no null values in this data set I decided to skip the cleaning process.

Here is a quick view of what the cleaned data contains:

Record	Gender	Age	Occupation	Sleep Duration	Quality of Sleep	Activness_lvl	Stress_lvl	BMI Category	Daily Steps	Sleep Disorder
1	Male	27	Software Engineer	6.1	6	42	6	1	4200	None
2	Male	28	Doctor	6.2	6	60	8	0	10000	None
3	Male	28	Doctor	6.2	6	60	8	0	10000	None
4	Male	28	Sales Representative	5.9	4	30	8	2	3000	Sleep Apnea
5	Male	28	Sales Representative	5.9	4	30	8	2	3000	Sleep Apnea
6	Male	28	Software Engineer	5.9	4	30	8	2	3000	Insomnia
7	Male	29	Teacher	6.3	6	40	7	2	3500	Insomnia
8	Male	29	Doctor	7.8	7	75	6	0	8000	None

And here is the workflow of the cleaning process:



Exploratory Data Analysis

As mentioned earlier, the goal of this EDA is to examining different asspects of the individuals and visualize the data to gain more insights, discover patterns, and identify potential issues.

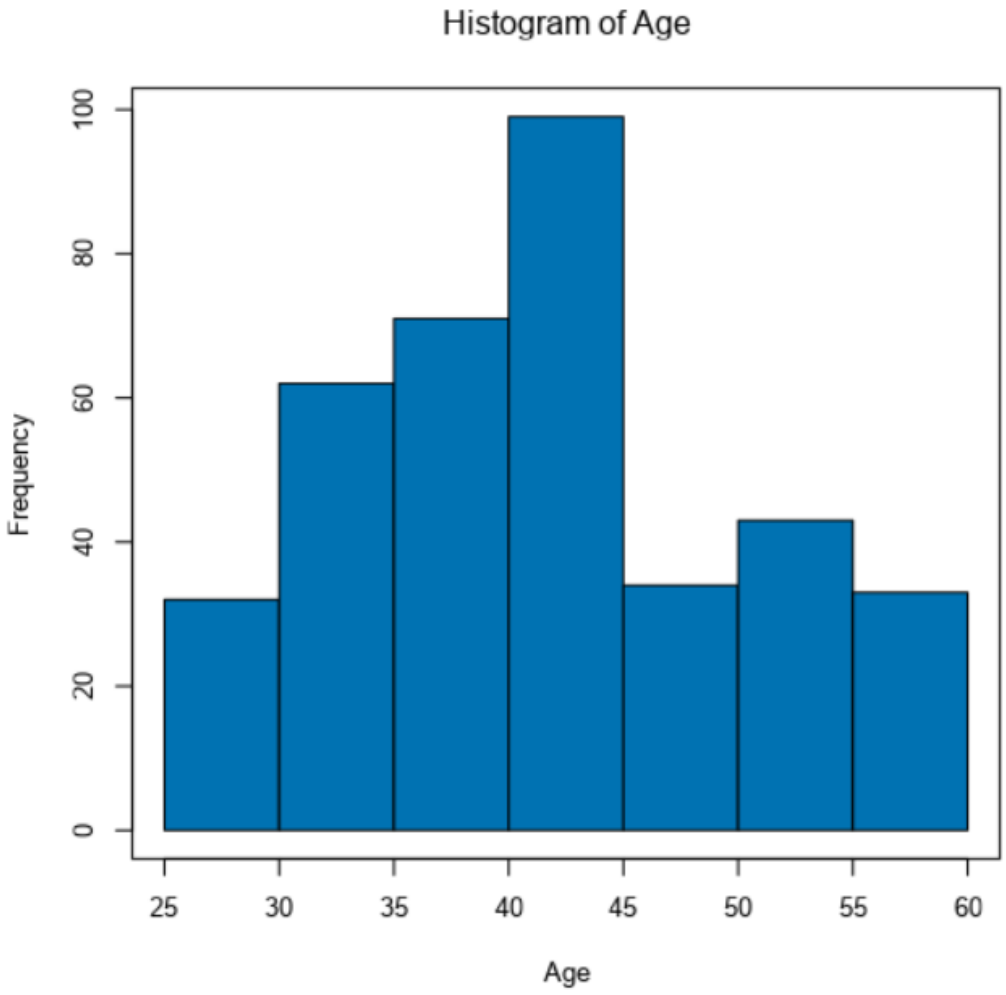
Descriptive Statistics:

