Sleep Data Analysis: From Raw Data to Insights

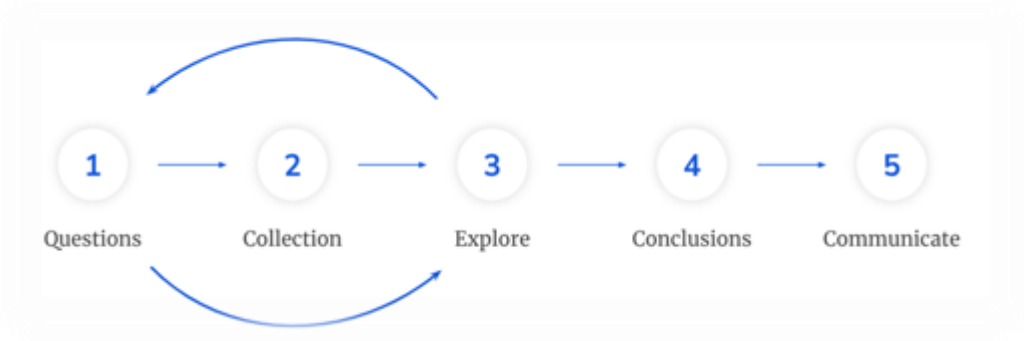
By Diana Nicutari

Executive Summary

A data-driven investigation into sleep health patterns and their correlations with lifestyle factors, conducted as part of the BrainStation's Data Analytics program. This analysis aims to explore the relationships between sleep quality, duration, and various health metrics.

Background

As a biologist transitioning into health data analytics, this project focuses on analyzing sleep health data to understand the factors affecting sleep quality and duration. The analysis follows the Process Framework: Questions → Collection → Explore → Conclusions → Communicate



Problem Definition & Research Questions

According to "The Global Problem of Insufficient Sleep and Its Serious Public Health Implications" (Healthcare, 2018):

- Sleep quality is essential for physical and mental health

- Insufficient sleep is a prevalent issue in modern society

– Medical professionals need to understand common sleep disruptors

Primary Research Question: What impacts sleep quality and duration?

- *What health and lifestyle metrics may influence sleep quality/duration or vice versa?*

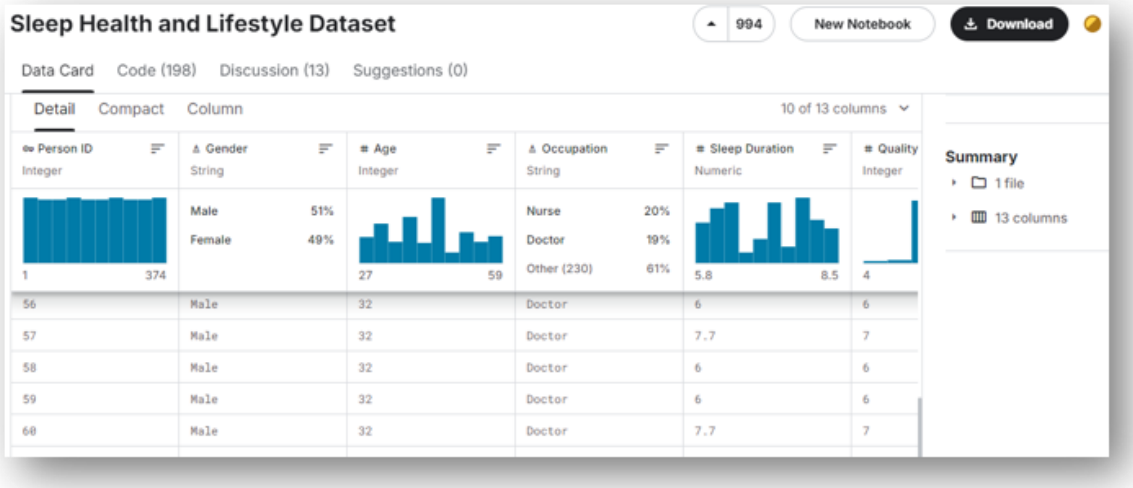
Project Goal: Test the hypothesis that sleep quality and duration positively correlate with a healthy lifestyle.

Data Collection & Methodology

Dataset Overview

- Source: Sleep Health and Lifestyle Dataset (Kaggle)
- Format: CSV file
- Size: 375 rows × 13 columns
- Type: Synthetic data created for illustrative purposes
- Source: Kaggle Dataset Link

Key Measurements



Demographics

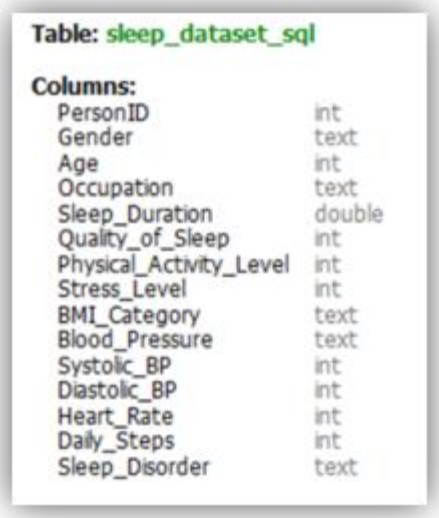
- Person ID
- Gender (Male/Female)
- Age (years)
- Occupation

