Angel Gutierrez Sanjuan February 11,2025

Sleep Health

Objective: In this project I aim to continue my work from hw3, this means I will be using the same dataset with the aim to further explore its analyzes on individuals' sleeping lifestyle. I want to explore any possible trends and lifestyles that can impact an individual's sleeping habits.

Acquisition:

The dataset was acquired from kaggle at https://www.kaggle.com/datasets/uom190346a/sleep-health-and-lifestyle-dataset . This dataset was collected by an individual named Laksika Tharmalingam, the dataset has a gold ranking with 86,149 downloads; this backs up the reliability of the dataset. The goal of using this dataset is to analyze individuals' sleeping habits and observe any conditions that may impede restful nights sleep. This can include the amount of physical activity all the way to occupation.

Early Analyzes: To initiate any work on the dataset I will break down the types of data given. Looking at the dataset we get a total of 374 observations with 13 variables. The dataset has 8 numerical categories and 5 categorical categories.

Numerical Categories:

- Person ID - Ascending values in order to keep track of the number of individuals.
- Age - Individuals age
  - Mean : 42.18
  - Median : 43
  - Standard Deviation : 8.67
  - Minimum: 27
  - Maximum: 59
  - Quartiles: 0% -> 27 25% -> 35.25 50% -> 43.00 75% -> 50.00 100% -> 59

- Sleep Duration - This value is significant to the amount of sleep received.
  - Mean: 7.13
  - Median: 7.2
  - Standard Deviation: 0.79
  - Minimum: 5.8
  - Maximum: 8.5
  - Quartiles: 0% -> 5.8 25% -> 6.4 50% -> 7.2 75% -> 7.8 100% -> 8.5

- Quality of Sleeps - This significant value signifies the quality of an individual's sleep from a rating of 1 to 10 with 1 being the worst and 10 being the best.
  - Mean: 7.31
  - Median: 7

    - Standard Deviation: 1.19
    - Minimum: 4
    - Maximum: 9
    - Quartiles: 0% -> 4 25% -> 6 50% -> 7 75% -> 8 100% -> 9

- Physical Activity Level - Significant values that signifies the level of physical activity from a range of 0 to 100 with 0 being the lowest level of physical activity and 100 being the highest level of physical activity.
    - Mean: 59.13
    - Median: 60
    - Standard Deviation: 20.83
    - Minimum: 30
    - Maximum: 90
    - Quartiles: 0% -> 30 25% -> 45 50% -> 60 75% -> 75 100% -> 90
- Stress Level - The values signify the stress in individuals, the rating ranges from 0 to 10 with 0 being the least stressed and 10 being really stressed.
    - Mean: 5.38
    - Median: 5
    - Standard Deviation: 1.77
    - Minimum: 3
    - Maximum: 8
    - Quartiles: 0% ->3 25% ->4   50% ->5 75% ->7 100% ->8

- Blood Pressure - Blood pressure is taken of each individual, but in order to get the values we must convert the values using MAP = DBP + (⅓) * (SBP - DBP)
    - Mean: 128.55
    - Median: 130
    - Standard Deviation: 7.74
    - Minimum: 115
    - Maximum: 142
    - Quartiles: 0% ->115 25% ->125   50% ->130 75% ->135 100% ->142

- Heart Rate - It is the frequency of the heartbeat per minute, this lets us know if there are any abnormalities based on a person's weight.
    - Mean: 70.16
    - Median: 70
    - Standard Deviation: 4.13
    - Minimum: 65
    - Maximum: 86
    - Quartiles: 0% ->65 25% ->68   50% ->70 75% ->72 100% ->86

- Daily Steps - A count of how often individuals walk on a daily basis
    - Mean: 6816.84
    - Median: 7000
    - Standard Deviation: 1617.91
    - Minimum: 3000
    - Maximum: 10000
    - Quartiles: 0% ->3000 25% ->56000 50% ->7000 75% ->8000 100% -> 10000

Our Numerical Categories allow us to get an insight on the amount of physical activity on average is done. It helps us understand the data in its raw form before seeing visuals that will demonstrate different information.

Categorical Categories:

- Gender - This category lets us know if the individuals are either male or female
- Occupation - This lets us know an individual's occupation, trends could be potentially found between weight and occupation.
- BMI Category - It lets us know if individuals are either “Overweight”, “Normal”, or “Obeses”
- Sleep Disorder - It indicates if individuals have “None” (no sleeping disorder), “Sleep apnea” or “Insomnia”.