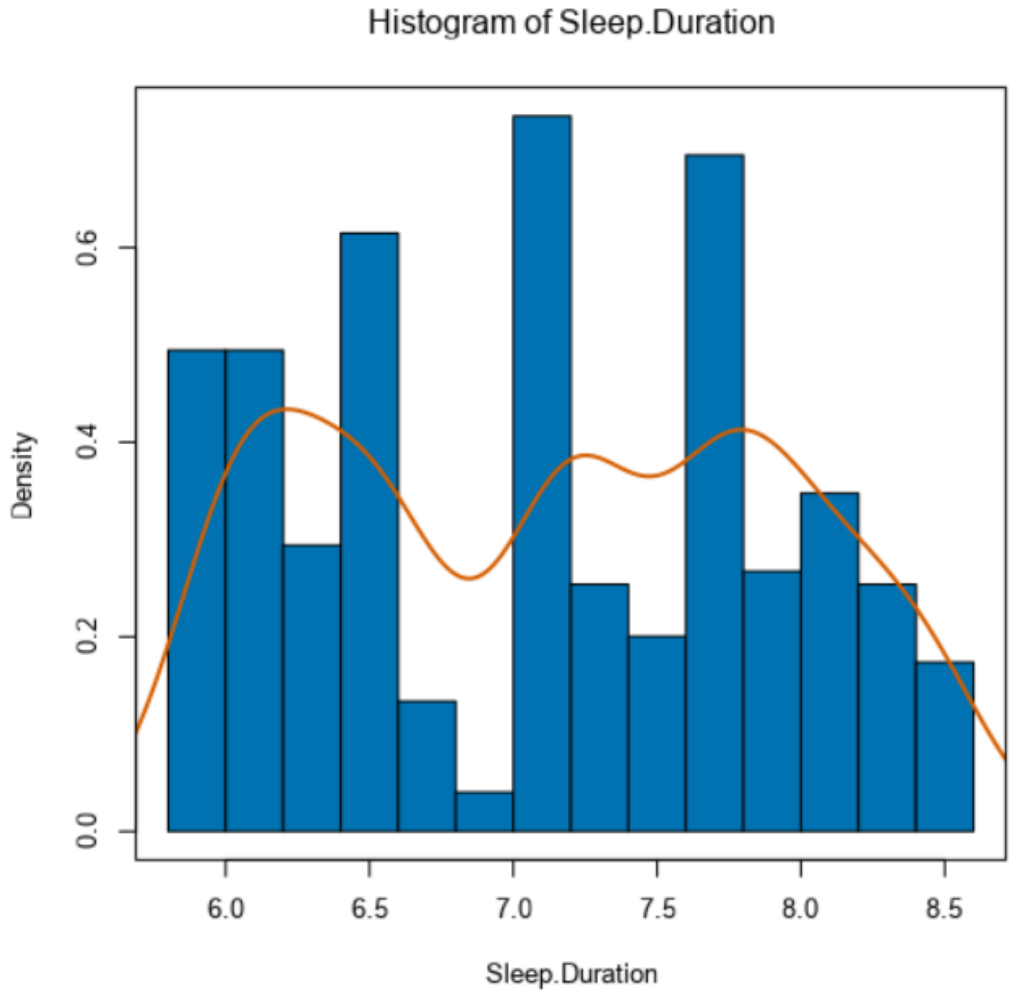
Summary statistics:

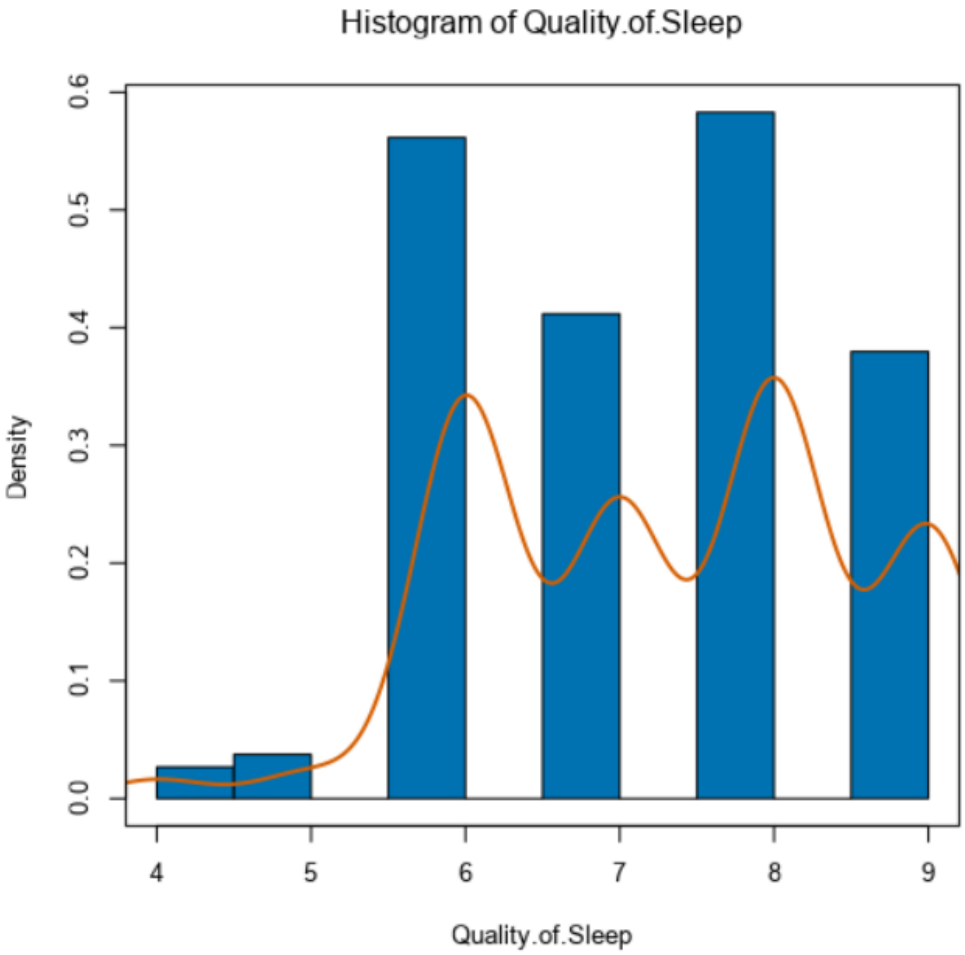
After cleaning the data I wanted to know what is the range of the data that I'm dealing with as well as how many unique values each column had and for that I used the Field Summary tool:

Record	Name	Field Category	Min	Max	Median	Std. Dev.	Percent Missing	Unique Values	Mean
1	BMI Category	Numeric	0	2	0	0.642727	0	3	0.561497
2	Stress_lvl	Numeric	3	8	5	1.774526	0	6	5.385027
3	Quality of Sleep	Numeric	4	9	7	1.196956	0	6	7.312834
4	Sleep Duration	Numeric	5.8	8.5	7.2	0.795657	0	27	7.132086
5	Activness_lvl	Numeric	30	90	60	20.830804	0	16	59.171123
6	Age	Numeric	27	59	43	8.673133	0	31	42.184492
7	Occupation	String	[Null]	[Null]	[Null]	[Null]	0	11	[Null]

Frequency distributions: Based on the summary I can now focus on the important factors that I will be using to reach some insights. I decided to figure out what age group this data relates to and how sleep quiality and duration is distribured.