Sleep Metrics

- Sleep Quality (scale: 1-10)
- Sleep Duration (hours/day)
- Sleep Disorder (None, Insomnia, Sleep Apnea)

Health & Lifestyle Indicators

- Physical Activity Level (minutes/day)
- Daily Steps
- BMI Category (Underweight, Normal, Overweight)
- Blood Pressure (systolic/diastolic)
- Heart Rate (bpm)
- Stress Level (scale: 1-10)



Secondary Research Questions

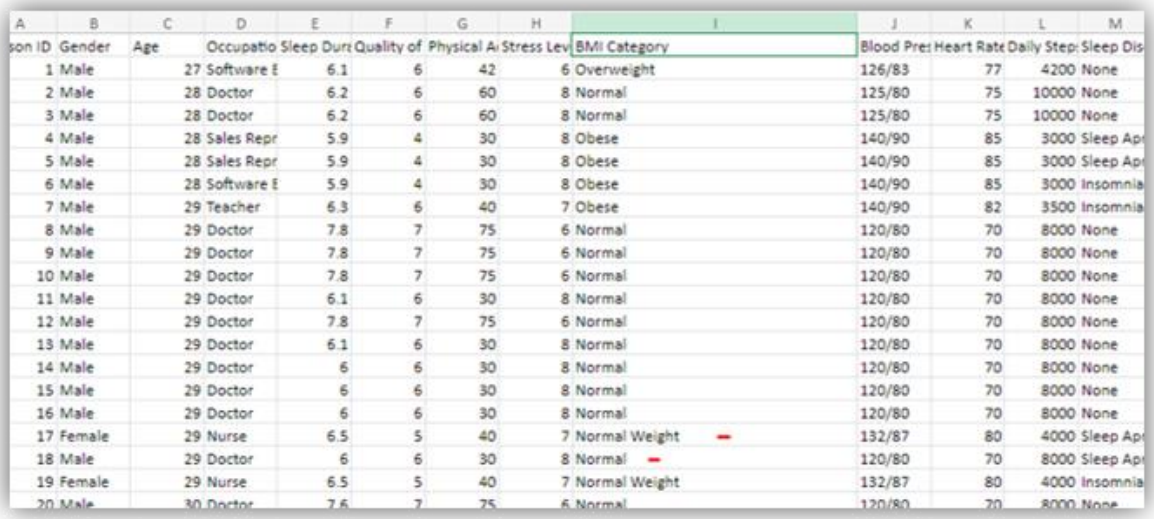
1. Gender Differences
   - Is there a difference between genders in terms of sleep quality?
   - Is there a difference between genders in terms of sleep duration?
2. Physical Activity Impact
   - Do people with higher physical activity levels experience better sleep quality?
   - Is there a relationship between daily steps and sleep duration?
3. Occupation & Stress Effects
   - Do people with higher stress levels sleep more or less?
   - Is there a correlation between occupation, stress levels and sleep quality?
4. Health Indicators
   - How do blood pressure levels correlate with sleep patterns?
   - Is there a relationship between BMI category and sleep disorders?

Data Analysis

Data Preparation

1. BMI Category Standardization

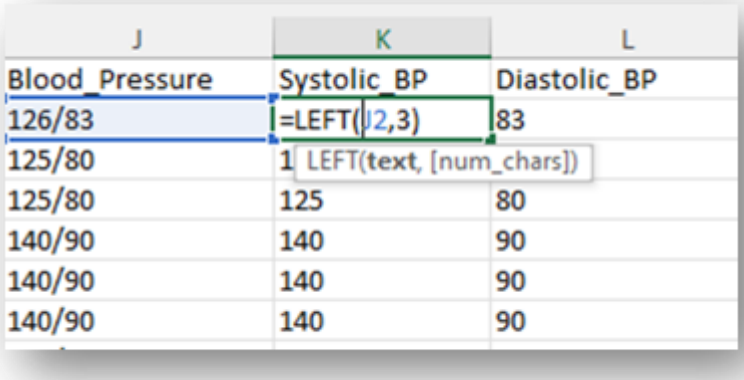
- Used Excel's 'Split Text to Columns' feature to standardize BMI categories
- Corrected inconsistencies between "Normal" and "Normal Weight"



