Kelley Denny
Student ID 800563382
Intro to Data Mining
Project 3 - Regression Modeling

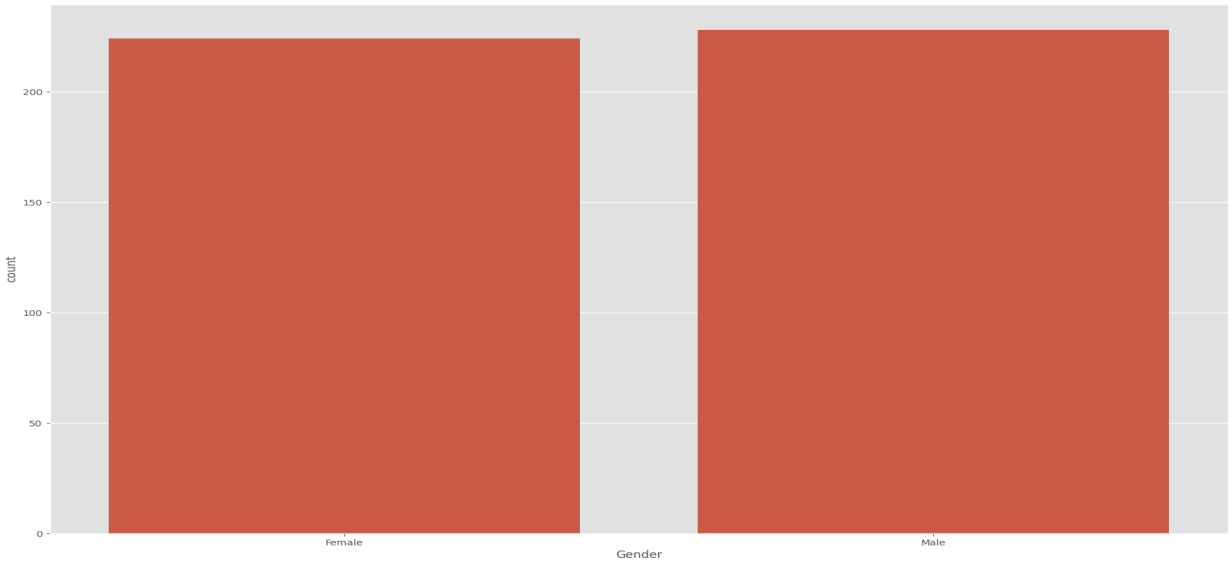
Dataset: Sleep Efficiency

Sleep is an important factor in bettering physical and mental health. It is the time in which the body and brain repair and remove toxins that contribute to diseases and disorders (NIH, 2021). There are a lot of factors that can decrease the overall amount of sleep, or decrease the quality of sleep that we do get. Some of these factors include busy schedules, anxiety, technology, and family (UC Davis, 2024). Understanding the factors that contribute to a better overall sleep quality score can help people to establish better sleep habits or to reinforce existing habits that contribute to better overall sleep quality.

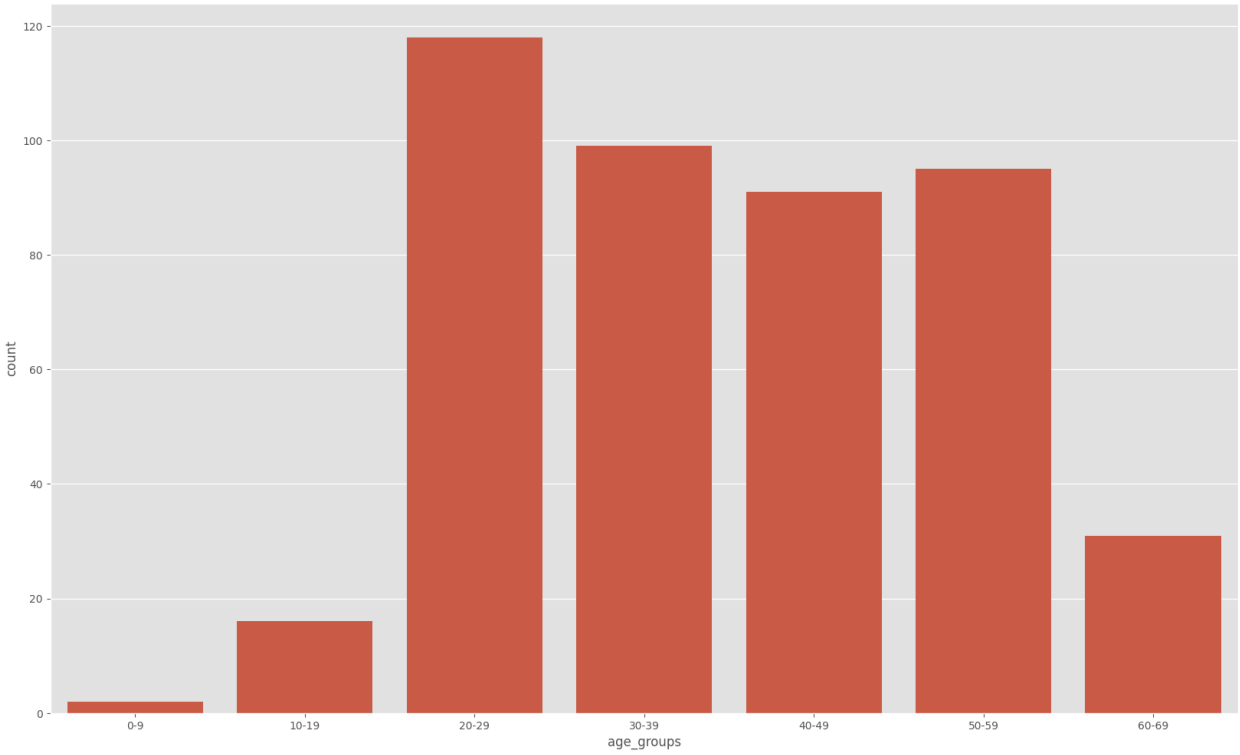
The dataset for this project is obtained from kaggle, it is a sleep efficiency dataset with 452 observations and 15 columns. The columns include data such as age, gender, bedtime and wake-up time, light/REM/deep sleep percentages, how many awakenings through the night a person had, whether the person smokes, alcohol consumption, caffeine consumption, and frequency of exercise. These are different factors that either affect the quality of sleep (like smoking and alcohol consumption, exercise frequency) or contribute to a lower or higher sleep efficiency score (like sleep duration, light/REM/deep sleep percentage, bedtime and wake-up time).

The purpose of this analysis is to create several regression models to predict the sleep efficiency score based on the given independent variables. Regression is a type of machine learning based on statistical analysis that seeks to determine the relationship between the predictor variables and the target variable. It is used to predict a continuous value and to determine the highest correlation between the variable and the target. Linear regression is graphically depicted by a straight line that should be as close as possible to all the given data points in a scatter plot, the slope of the line communicates how the change in one variable affects the other(s), and the y-intercept of the line shows what the value of the dependent (target) variable will be when the independent (predictor) variable is 0. The formula for simple linear regression is $Y = a + bX + u$ where Y is the dependent variable meant to be predicted, a is the y-intercept, b is the slope of the line, X is the independent variable, and u is the residuals, or errors between the predicted value and the actual observed values. For multiple linear regression the formula is similar, but there are multiple bX values in the equation, up to n number depending on how many independent variables are being studied to predict the dependent variable.

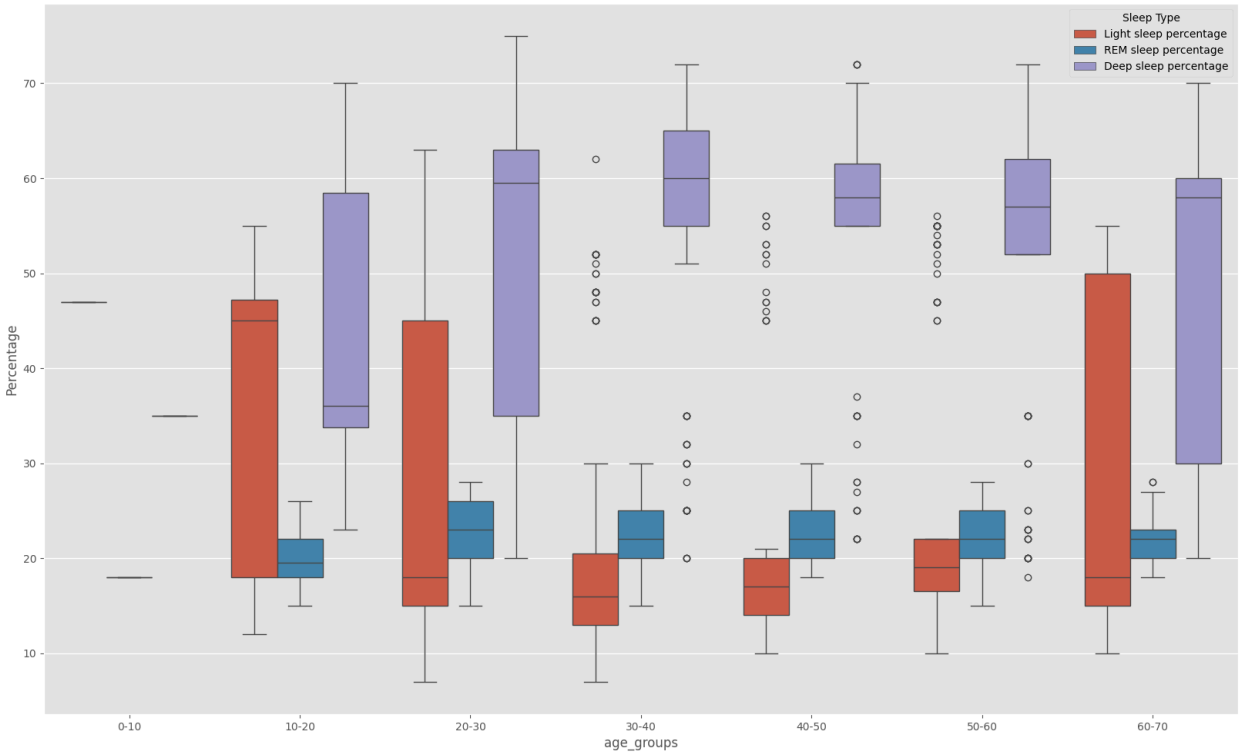
To understand the dataset, I first chose to look at the distribution of genders. In this dataset there were only two genders observed, males and females. The dataset was almost evenly distributed, with 228 males and 224 females observed.



I then broke the ages down into age groups with bins of 10 year ranges. I then plotted the observations into these age group bins and found that the ages of 20-59 being the vast majority of observations, and 0-9 years having the least amount of observations (2).



I then created pair plots for the whole dataset and found that most of the features did not seem to have a relationship with each other. The exception to this was that deep sleep percentage and light sleep percentage did have somewhat of a linear relationship, as did deep sleep percentage and sleep efficiency. Next I wanted to see a boxplot of the breakdown of sleep percentages among the different age groups to see what type of sleep the different age groups were getting.



I noticed that there were quite a few outliers in the deep and light sleep categories in the older age groups (ages 30 +) and that younger ages and the oldest age group got more light sleep and deep sleep while REM sleep seemed to decrease with age. This was an interesting insight to uncover while in the exploratory data analysis stage.

For data pre-processing I checked the data set for null and duplicate values. I found that there were minimal null values or missing data, so I chose to fill the missing values with the mean value of the column, and found that there were no duplicate values. The gender and smoking status columns had to be changed to numerical values in order to use them in regression analysis, so I used one-hot encoding to change them to numeric values. The awakenings and caffeine consumption columns were up to 8 decimal places so I rounded those to 2 decimal places for easier readability. I also noticed that the bedtime and wakeup times were in date format, and for this type of analysis it makes more sense to only see the hour of the bedtime and wake-up time so I pulled the hour out of the date and kept it while dropping the rest of the date info. Finally I did not have any use for the id column as it merely served as an index so that column was dropped.

For experiment 1 I ran a linear regression model, choosing the sleep efficiency as the target, and leaving all the other variables as the predictor variables. I chose to view MSE, RMSE, and R2 scores to evaluate the model performance. I ran this regression model with the raw, original data and did not check for correlations between the features aside from looking at the pairplots. Once I ran the regression model, the metrics obtained were MSE of .0047, RMSE of .0686, and an R2 score of .7429. This shows that the error between the actual and predicted values (MSE and RMSE) are quite small, meaning the model did quite well at predicting the sleep efficiency score. An R2 score of .7429 says that about 74% of the variance in the sleep

efficiency scores can be explained by the interplay of all the other features or variables of the dataset.

For experiment 2 I chose to make a heatmap of the features to see which ones had the highest correlation and found that light sleep percentage was highly correlated with deep sleep percentage and sleep efficiency so I chose to drop this variable. I also chose to drop the wakeup time column because there were NAN values showing after changing the values to numerical with one-hot encoding so this was giving some issues with the modeling. I then chose to run another linear regression model for this experiment, selecting a different random state to get a different split of the training and testing data.I then obtained the evaluation metrics. These metrics were as follows: MSE of .0039, RMSE of .0624, and R2 score of .7909. These metrics were even better than the previous linear regression model, showing that perhaps the highly correlated feature of light sleep percentage was causing some extra “noise” in the model.

Finally for experiment 3 I ran a Lasso regression model, which works by adding what is known as a regularization term (or penalty) to the loss function of ordinary linear regression, meaning that it changes some of the coefficients (or variables) down to zero, which then acts as a means of feature selection for the model. It essentially drops the features whose coefficients are reduced to zero with the addition of the penalty for loss, and then runs the model with only the features that are left, which in theory will reduce overfitting of the model. The lasso regression model again was given a different random state to change the split. The evaluation metrics for the Lasso regression model were quite a bit lower or worse than those of the linear regression models, with an MSE of .0112, an RMSE of .1058 and an R2 score of .3946. The lasso regression model could only explain about 40% of the variance of the sleep efficiency score based on the features that were selected in the Lasso regression model.

The impacts that this analysis could have are pretty numerous and there could be some good and some negative impacts. Firstly, the positive influence that this type of analysis could have on people is that it can show people which factors both positively and negatively affect their overall sleep quality, which can lead to those people changing their habits or introducing new and better habits to improve their sleep, thereby improving their overall health. This type of analysis can also be useful for companies that offer fitness trackers as offering suggestions to users to improve sleep quality could be beneficial. Discussing the factors that negatively affect sleep would be more useful if the analysis correctly identifies the factors that both positively and negatively affect the sleep quality scores. As for the negative impacts, this study could include an overfitted model that would not perform well on any additional, new data and could be far less accurate when predicting sleep quality of users of fitness trackers for example. This type of study also includes some personal health data from users and if there are not adequate checks in place to ensure privacy of those studied, privacy can be easily compromised.

In conclusion, this study was quite interesting to analyze, even though the dataset was smaller in comparison to others that I have studied in the past. With more observations, I feel that these models can be trained and tuned to be more accurate with better predictions. Linear regression models seemed to perform better on the amount of data given, and my assumption on this is that there were not very many highly correlated features and adding the penalty to some of these features in the Lasso regression model seemed to lose some important information. This made the model less able to explain the variation in the sleep efficiency score as shown in the R2 score of the Lasso regression model.

References

UC Davis. (2024, September 20). *Better sleep: Why it's important for your health and tips to sleep soundly*. health. https://health.ucdavis.edu/blog/cultivating-health/better-sleep-why-its-important-for-your-health-and-tips-to-sleep-soundly/2023/03

Wein, H., & Hicklin, T. (Eds.). (2024, June 18). Good sleep for good health. National Institutes of Health. https://newsinhealth.nih.gov/2021/04/good-sleep-good-health