Individual Project

Adyasha Mishra

2024-03-02

**HOW DOES QUALITY OF SLEEP GET AFFECTED? Checking through various factors such as Physical Activity Level, Age, Gender, Daily Steps.**

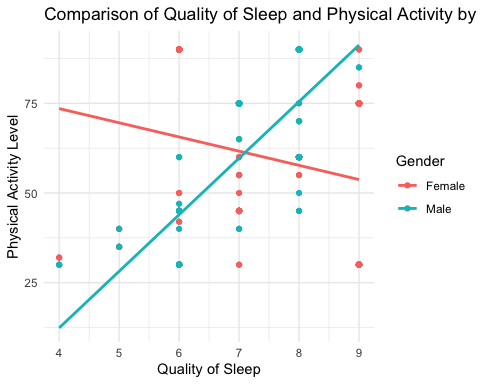
library(readr)  
Sleep\_Health\_Lifestyle <- read\_csv("~/Downloads/Business Intelligence/Sleep\_Health\_Project/Sleep Health Lifestyle.csv")

## Rows: 373 Columns: 13  
## ── Column specification ────────────────────────────────────────────────────────  
## Delimiter: ","  
## chr (5): Gender, Occupation, BMI Category, Blood Pressure, Sleep Disorder  
## dbl (8): Person ID, Age, Sleep Duration, Quality of Sleep, Physical Activity...  
##   
## ℹ Use `spec()` to retrieve the full column specification for this data.  
## ℹ Specify the column types or set `show\_col\_types = FALSE` to quiet this message.

library(ggplot2)

# Create a ggplot scatterplot comparing Quality of Sleep and Physical Activity Level  
ggplot(Sleep\_Health\_Lifestyle, aes(x = `Quality of Sleep`, y = `Physical Activity Level`, color = factor(Gender))) +  
 geom\_point() + # Add points to the plot  
 geom\_smooth(method = "lm", se = FALSE) + # Add linear trend line without standard error band  
 labs(x = "Quality of Sleep", y = "Physical Activity Level", color = "Gender") + # Add axis and legend labels  
 ggtitle("Comparison of Quality of Sleep and Physical Activity by Gender") + # Add title  
 theme\_minimal() # Set minimal theme

## `geom\_smooth()` using formula = 'y ~ x'

 For males, as the physical acitvity level increases, quality of sleep increases. We see a positive correlation here. For females, as the physical activity decreases, quality of sleep increases.

library(dplyr)

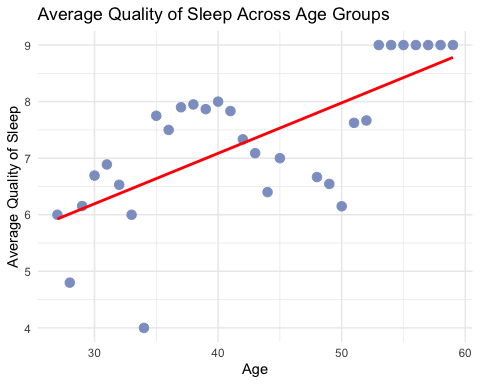
##   
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':  
##   
## filter, lag

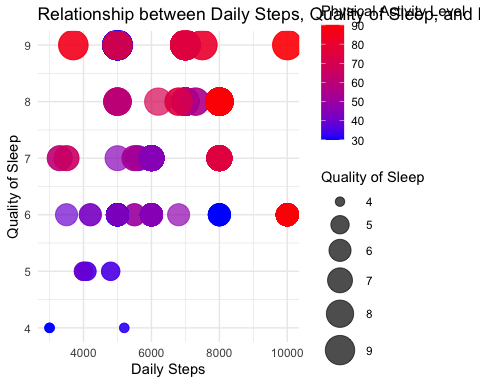
## The following objects are masked from 'package:base':  
##   
## intersect, setdiff, setequal, union

avg\_quality\_sleep <- Sleep\_Health\_Lifestyle %>%  
 group\_by(Age) %>%  
 summarize(avg\_quality\_sleep = mean(`Quality of Sleep`, na.rm = TRUE))  
  
# Plot the average quality of sleep against age with a trend line  
ggplot(avg\_quality\_sleep, aes(x = Age, y = avg\_quality\_sleep)) +  
 geom\_point(color = "#8da0cb", size = 3) +  
 geom\_smooth(method = "lm", se = FALSE, color = "red") + # Add trend line  
 labs(x = "Age", y = "Average Quality of Sleep") +  
 ggtitle("Average Quality of Sleep Across Age Groups") +  
 theme\_minimal()

## `geom\_smooth()` using formula = 'y ~ x'

 We can interpret that average quality of sleep increases as Age increases

ggplot(Sleep\_Health\_Lifestyle, aes(x = `Daily Steps`, y = `Quality of Sleep`, size = `Quality of Sleep`, color = `Physical Activity Level`)) +  
 geom\_point(alpha = 0.7) + # Add points to the plot with transparency  
 scale\_size\_continuous(range = c(3, 10)) + # Set the range of bubble sizes  
 labs(x = "Daily Steps", y = "Quality of Sleep", color = "Physical Activity Level", size = "Quality of Sleep") + # Add axis labels  
 ggtitle("Relationship between Daily Steps, Quality of Sleep, and Physical Activity Level") + # Add title  
 theme\_minimal() + # Set minimal theme  
 scale\_color\_gradient(low = "blue", high = "red") # Set color gradient

 As Physical Activity increases with daily steps, we see that the quality of sleep is better and increases significantly.

# Load the required libraries  
library(rpart)  
library(rpart.plot)  
  
# Check the structure of your dataset  
str(Sleep\_Health\_Lifestyle)

## spc\_tbl\_ [373 × 13] (S3: spec\_tbl\_df/tbl\_df/tbl/data.frame)  
## $ Person ID : num [1:373] 1 2 3 4 5 6 7 8 9 10 ...  
## $ Gender : chr [1:373] "Male" "Male" "Male" "Male" ...  
## $ Age : num [1:373] 27 28 28 28 28 28 29 29 29 29 ...  
## $ Occupation : chr [1:373] "Software Engineer" "Doctor" "Doctor" "Sales Representative" ...  
## $ Sleep Duration : num [1:373] 6.1 6.2 6.2 5.9 5.9 5.9 6.3 7.8 7.8 7.8 ...  
## $ Quality of Sleep : num [1:373] 6 6 6 4 4 4 6 7 7 7 ...  
## $ Physical Activity Level: num [1:373] 42 60 60 30 30 30 40 75 75 75 ...  
## $ Stress Level : num [1:373] 6 8 8 8 8 8 7 6 6 6 ...  
## $ BMI Category : chr [1:373] "Overweight" "Normal" "Normal" "Obese" ...  
## $ Blood Pressure : chr [1:373] "126/83" "125/80" "125/80" "140/90" ...  
## $ Heart Rate : num [1:373] 77 75 75 85 85 85 82 70 70 70 ...  
## $ Daily Steps : num [1:373] 4200 10000 10000 3000 3000 3000 3500 8000 8000 8000 ...  
## $ Sleep Disorder : chr [1:373] "None" "None" "None" "Sleep Apnea" ...  
## - attr(\*, "spec")=  
## .. cols(  
## .. `Person ID` = col\_double(),  
## .. Gender = col\_character(),  
## .. Age = col\_double(),  
## .. Occupation = col\_character(),  
## .. `Sleep Duration` = col\_double(),  
## .. `Quality of Sleep` = col\_double(),  
## .. `Physical Activity Level` = col\_double(),  
## .. `Stress Level` = col\_double(),  
## .. `BMI Category` = col\_character(),  
## .. `Blood Pressure` = col\_character(),  
## .. `Heart Rate` = col\_double(),  
## .. `Daily Steps` = col\_double(),  
## .. `Sleep Disorder` = col\_character()  
## .. )  
## - attr(\*, "problems")=<externalptr>

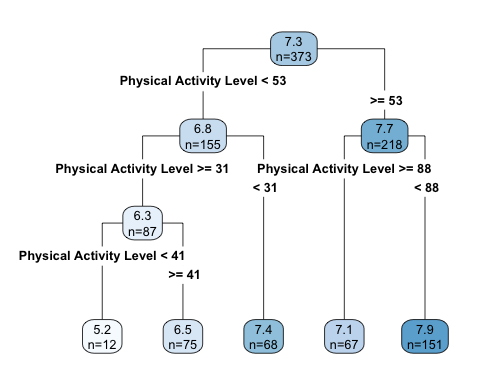
# Create and train the Decision Tree model  
tree\_model <- rpart(`Quality of Sleep` ~ `Physical Activity Level`, data = Sleep\_Health\_Lifestyle)  
print(tree\_model)

## n= 373   
##   
## node), split, n, deviance, yval  
## \* denotes terminal node  
##   
## 1) root 373 531.54420 7.308311   
## 2) Physical Activity Level< 52.5 155 267.97420 6.787097   
## 4) Physical Activity Level>=31 87 51.65517 6.344828   
## 8) Physical Activity Level< 41 12 10.25000 5.250000 \*  
## 9) Physical Activity Level>=41 75 24.72000 6.520000 \*  
## 5) Physical Activity Level< 31 68 177.52940 7.352941 \*  
## 3) Physical Activity Level>=52.5 218 191.52290 7.678899   
## 6) Physical Activity Level>=87.5 67 72.62687 7.074627 \*  
## 7) Physical Activity Level< 87.5 151 83.57616 7.947020 \*

# Summary of the model  
summary(tree\_model)

## Call:  
## rpart(formula = `Quality of Sleep` ~ `Physical Activity Level`,   
## data = Sleep\_Health\_Lifestyle)  
## n= 373   
##   
## CP nsplit rel error xerror xstd  
## 1 0.13554301 0 1.0000000 1.0050237 0.05804080  
## 2 0.07297532 1 0.8644570 0.8777746 0.06000389  
## 3 0.06644774 2 0.7914817 0.8706260 0.05587558  
## 4 0.03139000 3 0.7250339 0.7502756 0.04805129  
## 5 0.01000000 4 0.6936439 0.7196532 0.04669015  
##   
## Variable importance  
## Physical Activity Level   
## 100   
##   
## Node number 1: 373 observations, complexity param=0.135543  
## mean=7.308311, MSE=1.425052   
## left son=2 (155 obs) right son=3 (218 obs)  
## Primary splits:  
## Physical Activity Level < 52.5 to the left, improve=0.135543, (0 missing)  
##   
## Node number 2: 155 observations, complexity param=0.07297532  
## mean=6.787097, MSE=1.728866   
## left son=4 (87 obs) right son=5 (68 obs)  
## Primary splits:  
## Physical Activity Level < 31 to the right, improve=0.1447513, (0 missing)  
##   
## Node number 3: 218 observations, complexity param=0.06644774  
## mean=7.678899, MSE=0.8785456   
## left son=6 (67 obs) right son=7 (151 obs)  
## Primary splits:  
## Physical Activity Level < 87.5 to the right, improve=0.1844161, (0 missing)  
##   
## Node number 4: 87 observations, complexity param=0.03139  
## mean=6.344828, MSE=0.5937376   
## left son=8 (12 obs) right son=9 (75 obs)  
## Primary splits:  
## Physical Activity Level < 41 to the left, improve=0.3230107, (0 missing)  
##   
## Node number 5: 68 observations  
## mean=7.352941, MSE=2.610727   
##   
## Node number 6: 67 observations  
## mean=7.074627, MSE=1.083983   
##   
## Node number 7: 151 observations  
## mean=7.94702, MSE=0.5534845   
##   
## Node number 8: 12 observations  
## mean=5.25, MSE=0.8541667   
##   
## Node number 9: 75 observations  
## mean=6.52, MSE=0.3296

# Plot the decision tree  
rpart.plot(tree\_model, type = 4, extra = 1)



# Make predictions on the training data  
predictions <- predict(tree\_model, Sleep\_Health\_Lifestyle)  
  
# Compare predictions with actual values  
comparison <- data.frame(Actual = Sleep\_Health\_Lifestyle$`Quality of Sleep`, Predicted = predictions)  
head(comparison)

## Actual Predicted  
## 1 6 6.520000  
## 2 6 7.947020  
## 3 6 7.947020  
## 4 4 7.352941  
## 5 4 7.352941  
## 6 4 7.352941

The decision tree suggests that higher levels of physical activity are generally associated with better quality of sleep.

If “Physical Activity Level” < 52.5: Further splits based on different thresholds: < 31: Predicted quality of sleep ≈ 7.35. 31 - 41: Predicted quality of sleep ≈ 6.52. ≥ 41: Predicted quality of sleep ≈ 5.25. If “Physical Activity Level” ≥ 52.5: Further splits: < 87.5: Predicted quality of sleep ≈ 7.95. ≥ 87.5: Predicted quality of sleep ≈ 7.07.