A	B	C	D	E	F	G	H	I	J	K	L	M
son ID	Gender	Age	Occupatio	Sleep Dur	Quality of	Physical A	Stress Lev	BMI Category	Blood Pre	Heart Rate	Daily Step	Sleep Dis
1	Male	27	Software E	6.1	6	42	6	Overweight	126/83	77	4200	None
2	Male	28	Doctor	6.2	6	60	8	Normal	125/80	75	10000	None
3	Male	28	Doctor	6.2	6	60	8	Normal	125/80	75	10000	None
4	Male	28	Sales Repr	5.9	4	30	8	Obese	140/90	85	3000	Sleep Ap
5	Male	28	Sales Repr	5.9	4	30	8	Obese	140/90	85	3000	Sleep Ap
6	Male	28	Software E	5.9	4	30	8	Obese	140/90	85	3000	Insomnia
7	Male	29	Teacher	6.3	6	40	7	Obese	140/90	82	3500	Insomnia
8	Male	29	Doctor	7.8	7	75	6	Normal	120/80	70	8000	None
9	Male	29	Doctor	7.8	7	75	6	Normal	120/80	70	8000	None
10	Male	29	Doctor	7.8	7	75	6	Normal	120/80	70	8000	None
11	Male	29	Doctor	6.1	6	30	8	Normal	120/80	70	8000	None
12	Male	29	Doctor	7.8	7	75	6	Normal	120/80	70	8000	None
13	Male	29	Doctor	6.1	6	30	8	Normal	120/80	70	8000	None
14	Male	29	Doctor	6	6	30	8	Normal	120/80	70	8000	None
15	Male	29	Doctor	6	6	30	8	Normal	120/80	70	8000	None
16	Male	29	Doctor	6	6	30	8	Normal	120/80	70	8000	None
17	Female	29	Nurse	6.5	5	40	7	Normal Weight	132/87	80	4000	Sleep Ap
18	Male	29	Doctor	6	6	30	8	Normal	120/80	70	8000	Sleep Ap
19	Female	29	Nurse	6.5	5	40	7	Normal Weight	132/87	80	4000	Insomnia
20	Male	30	Doctor	7.6	7	75	6	Normal	120/80	70	8000	None

2. Blood Pressure Column Processing

- Split BP text string into two numerical columns:
  - Systolic Blood Pressure
  - Diastolic Blood Pressure
- Used LEFT() and RIGHT() functions for separation



J	K	L
Blood_Pressure	Systolic_BP	Diastolic_BP
126/83	=LEFT(J2,3)	83
125/80	1	
125/80	125	80
140/90	140	90
140/90	140	90
140/90	140	90

LEFT(**text**, [num_chars])

3. Age Group Classification Total age range: 27-59 years Created three age groups:

- Group 1: Late 20's to mid-30's (27-36)
- Group 2: Late 30's to mid-40's (37-46)
- Group 3: Late 40's to late-50's (47-59)

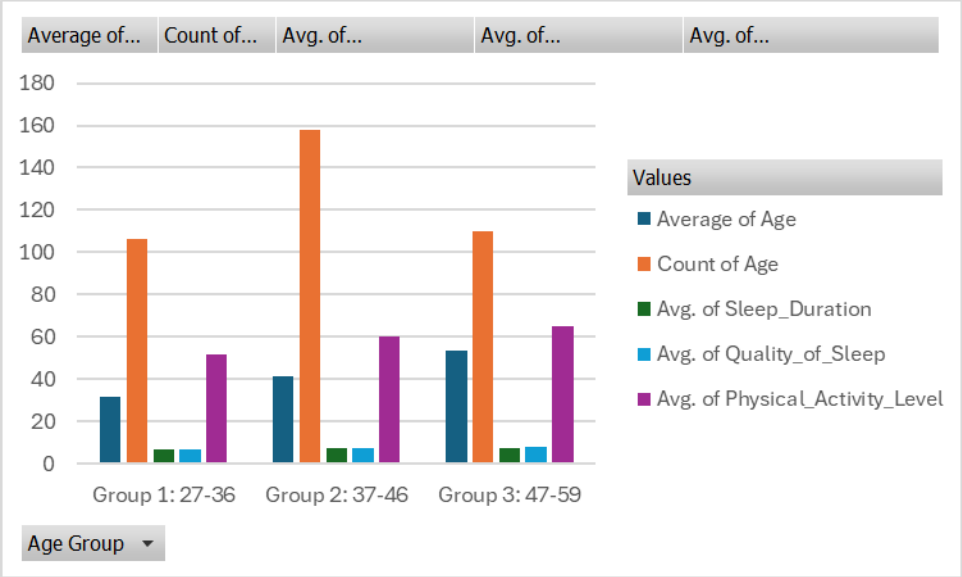
Excel formula used:

```
=IF([@Age]<37, "Group 1: 27-36", IF([@Age]<47, "Group 2: 37-46", "Group 3:
47-59"))
```

Row Labels	Average of Age	Count of Age	Avg. of Sleep_Duration	Avg. of Quality_of_Sleep	Avg. of Physical_Activity_Level
Group 1: 27-36	31.93396226	106	6.878301887	6.613207547	51.89622642
Group 2: 37-46	41.31012658	158	7.049367089	7.329113924	60.1835443
Group 3: 47-59	53.31818182	110	7.495454545	7.963636364	64.72727273
Grand Total	**42.18449198**	**374**	**7.132085561**	**7.312834225**	**59.17112299**

Initial Findings – Pivot Table Analysis

- Data shows unbalanced distribution across age groups (Group 2 has 158 individuals vs 106 and 110 in other groups)
- Sleep duration shows slight increase with age
- Sleep quality increases more notably with age
- Physical Activity levels are highest in the older group (47-59)



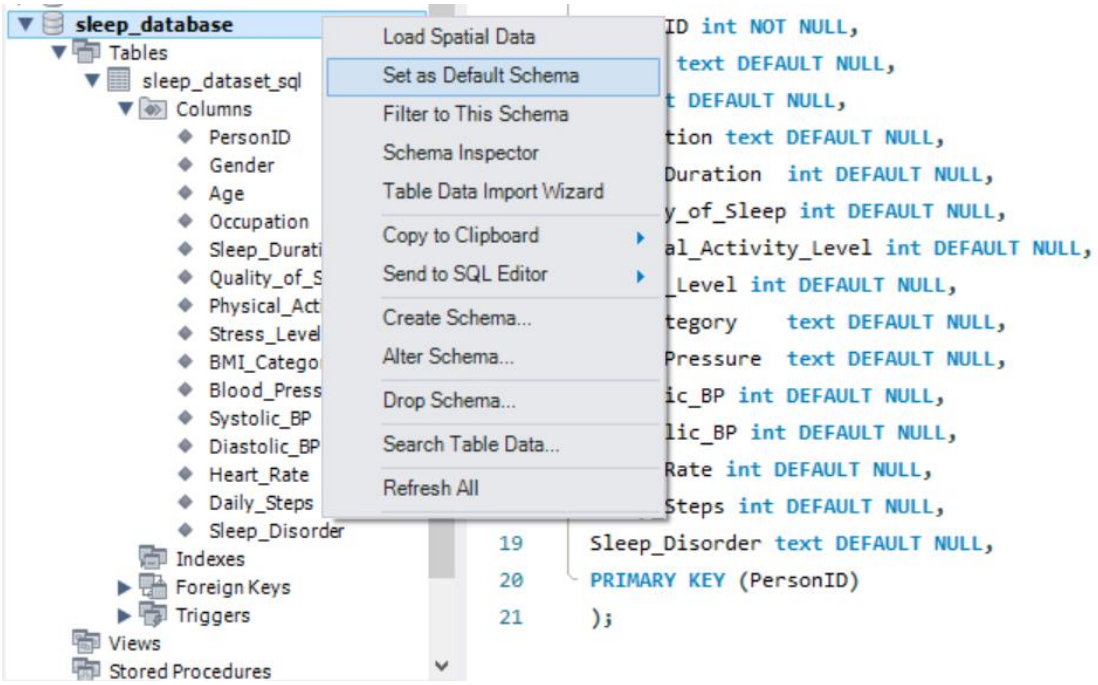
Database Analysis with SQL

Initial Database Setup

```
CREATE SCHEMA sleep_database;
USE Sleep_dataset_SQL;

CREATE TABLE sleepdata (
PersonID int NOT NULL,
Gender text DEFAULT NULL,
Age int DEFAULT NULL,
Occupation text DEFAULT NULL,
Sleep_Duration double DEFAULT NULL,
Quality_of_Sleep int DEFAULT NULL,
Physical_Activity_Level int DEFAULT NULL,
Stress_Level int DEFAULT NULL,
BMI_Category text DEFAULT NULL,
Blood_Pressure text DEFAULT NULL,
```

```
Systolic_BP int DEFAULT NULL,
Diastolic_BP int DEFAULT NULL,
Heart_Rate int DEFAULT NULL,
Daily_Steps int DEFAULT NULL,
Sleep_Disorder text DEFAULT NULL,
PRIMARY KEY (PersonID)
);
```



Data Analysis Queries and Results

Gender Distribution Analysis

```
SELECT COUNT(PersonID), Gender FROM sleep_database.sleep_dataset_sql
GROUP BY Gender;
```

Result Grid | Filter Rows:

COUNT(PersonID)	Gender
189	Male
185	Female

```
SELECT AVG(Age), Gender FROM sleep_database.sleep_dataset_sql
group by Gender;
```

AVG(Age)	Gender
37.0741	Male
47.4054	Female

Notable Gender-Based Findings:

1. Age Distribution

- Male Average Age: 37.0741 years

- Female Average Age: 47.4054 years

2. Stress Level Analysis

```
SELECT AVG(Stress_Level), Gender FROM sleep_database.sleep_dataset_sql
group by Gender;
```

Results:

- Male Average Stress: 6.0794

- Female Average Stress: 4.6757

3. General Health Metrics

```
SELECT COUNT(DISTINCT BMI_Category), AVG(Systolic_BP), AVG(Diastolic_BP),
       AVG(Daily_Steps), AVG(Physical_Activity_Level), Gender
FROM sleep_database.sleep_dataset_sql
group by Gender;
```