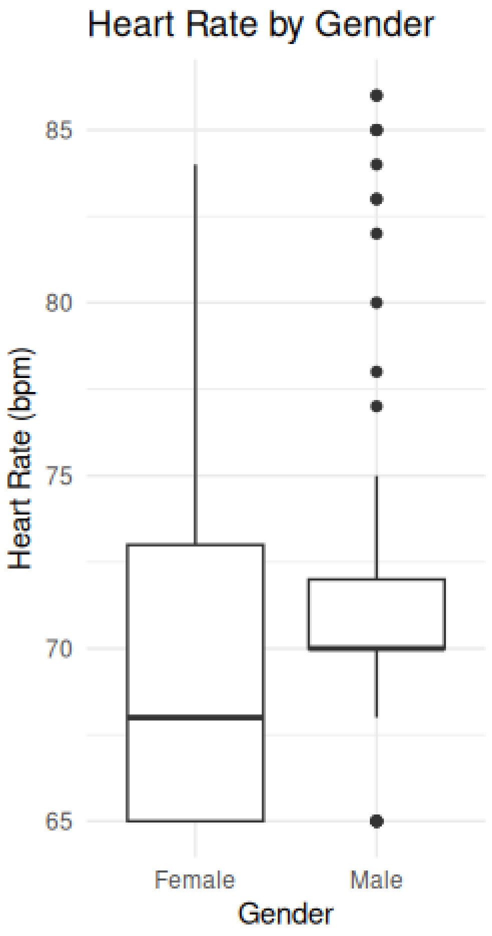
Our Categorical values give us insight on how people can be placed whether it be by occupation, sleeping disorder, gender etc.. It allows us to break data down with numerical values to find correlations.

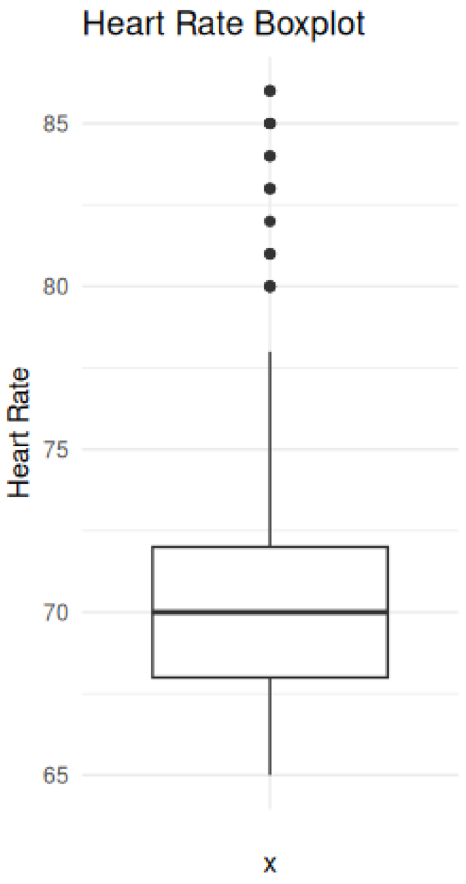
Cleansing: The next step of the data analysis pipeline is to make sure that the integrity of the data is not compromised. In order to this we will employ techniques to clean, fill in or make sure no duplicates are found. My initial step here is to make sure that there are no missing values, if the dataset has missing values then they must be removed or replaced. While looking at the data I made sure it came from a relevant source, checking kaggle this data set had 86.2 thousand downloads and most recently it had 3230 downloads in the last month. Using R programming I checked for missing values and received no missing values. In addition to this I had to make sure that the data was not compromised with extreme outliers, extreme outliers can skew any visualizations; thus invalidating the integrity of the visualization. Finally I also checked for any duplicate values that might skew data visualization in condensed groups, this will prevent any wrong clustering from happening. It is normal to see some clustering but duplicate values will have overlapping clustering. These are the steps I took in order to make sure that the dataset was secured, first I checked the reliability of the dataset, next I checked for missing values, afterwards I checked for extreme outliers that might skew any graphs and finally I checked for any duplicate values. While making the histograms I realized that under the BMI. Category I had 4 categories; normal weight, normal, obese and overweight. This is contradictory when making a visual since it divides a variable into two different sections. In order to fix it, using R programming I switched the variables normal weight into normal.

Transformation: Transformation in this project is done by fixing any values that will normalize values. In this project I ran into two instance where the data variables had to transform; firstly I had to change the values of individuals blood pressure since it is formatted as a fraction this made it into char value, next I had to fix the column of BMI since it regurgitated the same information twice using both "normal weight" and "normal" as different values. This can be found in the R script.

Analyzes (PART 1):

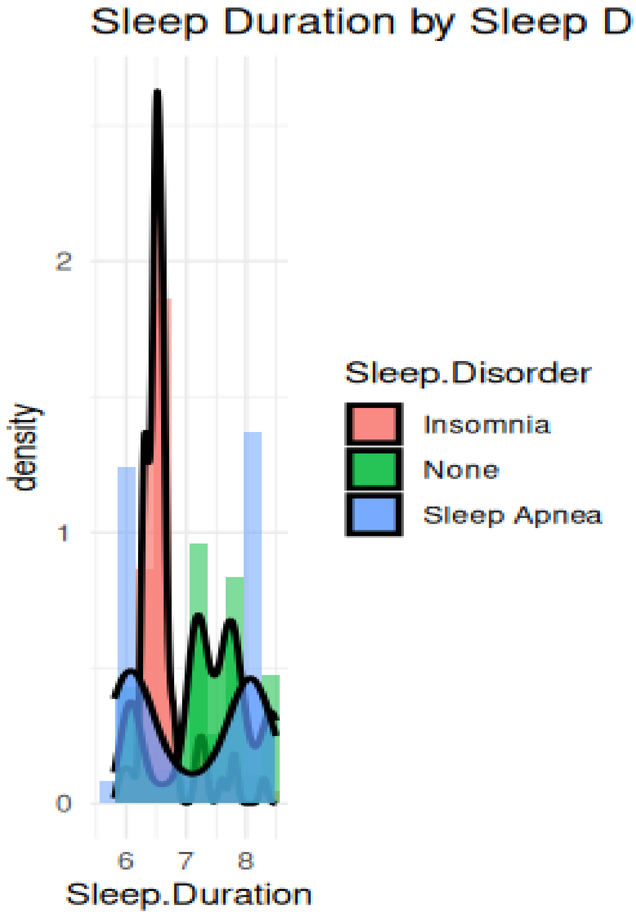
Boxplot of heart rate, since we are looking at individuals' quality of sleep getting insight of the heart rate might help determine trends. Medically individuals with higher heart rates may have higher body mass, athletes on the other hand have a lower heart rate.

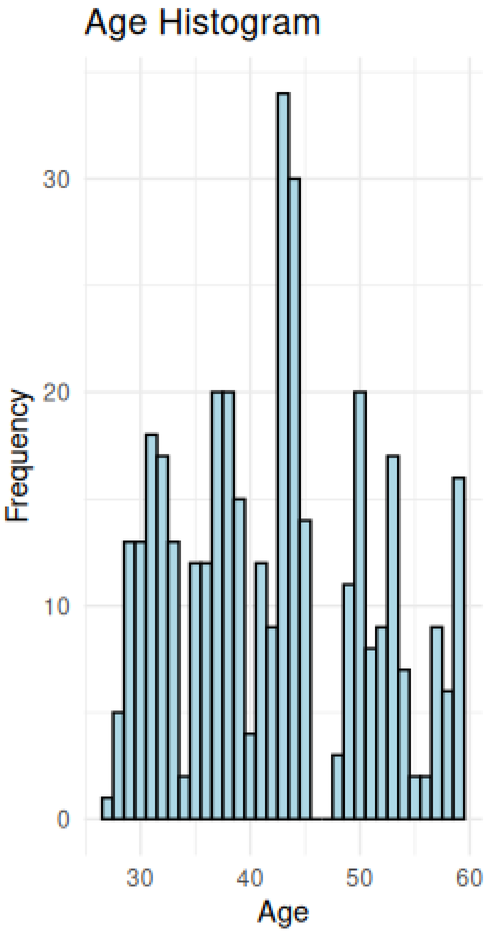




Getting the general heart rate average for all individuals shows that outliers above 80 beats per minute to their heart rate. When looking at the boxplot that separates the gender it shows that females have a general mean around 68 beats per minute and males on average are around 73 bpm. This can indicate that females may have a healthier lifestyle compared to males. The whiskers for the female boxplot show a larger heavier amount of females while the males have higher and in this case heavier individuals in the outliers.

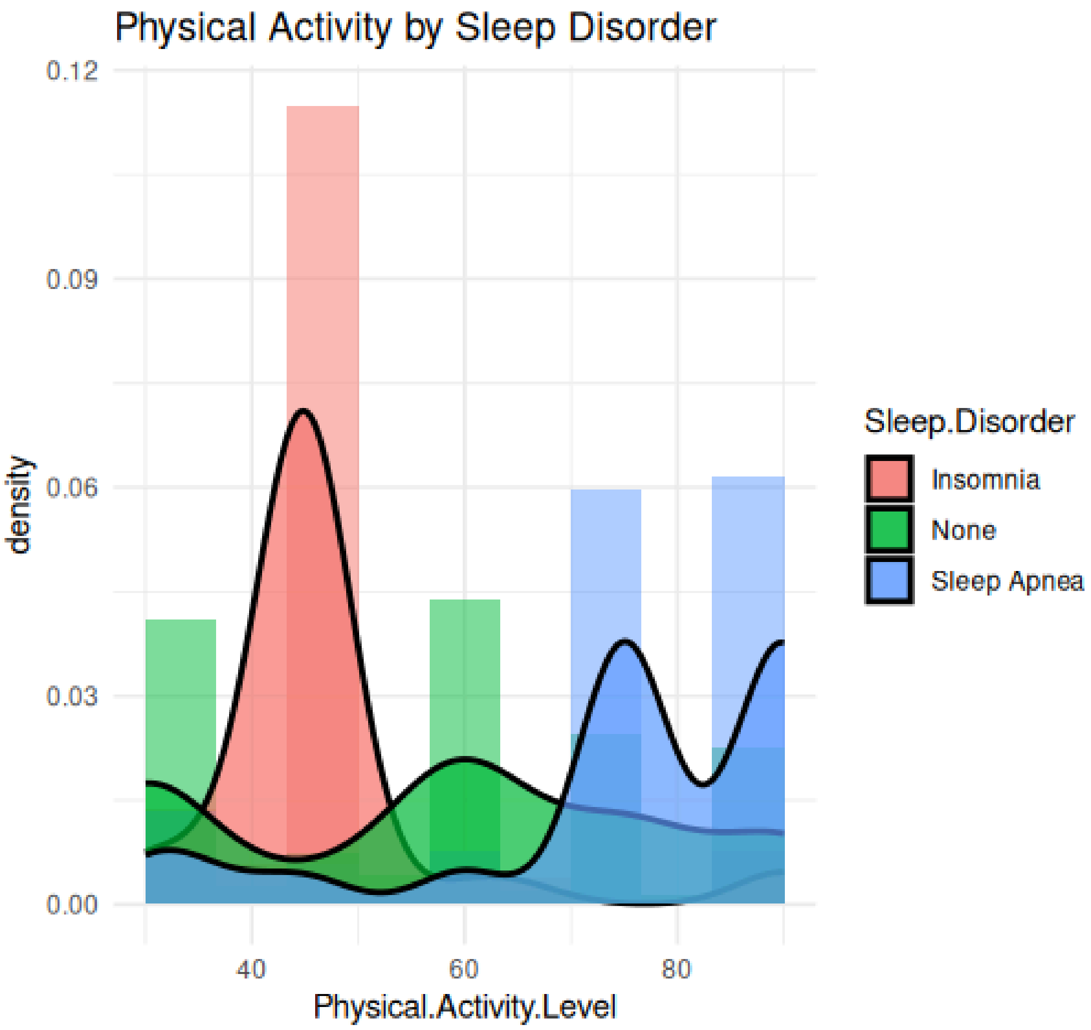
When looking at the histograms I want to make sure it is relevant to the dataset so I took a histogram from hw3 that seemed appropriate for this project. In the histogram we can see that the density of individuals with either Insomnia, No disorders and individuals with sleep apnea are contrasted with their sleep duration.



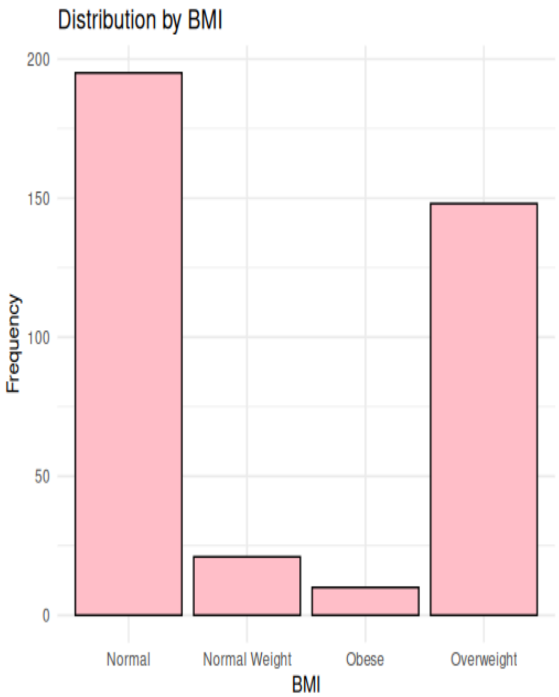


Angel Gutierrez Sanjuan February 11,2025

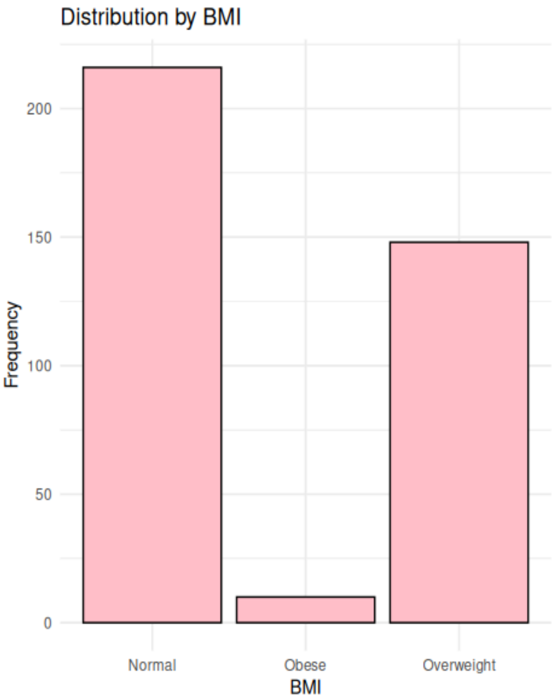
In addition to using some hw3 histogram to make some new ones, I made a histogram that checks the density or frequency of individuals by age. Based on the histogram we can conclude that the highest density are individuals around their mid 40’s. This information can be relevant since it gives us a probable estimate of expected health, this just means that older individuals might on average have more care about their health but less physical activities. The histogram covers physical activity; we are made aware that individuals with less physical activities are most likely to have insomnia.



Angel Gutierrez Sanjuan February 11,2025

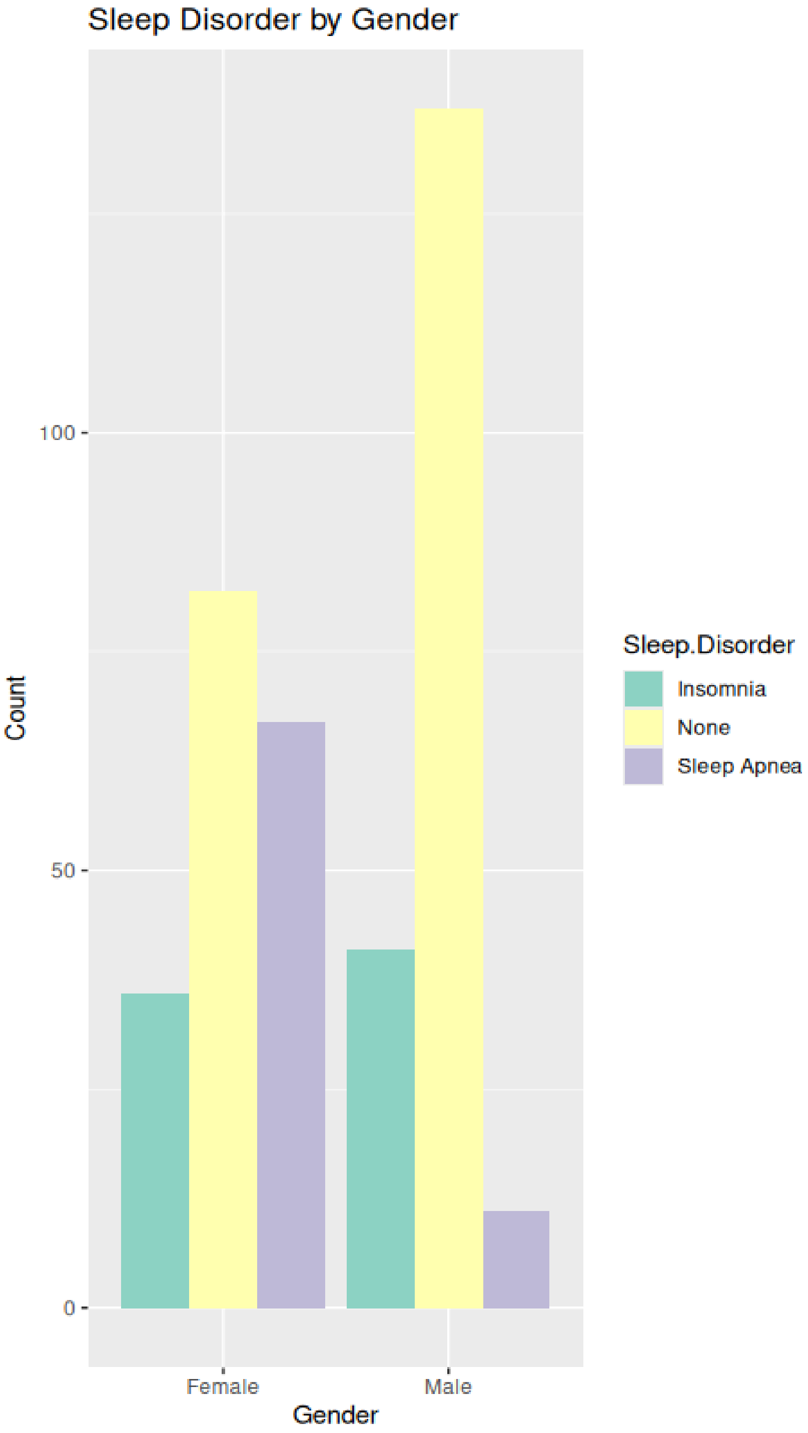


FIX
===>



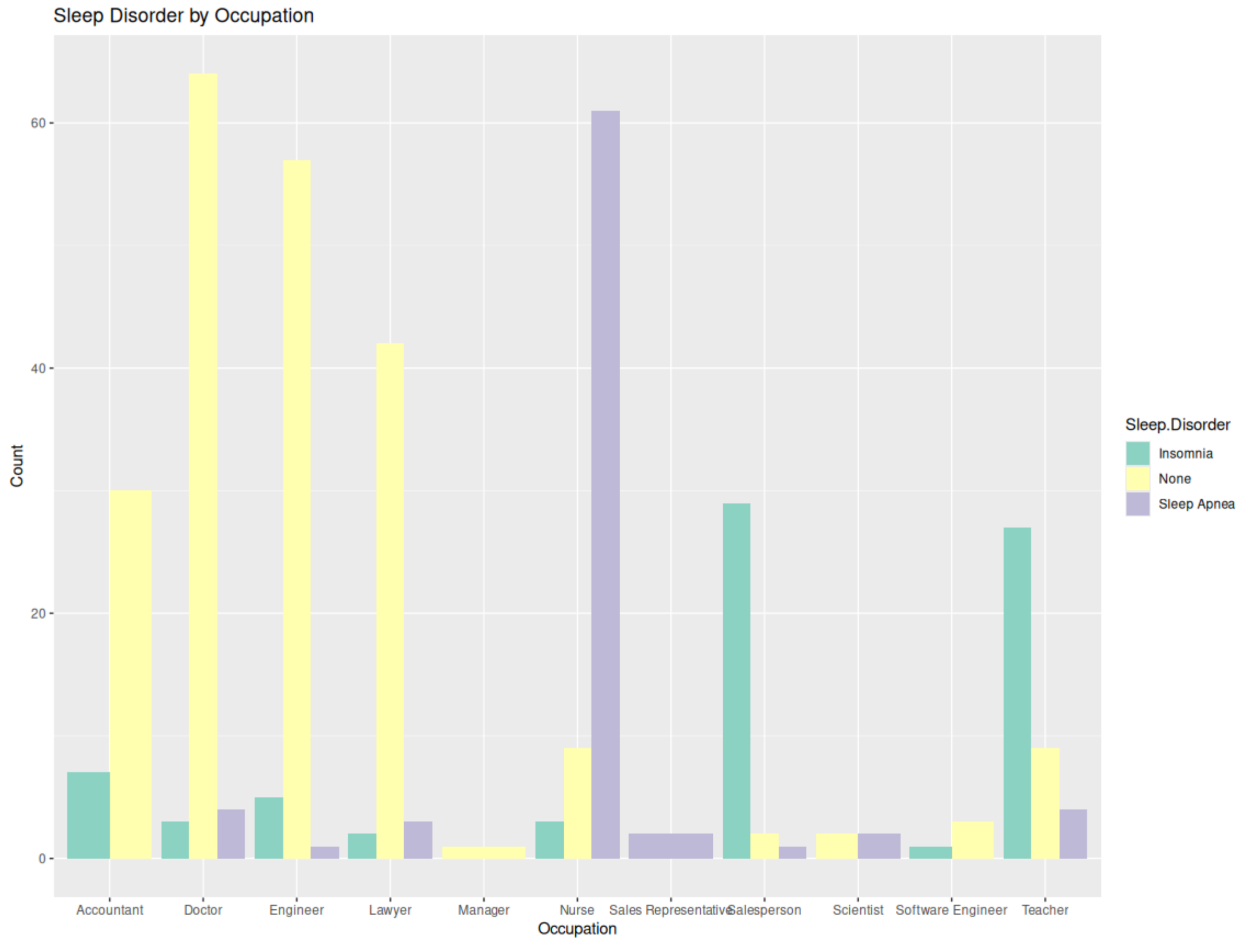
The Histograms above show the amount of physical activity done by the density of individuals, alongside it there's another histogram that conveys the amount of individuals with both normal weights and abnormal weights. Physical Activity covered by the amount of frequent individuals, the amount of physical activity level is distributed across individuals with sleeping disorders. We can make some assumptions with this data, first we can see that people with insomnia have the least amount of physical activity, while people with sleep apnea have the most physical activity . People with no sleeping disorder have a wide range of physical activities ranging from least amount to the most amount. On the other histograms it demonstrates the original histogram while the fixed histogram corrects any errors, we can observe that people are determined by Normal, Obese and Overweight. When we talk about people's BMI weight it is determined by different ranges that are calculated by body weight relative to individuals height , Normal Weight is measured by 18.5 - 24.9 , Overweight is measured by 25 - 29.9 and Obesity is measured by 30. The Histogram shows that the majority of individuals have a normal weight (190~), while obese individuals (20~) are the least amount and overweight individuals are the second highest majority with 150 people.

Barplots used in this project aim to highlight differences in groups and categories here are some of the boxplots done. This bar plot contrasts the Sleeping disorder by gender; immediately we can observe that males on average tend to have no sleeping disorder with a range of 130~ individuals but we can also note that on average most females also don't have sleeping disorders. It also stands out that females have a higher chance of having sleep apnea. It is also notable that both females and males have on average the same amount of insomnia cases.

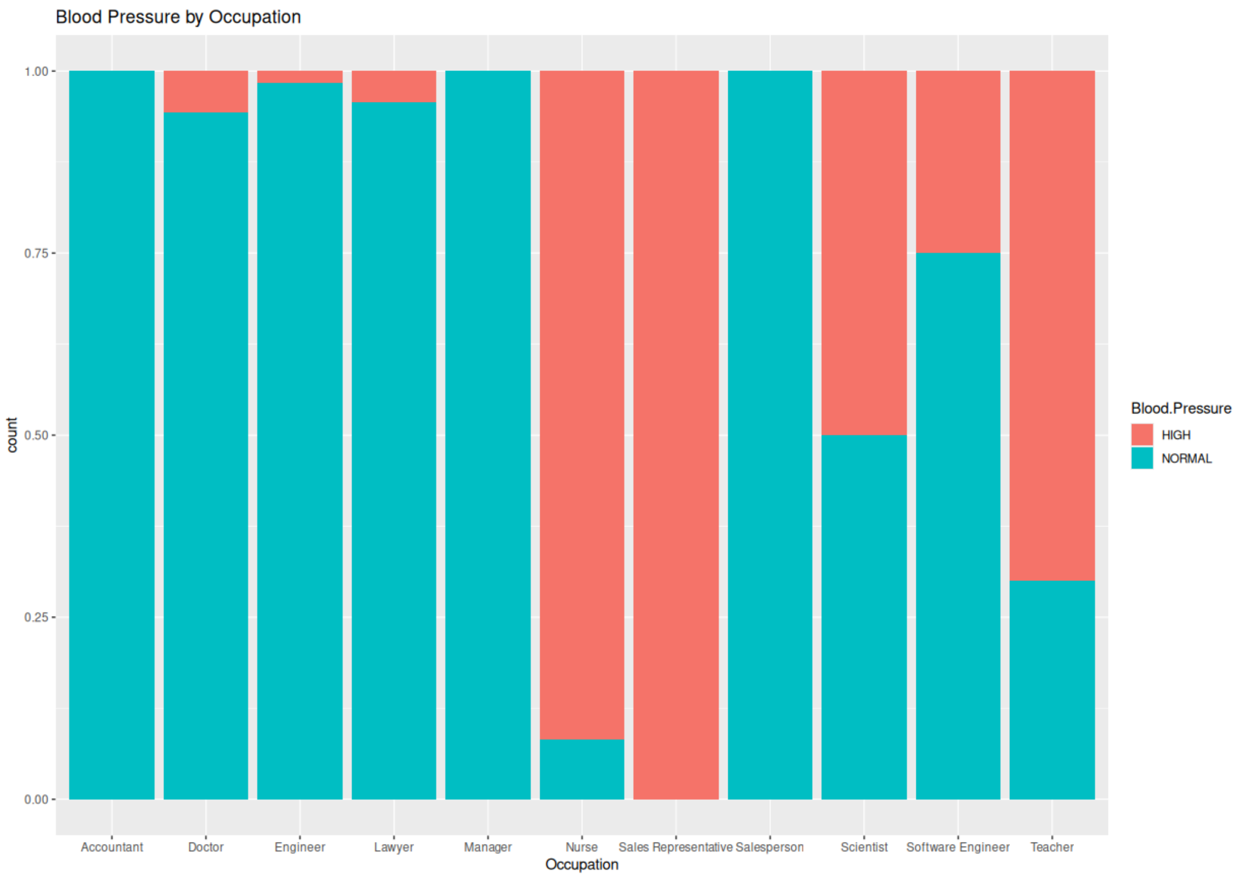


Angel Gutierrez Sanjuan February 11,2025

In this barplot we get a visual that marks sleeping disorder by occupation. Initially we can see that accountants, doctors, engineers, lawyers and managers have the lowest reported cases of any sleeping disorder. Next we observe a majority increase of sleep apnea with nurses; nurses by far the most dominant group in the data set to have this condition out of all occupations. Additionally we can observe that the majority of insomnia cases are found in school teachers and sales persons. We can now make an educated conclusion in what occupations may have the highest stressors to lead to conditions; those being Nurses, School Teachers and Sales Persons.



Here we have a bar plot that demonstrates blood pressure between different occupations. The red signifies High Blood pressure while the blue holds normal blood pressure. Observing this barplot we immediately noticed that the sales representative all had high blood pressure, following this along we have nurses with the second highest blood pressure. The individual groups of accountants and salespersons remain the only groups without any high blood pressure. Having a salesperson not having high blood pressure gives us an interesting insight especially since they are the occupation with the highest insomnia levels (previous barplot) . We can hypothesize that most salespeople don't really have stressors to increase the blood pressure but the lifestyle can cause insomnia to develop. Scientists seem to have an even number of individuals that have a normal blood pressure and those with high blood pressure. Teachers also seem to have an abnormally high amount of blood pressure. A leading trend of the highest groups with high blood pressure are professions that deal with a high amount of people. In fifth place we have software engineers with a high amount of cases with high blood pressure.



Angel Gutierrez Sanjuan February 11,2025

Conclusion: This project was really fun. It allowed me to learn a lot lot more, it helped improve my analytical skills. I think this dataset was challenging enough where I did have to go back and fix values. It took me a while when making some graphs to realize my mistakes. I overlooked some initial stuff during the transformation process of the data processing pipeline. Next project I will try and get a tougher dataset, in order to further sharpen my skills.