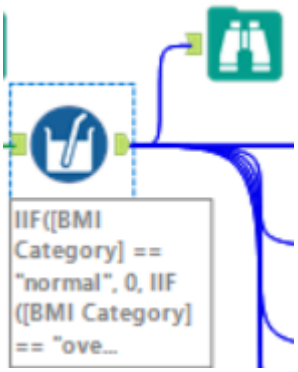



Histogram of Sleep.Duration
Density
0.0
0.2
0.4
0.6
6.0
6.5
7.0
7.5
8.0
8.5
Sleep.Duration



Data Transformation:

Data encoding: Since I wanted to investigate the relationship between BMI and sleep quality I had to convert BMI from a categorical variables into numerical format using Multi-Field Formula tool. There, I assigned each value in the 'BMI Category' a number so that I can use it in my analysis. There were 3 unique values, Normal, Overweight, and Obese with each being assigned 0, 1, and 2 respectively.



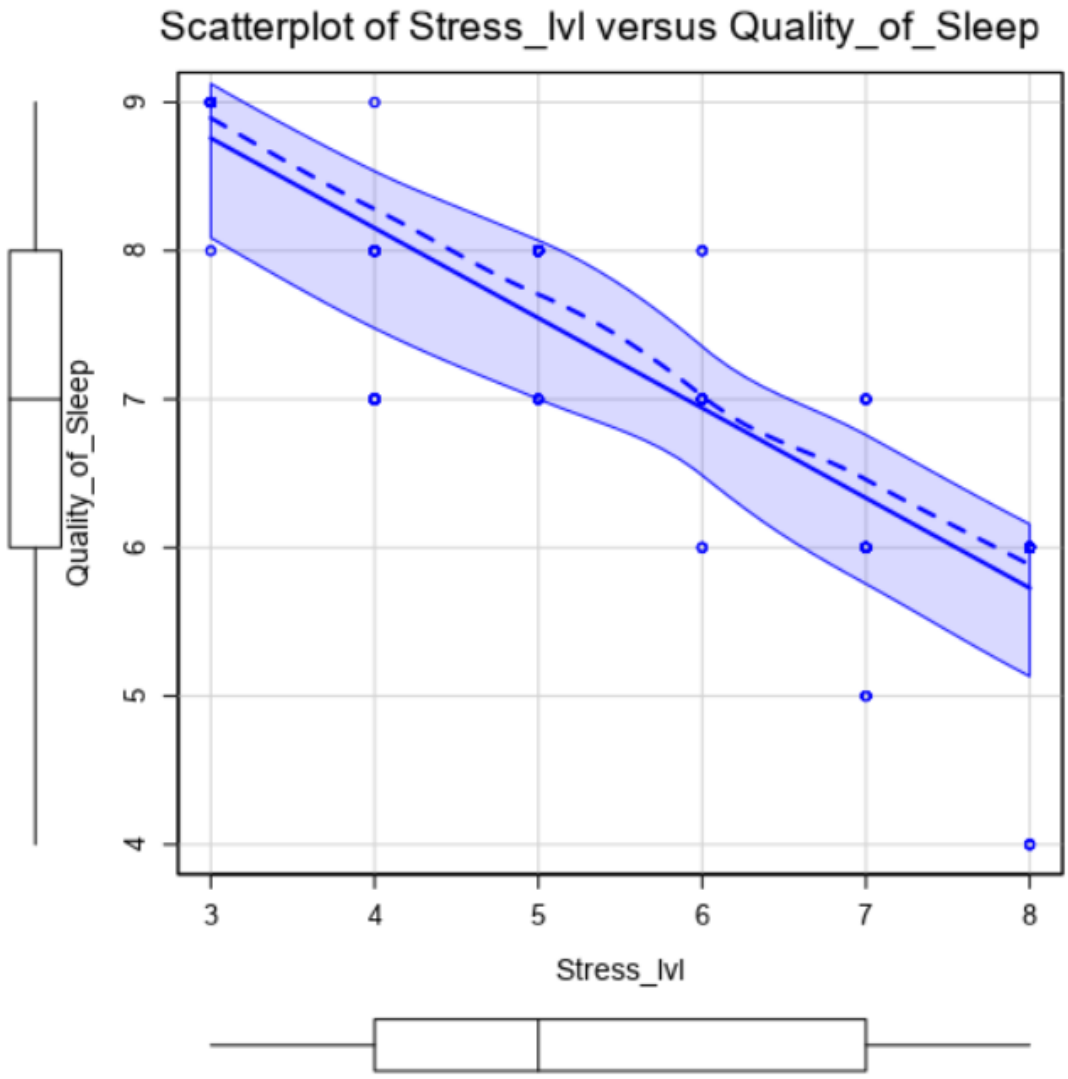
This is how the column used to represent the data vs how it is representing it now:

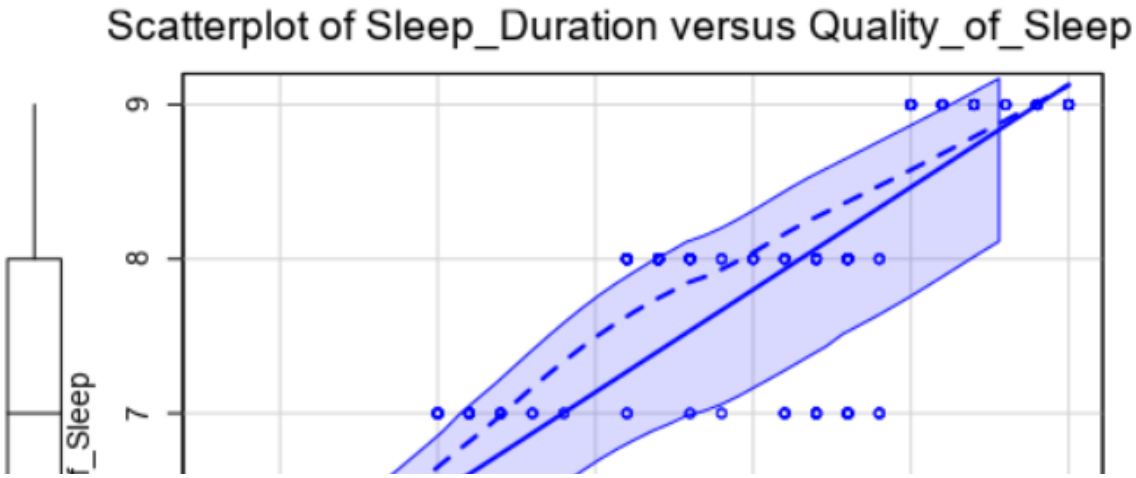
BMI Category
Overweight
Normal
Normal
Obese
Obese
Obese
Obese
Normal

BMI Category
1
0
0
2
2
2
2
0

Outlier Detection and Relation

Visual inspection: To answer my questions, I used a scatterplot to find outliers and relations by comparing 'Quality of Sleep' with 'Sleep Duration' and 'Stress_lvl' as well as comparing 'Sleep Duration' with 'Stress_lvl':


Scatterplot of Stress_lvl versus Quality_of_Sleep
Quality_of_Sleep
4
5
6
7
8
9
3
4
5
6
7
8
Stress_lvl



Conclusion

In conclusion, we successfully highlighted the significant impact of stress on sleep duration and quality. We ruled out what factors cause us to sleep less and how it can affect our daily life and performance. It is advisable to consider stress-reduction strategies and establish consistent sleep routines to combat sleep quality issues. Also, even if there is no correlation between BMI and sleep quality, it is best to lead an active lifestyle to help manage stress and boost your confidence! I hope these insights empower individuals to take proactive steps toward better sleep, ultimately contributing to improved physical and mental well-being.

While there are more insights that I derived from this dataset, I decided to leave it out since it is not really related to the Business Proble.

Here is the complete workflow: