Dataset Exploration Using ANOVA Methods:

Sleep Health and Lifestyle

Wendy Mendoza Gutierrez

Metropolitan State University

Amina Hakim Mohamud

Metropolitan State University

Introduction

The Sleep Health and Lifestyle dataset contains information on 374 individuals. It aims to capture the relationship between sleep patterns and various demographic and lifestyle factors. For our analysis, we are focusing specifically on four variables: gender, physical activity levels, occupation, and sleep duration.

Dataset Selection

The Sleep Health and Lifestyle dataset was chosen for its relevance in everyday life and the desire to recognize sleep patterns. The key variables allow us to explore how personal characteristics, and daily habits may influence sleep patterns. This synthetic dataset, created by Kaggle user Laksika Tharmalingam for illustrative purposes, does not contain real patient data.

Research Questions

1. Is there a significant difference in mean sleep duration among the four occupation categories?
2. Is there a significant difference in mean sleep duration between gender and the physical activity groups, is there a significant difference between gender and physical activity group?

Key Variables for Analysis

Gender: An independent categorical variable with two levels, male and female. This variable allows us to examine the potential differences in sleep patterns among the genders.

Physical Activity Level: For analytics purposes, we have categorized physical activity into an independent categorical variable with two groups, low and high. Measured in minutes of daily physical activity, this categorization allows us to assess how different intensities of physical activity may impact sleep duration.

Occupation: An independent categorical variable that we have categorized into four levels: Business/Finance (Accountant, Sales Representative, Salesperson, Manager), Education/Social Science (Lawyer, Teacher), STEM (Software Engineer, Scientist, Engineer), and Healthcare (Doctor, Nurse). This variable enables us to investigate how different work demands and schedules may affect sleep patterns.

Sleep Duration: Measured in hours, sleep duration is the primary dependent continuous numeric variable in this analysis. Sleep duration allows us to assess how the other factors might influence the amount of sleep individuals receive daily.

Variable	N	Mean	Std. Dev.	Min	Max
Gender	374				
Male	189	7.037	0.693	5.9	8.1
Female	185	7.230	0.879	5.8	8.5
Physical Activity Level	374				
Low	155	6.808	0.872	5.8	8.5
High	219	7.362	0.646	5.9	8.2
Occupation	374				
Education/Social Science	87	7.079	0.439	6.3	7.9
Business/Finance	72	6.761	0.446	5.9	8.0
STEM	71	7.806	0.739	5.8	8.5
Healthcare	144	7.017	0.928	5.9	8.2
Sleep Duration	374	7.13	0.796	5.8	8.5

Hypothesis

One-way ANOVA analysis

$$H_0: \mu_{Education} = \mu_{SocialScience} = \mu_{Business} = \mu_{Stem} = \mu_{Healthcare}$$

H_a: At least two μ's are different

Two-way ANOVA analysis

Main effect of Gender:

$$H_0: \mu_{Male} = \mu_{Female}$$

$$H_a: \mu_{Male} \neq \mu_{Female}$$

Main effect of Physical Activity:

$$H_0: \mu_{Low} = \mu_{Moderate} = \mu_{High}$$

H_a: at least 2 μ's are different

Main effect of interaction between Gender and Physical Activity:

H_0: There is no interaction between Gender and Physical Activity

H_a: There is an interaction between Gender and Physical Activity

Methodology

Our analytical approach used Analysis of Variance (ANOVA) to determine whether there are statistically significant differences in sleep duration across various factors. ANOVA is a statistical method that compares means across multiple groups to identify significant differences. One-way ANOVA will be used to assess the effect of occupation on sleep duration. One-way ANOVA is appropriate because occupation is a single independent variable (IV) which is categorical with more than 2 categories (Education and Social Service, Business and Finance, STEM and Healthcare), and we want to compare mean differences in sleep duration across these groups.

Two-way ANOVA will be used to examine the effects of gender and physical activity on sleep duration, as well as the interaction between these two independent variables. Two-way ANOVA allows us to analyze two categorical IVs simultaneously and determine both their individual main effects and their potential interaction effect on sleep duration.

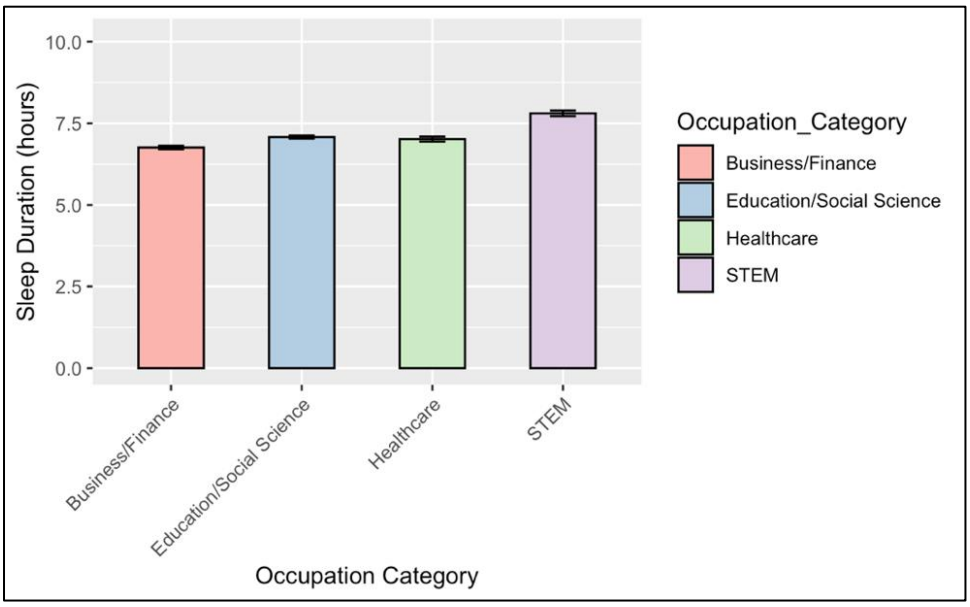
Both one-way and two-way ANOVA are used to compare the mean differences between groups. For our analysis, we can assume sleep duration is normally distributed, meeting a key assumption of the ANOVA test.

Our data preparation focused on appropriate categorization of key variables. Raw occupation data was standardized into defined categories to enable meaningful group comparisons. Similarly, Physical Activity Levels were transformed into 2 distinct categories (low and high) to create appropriate grouping for the ANOVA tests.

Conclusion

One-way ANOVA Analysis for Research Question 1

The analysis has an F statistic of 28.45 with a p-value of $< 2e-16$. Since the p-value is less than the significance level ($\alpha = 0.05$), we reject the null hypothesis (H_0). This provides sufficient evidence to support the claim that there is a significant difference between the mean Sleep Duration among the Occupation categories. The Tukey test further reveals that there is a significant difference between STEM and Business/Finance, STEM and Education/Social Science, and STEM and Healthcare.



This ANOVA analysis reveals that occupation significantly affects sleep duration, prompting us to reject the null hypothesis. The Tukey test proves this discrepancy by indicating that there is a significant difference between STEM and Business/Finance, STEM and Education/Social Science, as well as STEM and Healthcare. Interestingly, STEM workers report more sleep than expected, contradicting our assumption that they would have the fewest number of hours slept. The short error bars from the graph indicate low variability in sleep duration among these groups, suggesting that the values are highly concentrated. For this ANOVA analysis, no additional data is required to interpret the results.

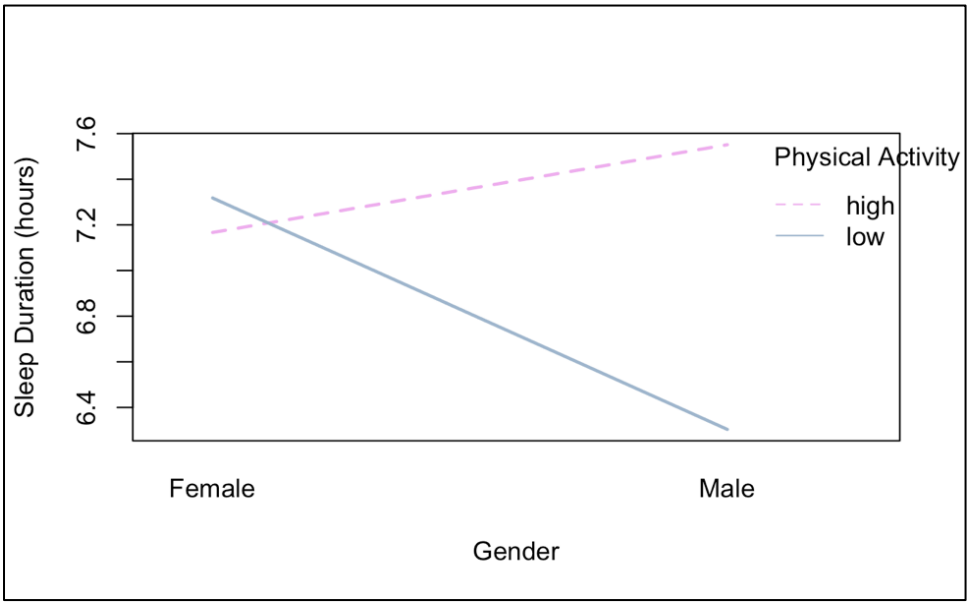
Two-way ANOVA Analysis for Research Question 2

The main effect of Gender shows an F statistic of 8.055 with a p-value of 0.00479. Since the p-value is less than the significance level ($\alpha = 0.05$), we reject the null hypothesis (H_0). There is enough evidence to support the claim that the mean Sleep Duration is significantly different between Males and Females.

The main effect of Physical Activity Level shows an F statistic of 64.430 with a p-value of $1.34e-14$. Since the p-value is less than the significance level ($\alpha = 0.05$), we reject the null hypothesis (H_0). There is enough evidence to support the claim that the mean Sleep Duration is significantly different between the physical activity groups.

For the Interaction test, $F = 102.485$ with a p-value of $< 2e-16$. Since the p-value is less than the significance level ($\alpha = 0.05$), we reject the null hypothesis (H_0). This suggests that there is a significant interaction between Gender and Physical Activity Level.

To examine where the significant interactions lie, for males: $F = 700$ with a p-value of $< 2e-16$, indicating significance difference in sleep duration between the two physical activity groups. In contrast, for females: $F = 1.337$ with a p-value of 0.249, indicating no significant difference in sleep duration between the two physical activity groups.



In this Two-way ANOVA analysis, we discovered that there exists a significant difference in sleep duration among the genders, indicating males and females experience different amounts of sleep. This finding challenges our initial assumption that gender would not significantly impact sleep duration, making this finding surprising. Additionally, we found a significant difference

between the physical activity groups and sleep duration, confirming our assumption that physical activity would have a significant impact on sleep duration. Incorporating a variable such as "intensity" for physical activity levels could further help researchers investigate this issue.

The interaction analysis suggests that both gender and physical activity levels must be considered to fully understand their combined impact on sleep duration. For males, there is a significant difference in sleep duration between the two physical activity groups, while for females, there is no significant difference. This is evident in the interaction plot, which visually illustrates this relationship. These findings indicate a significant interaction between physical activity and gender on sleep duration.