	COUNT(DISTINCT BMI_Category)	AVG(Systolic_BP)	AVG(Diastolic_BP)	AVG(Daily_Steps)	Gender
▶	3	130.2000	86.3189	6840.5405	Female
	3	126.9418	83.0159	6793.6508	Male

Results demonstrate balanced distribution across genders for: - BMI Categories (3 categories per gender) - Average Blood Pressure - Daily Steps - Physical Activity Levels

Blood Pressure Analysis

High Blood Pressure Investigation

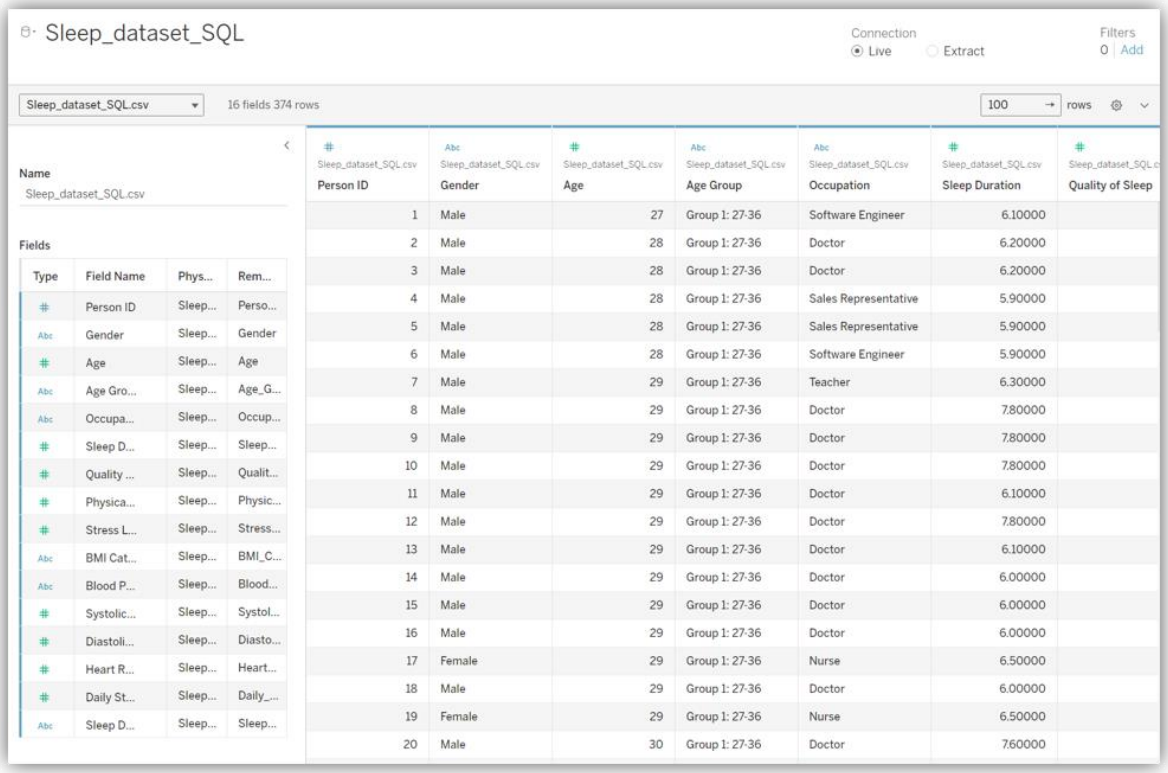
Definition criteria: - Stage 1 High BP: 130-139 mmHg/80-89 mmHg - Stage 2 High BP: ≥140/90 mmHg

```
-- Total High BP Cases
SELECT COUNT(*) FROM sleep_database.sleep_dataset_sql
WHERE (( Diastolic_BP > 80) OR (Systolic_BP > 130))
-- Result: 221 cases

-- Stage 2 High BP Cases
SELECT COUNT(*) FROM sleep_database.sleep_dataset_sql
WHERE (( Diastolic_BP > 90) OR (Systolic_BP > 140))
-- Result: 69 cases
```

```sql
-- High BP Gender Distribution
SELECT COUNT(*), Gender FROM sleep_database.sleep_dataset_sql
WHERE (( Diastolic_BP > 80) OR (Systolic_BP > 130))
group by Gender;
-- Results: Male: 108, Female: 113

-- Stage 2 High BP Gender Distribution
SELECT COUNT(*), Gender FROM sleep_database.sleep_dataset_sql
WHERE (( Diastolic_BP > 90) OR (Systolic_BP > 140))
group by Gender;
-- Results: Male: 4, Female: 65
```



Person ID	Gender	Age	Age Group	Occupation	Sleep Duration	Quality of Sleep
1	Male	27	Group 1: 27-36	Software Engineer	6.10000	
2	Male	28	Group 1: 27-36	Doctor	6.20000	
3	Male	28	Group 1: 27-36	Doctor	6.20000	
4	Male	28	Group 1: 27-36	Sales Representative	5.90000	
5	Male	28	Group 1: 27-36	Sales Representative	5.90000	
6	Male	28	Group 1: 27-36	Software Engineer	5.90000	
7	Male	29	Group 1: 27-36	Teacher	6.30000	
8	Male	29	Group 1: 27-36	Doctor	7.80000	
9	Male	29	Group 1: 27-36	Doctor	7.80000	
10	Male	29	Group 1: 27-36	Doctor	7.80000	
11	Male	29	Group 1: 27-36	Doctor	6.10000	
12	Male	29	Group 1: 27-36	Doctor	7.80000	
13	Male	29	Group 1: 27-36	Doctor	6.10000	
14	Male	29	Group 1: 27-36	Doctor	6.00000	
15	Male	29	Group 1: 27-36	Doctor	6.00000	
16	Male	29	Group 1: 27-36	Doctor	6.00000	
17	Female	29	Group 1: 27-36	Nurse	6.50000	
18	Male	29	Group 1: 27-36	Doctor	6.00000	
19	Female	29	Group 1: 27-36	Nurse	6.50000	
20	Male	30	Group 1: 27-36	Doctor	7.60000	

Tableau Visualization Analysis

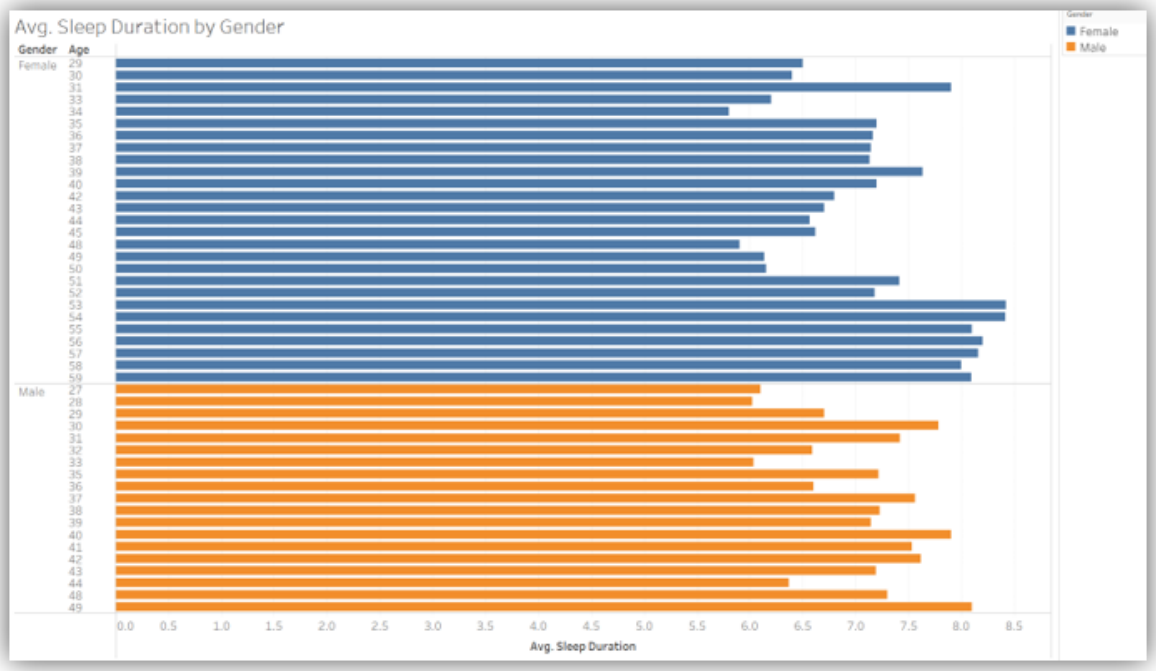
Gender-Based Sleep Patterns

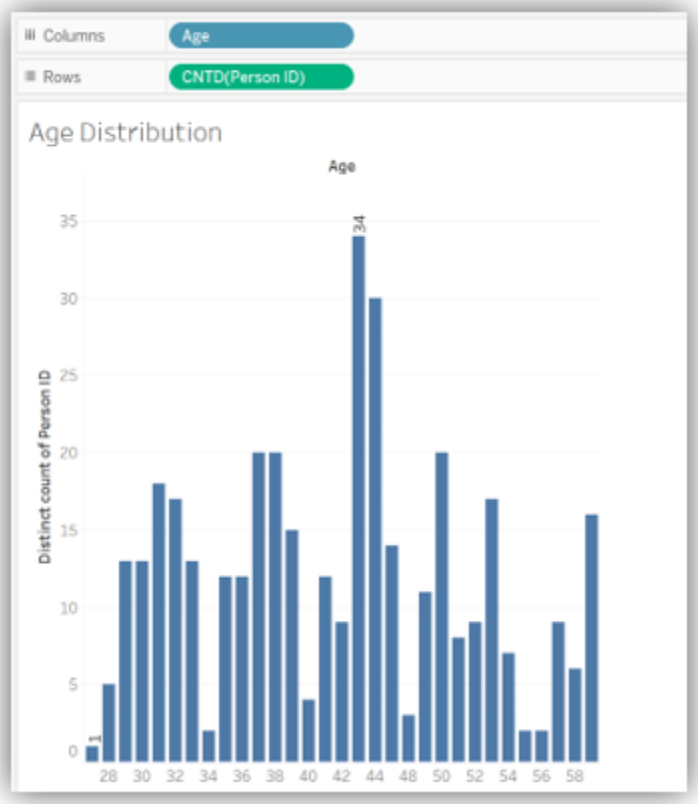
Analysis revealed:

- Women demonstrate slightly longer sleep duration

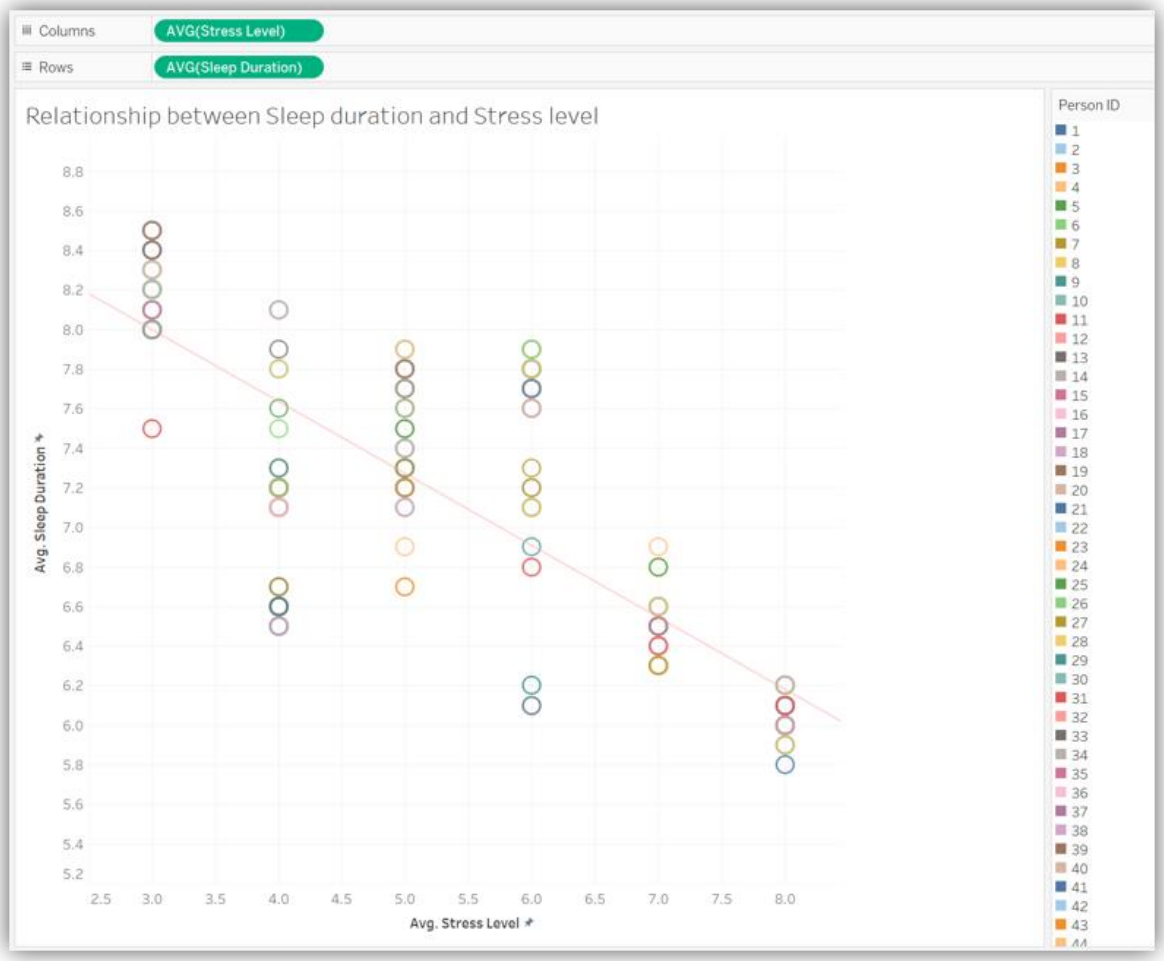
- Women report higher sleep quality ratings

- Note: Age distribution bias may influence these findings, as all participants over 49 are female


Avg. Sleep Duration by Gender
Gender
Age
Female
Male
Gender
Female
Male
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
4.0
4.5
5.0
5.5
6.0
6.5
7.0
7.5
8.0
8.5
Avg. Sleep Duration


Columns
Age
Rows
CNTD(Person ID)
Age Distribution
Age
Distinct count of Person ID
0
5
10
15
20
25
30
35
34
1
28
30
32
34
36
38
40
42
44
48
50
52
54
56
58

Stress Impact Analysis

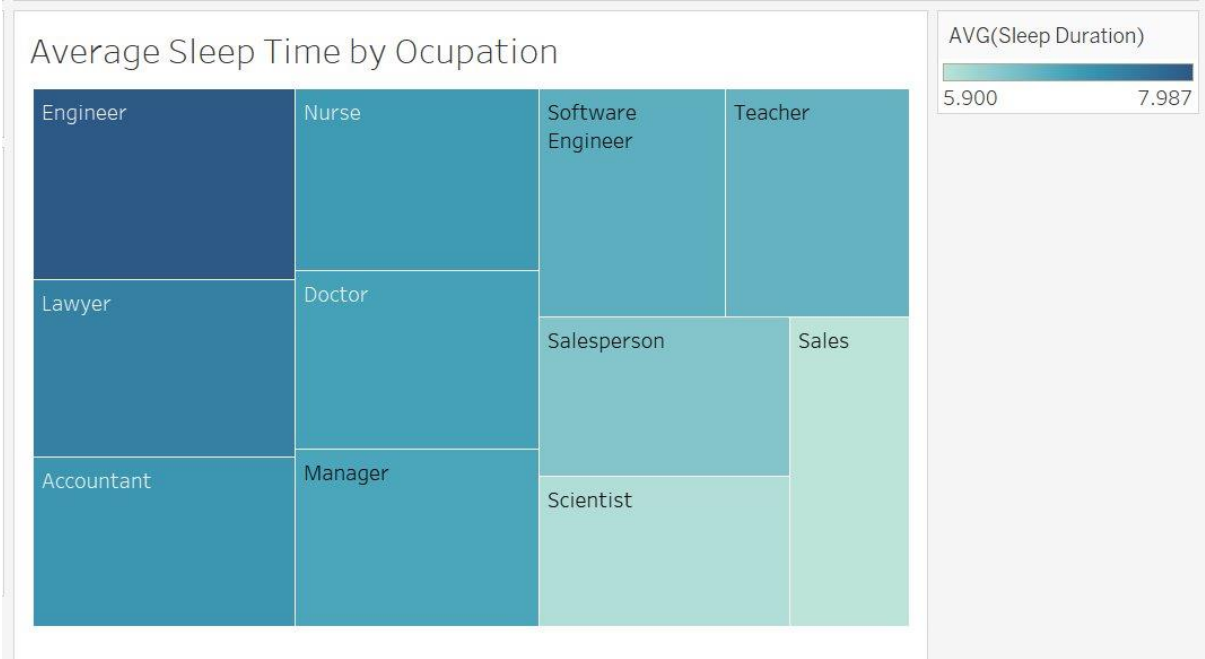


Visualization findings:

- Strong negative correlation between stress levels and sleep duration

- Clear trend showing decreased sleep duration with increased stress levels

Occupation and Physical Activity Impact

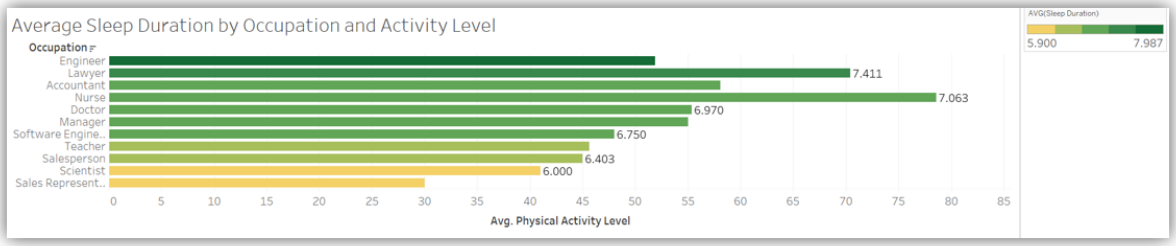


Key findings:

1. Occupation correlation with sleep duration:

- Lowest sleep duration: Sales professionals and Scientists

- Highest sleep duration: Engineers



2. Physical activity correlation:

- Positive correlation between activity level and sleep duration

- Nurses show highest activity levels

- Engineers present as outliers: high sleep duration despite moderate activity levels

Final Conclusions

Primary Findings

1. Positive correlation between sleep metrics and healthy lifestyle indicators:
   - Higher physical activity correlates with better sleep quality
   - Lower stress levels associate with longer sleep duration
   - Normal BMI category shows better sleep patterns
2. Occupation significantly influences sleep patterns
3. No significant correlation found between daily step count and sleep metrics

Data Quality Considerations

⚠ Identified Biases:

- Significant age gap between genders

- Uneven blood pressure distribution

- Non-uniform age group distribution

Recommendations for Future Research

1. Obtain larger population sample with normalized age distribution
2. Validate blood pressure distributions against established population norms
3. Ensure more balanced representation across demographic variables

---

Footnotes

1. Chattu VK, Manzar MD, Kumary S, Burman D, Spence DW, Pandi-Perumal SR. The Global Problem of Insufficient Sleep and Its Serious Public Health Implications. Healthcare (Basel). 2018 Dec 20;7(1):1. doi: 10.3390/healthcare7010001